Journal of Labor Research (2020) 41:1–33
https://doi.org/10.1007/s12122-020-09302-7

On the Sensitivity of Wage Gap Decompositions

Martin Huber[1] · **Anna Solovyeva**[2]

Published online: 7 May 2020

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 current values (as frequently seen in wage decompositions) entails smaller explained and larger unexplained components than when including both their current values and histories 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

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 suggested a linear method allowing disentangling the total gap into an explained part that is linked to differences in observed characteristics, for instance education,

✉ Martin Huber
martin.huber@unifr.ch

Anna Solovyeva
anna.solovyeva@unifr.ch

[1] Department of Economics, University of Fribourg, Fribourg, Switzerland

[2] NORC at the University of Chicago, Chicago, IL, USA

and an part that is linked to unobserved factors, for instance discrimination. Several studies proposed non-parametric decomposition methods dropping the linearity instance DiNardo et al. (1996), Barsky et al. (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 et al. (1993), DiNardo et al. (1996), Machado and Mata (2005), Melly (2005), Firpo et al. (2007), Chernozhukov et al. (2009), and Firpo et al. (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 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) 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 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 et al. (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 study, we investigate the sensitivity of average wage gap decompositions to various methods ignoring and considering endogeneity and sample selection, to

[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.

[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.

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. As it is common in the decomposition literature, we consider male wages to be the reference wages in the labor market. The analysed estimators include the Oaxaca-Blinder decomposition; semiparametric inverse probability weighting (IPW, see Hirano et al. 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.[4]

We find the estimates of the total wage gap as well as the explained and unexplained components to differ importantly across some, albeit not all methods. For instance, controlling or not controlling for potential confounders at birth in IPW (while disregarding sample selection) has only a limited impact on the results. However, this can admittedly be due to the omission of important confounders in our small set of observed control variables. In contrast, also tackling sample selection based on observables entails a non-negligibly larger total wage gap and unexplained component, thus reducing the relative importance of the explained component. 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 current values 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 current values 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 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 et al. (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

[4]The methods discussed in this paper may also be used for investigating causal mechanisms in other empirical contexts. Examples include the analysis of the health effect of education operating via specific health behaviors as in Brunello et al. (2016), the assessment of the cognitive and non-cognitive mechanisms through which childcare affects outcomes later in life as in Heckman et al. (2013) and Keele et al. (2015), or the investigation of the causal mechanisms through which job seeker counselling affects employment as in Huber et al. (2017).

parametric and nonparametric methods and 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 et al. (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. One major difference between our study and Goraus et al. (2015) is that they do not consider methods that control for confounders at birth to tackle the endogeneity of the observed characteristics. On the other Goraus et al. (2015) consider further estimators in addition to linear regression and IPW, e.g. matching as in Ñopo (2008), and decompositions at both the mean and quantiles of the wage distribution, using e.g. the method of Juhn et al. (1993), while our analysis is confined to average wage gaps.

The remainder of this paper is organized as follows. Section "Identification" 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 "Data" discusses the NLSY data, sample definition, and descriptive statistics. Section "Empirical Results" presents and interprets the estimation results. Section "Conclusion" concludes.

Identification

Fortin et al. (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, 2011), which in turn may entail discriminatory behavior.

For a formal definition of the causal mechanisms running through observed characteristics X and unobserved factors as parameters of interest, we denote by $Y(g)$

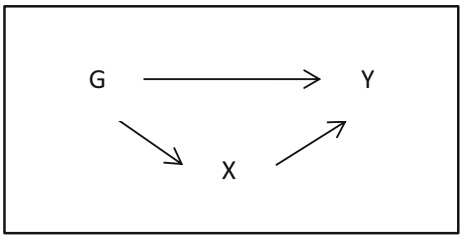


Fig. 1 A graphical representation of the decomposition under Assumption 1

and $X(g)$ the potential outcomes and characteristics when exogenously setting gender G to a specific g, with $g \in \{1, 0\}$.[5] $E(X(1)) - E(X(0))$ gives the average causal effect of G on X (represented by the arrow of G to X in Fig. 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 Fig. 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 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:

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

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

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

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. A graphical representation of this causal framework is given in Fig. 1, where arrows represent causal effects: G influences Y either through X, which corresponds to ψ in Eq. 1, or 'directly', which corresponds to η, but there are presumably no confounders jointly affecting at least two parameters out of G, X, Y.

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 Eqs. 2 and 3 correspond to the probability limits of the explained and unexplained components, respectively, in the Oaxaca-Blinder decompositions. For Eqs. 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.[6]

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):

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

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

Equation 5 is identical to the identification result for the average treatment effect on the non-treated (see Hirano et al. (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

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

G nor X given G are endogenous as postulated in Assumption 1. In the empirical application presented in Section "Empirical Results", we consider both the Oaxaca-Blinder decomposition and estimation based on the sample analogues of Eqs. 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 Fig. 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.

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 et al. (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. 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

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

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

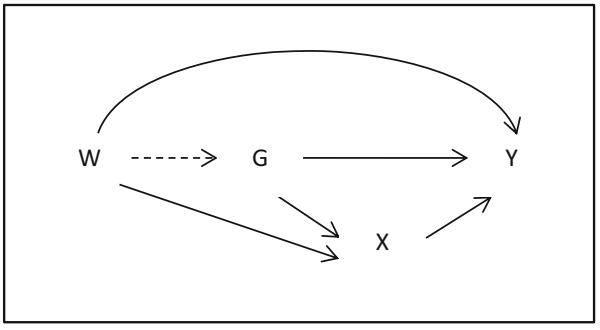


Fig. 2 A graphical representation of the decomposition under Assumption 2

Estimation of (ethnic) wage gaps based on Eqs. 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). 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, 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.

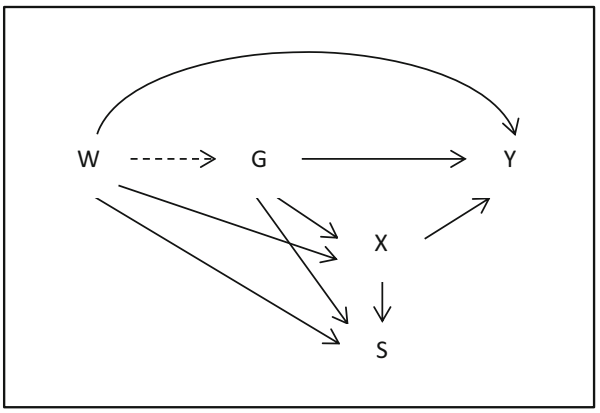


Fig. 3 A graphical representation of the decomposition under Assumption 3

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):

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

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.

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,

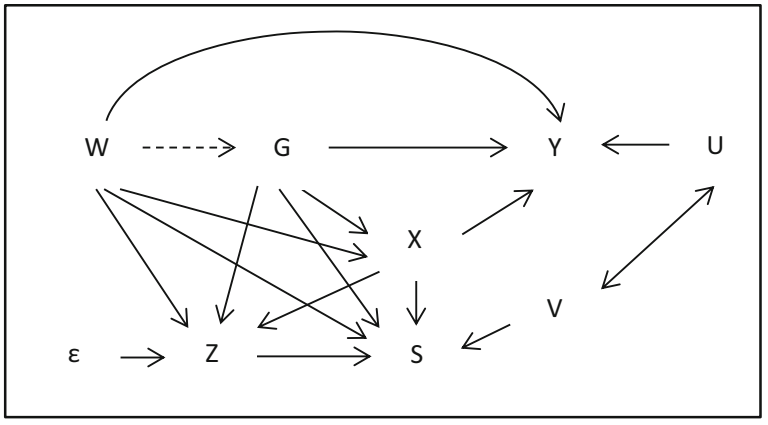


Fig. 4 A graphical representation of the decomposition under Assumption 4

(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 $E[Y(g, X(1)) - Y(g, X(0))|W = w, V = v, S = 1] = E[Y(g, X(1)) - Y(g, X(0))|W = w, V = v]$, 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) G and $\{Y(g, x), X(g')\}$ and (ii) X and $Y(g, x)$ 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:

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

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.[7] 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,[8] our evaluation sample consists of 6,658 individuals (3,162 men and 3,496 women). Table 3 in Appendix 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. This definition of G implies that men earn the reference wages in the labor market, a choice commonly made in the decomposition literature, in which male wages are frequently regarded as the fair or non-discriminatory wages. Sloczynski (2013), however, challenges such an interpretation, as it is conventionally not backed by a theoretical model supporting that male wages correspond to the general equilibrium wages in a world without discrimination. For instance, a weighted version of male and female wages could possibly better reflect the equilibrium wage, if the latter is to be defined as reference wage. In general, different (weights of) reference groups (in terms of their wages) change the magnitudes of the explained and unexplained components and thus, the interpretation of the counterfactual analysis, due to interaction effects between gender and observed characteristics in determining the wage. Analogous issues in defining the reference group arise in racial or other kinds of wage gap decompositions.

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 previous calendar year. This is the case for 87% of males and 70% of females. 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

[7]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 2001).

[8]Specifically, we excluded 502 persons who reported to have worked 1,000 hours or more in the previous calendar year, but whose average hourly wages in the previous 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 previous calendar year. Furthermore, 608 observations with missing values in observed characteristics (see Table 3 for the full list of characteristics) 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.

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 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 along with the history of health problems. 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 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 et al. 2011) and sorting into traditionally male-dominated occupations (Manning et al. 2010; 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, and their interactions with gender as instruments Z for selection into our employment indicator, which allows for differential effects of Z on S across G. 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 the invalidity of the instruments can therefore not be ruled out, 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 and its interaction with gender as additional instruments for selection. This, however, yields quite similar point estimates based on Eqs. 10 and 11 as when using the children-based instruments alone, see the discussion below.

Empirical Results

We decompose the gender wage gap based on the five approaches outlined in Section "Identification". 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, as well as several higher-order and interaction terms.[9] The standard Oaxaca-Blinder decomposition (Oaxaca-Bl.) based on Eqs. 2 and 3 as well as IPW (IPW no W) based on Eqs. 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 Eqs. 6 and 7 and includes potential confounders W listed in Table 3 in Appendix (IPW with W) to tackle endogeneity. Under Assumption 3, IPW based on Eqs. 8 and 9 uses these covariates to control for both endogeneity and selection (IPW MAR). Finally, under Assumption 4, IPW based on Eqs. 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).

[9]The included higher-order terms are marriage history squared and cubed, tenure squared and cubed, and years in current occupation squared and cubed. The interaction terms are between binary indicators for region in 1998 and urban residency, first job before 1975, first job in 1976-79, industry indicators, and employment in 1998; between education indicators and occupation indicators, years in current occupation, and the employment indicator 1998; and between tenure and the urban indicator, occupation indicators, years in current occupation, and the full-time employment indicator in 1998.

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	0.018	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	0.018	0.000	0.109	0.043	0.012	37.1%	0.184	0.043	0.000	62.9%	28	0.5%
IPW with *W*	0.264	0.017	0.000	0.093	0.041	0.024	35.2%	0.171	0.042	0.000	64.8%	28	0.5%
IPW MAR	0.365	0.035	0.000	0.092	0.045	0.040	25.3%	0.273	0.054	0.000	74.7%	90	1.4%
IPW IV	0.137	0.152	0.368	–0.014	0.255	0.956	–10.3%	0.151	0.237	0.523	110.3%	553	8.3%

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[10] of 0.299 is attributed to differences in the included post-group characteristics X, while about 72% (0.215) remains unexplained. All estimates are highly statistically significant.[11] 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, 6, 7, 8, 9, 10, 11, 12 and 13 and Tables 4 and 5 in Appendix resent, respectively, histograms and summary statistics (minimum, mean, and maximum) of the within-group propensity scores used in our IPW-based estimations.[12] 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 below 0.01, we drop 28 units, or 0.5%, from the sample. Compared to the Oaxaca-Blinder decomposition, the explained component is 0.03 log points larger and the unexplained component is by the same amount smaller, while total wage gap remains almost unchanged. For IPW including potential confounders W, Figs. 6 and 7 in Appendix display the histograms of the logit-based estimates of $\Pr(G = 1|W)$ and $\Pr(G = 1|X, W)$ and point to decent common 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 somewhat smaller estimates of the total wage gap as well as the explained and unexplained components when compared to IPW without controls. The results appear, however, quite robust to (not) conditioning on W in our application. We advise against interpreting this too boldly as evidence for negligible endogeneity in wage decompositions, as our control variables might lack important confounders and because sample selection has not been accounted for.

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.[13] Figure 10 in Appendix presents histograms of estimated selection probabilities for individuals who worked less than 1,000 hours in the previous 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

[10] Among the methods considered, the differences in the estimates of the total wage gap are statistically significant at the 10% level between Oaxaca-Blinder and IPW IV, IPW without and with controlling for W, IPW without W and IPW MAR, IPW without W and IPW IV, IPW with W and IPW MAR, IPW with W and IPW IV, and IPW MAR and IPW IV.

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

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

[13] Huber and Solovyeva (2018) provide a simulation study in which the finite sample behavior of IPW with W, IPW MAR, and IPW IV is investigated. The findings suggest that ignoring or not appropriately controlling for sample selection can entail substantial biases, see Tables 1 and 2 therein.

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 unexplained component (0.273 log points) are considerably larger than under IPW controlling for W (but ignoring selection). In contrast, the magnitude of the explained component (0.092 log points) is hardly affected, resulting in an overall drop of its share in the total wage gap to 25%.[14] Any estimates discussed so far are statistically significant at the 5% level.

In addition to controlling for observables, our last estimator, IPW IV, uses the number of children under 15 and under 6 years as well as their interactions with gender as instruments to control for selection, thus allowing for heterogeneous effects of the instruments on the employment decision S across females and males. 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 553 cases (8.3%). This needs to be kept in mind when interpreting the results, as trimming generally changes the target population for which the parameters are estimated. Any IPW IV estimates are far from being statistically significant at any conventional level, pointing to a weak instrument problem. Likelihood ratio tests reveal that the instruments only statistically significantly affect the employment decision of females, but not of males, entailing very low statistical power. Furthermore, the close to zero, but even negative estimate of the explained component appears implausible, which might be driven by the weakness or the potential invalidity of the instruments.

Table 8 in Appendix presents the results for all IPW estimators without trimming extreme propensity scores. The results are quite similar to those in Table 1, with the exception of the point estimates for IPW IV, which are, however, again insignificant. As a further variation in our approach, Table 9 in the appendix presents the results when defining S as a dummy for S working at least 1700 (rather than 1000) hours in the previous year, which roughly reflects the average of full time working individuals (trimming is 0.01). Most findings for the estimators are qualitatively similar to those in Table 1, only the results for IPW IV are again quite different, implausible, and very imprecisely estimated. Furthermore, Table 10 in the appendix provides the estimates when S is a dummy for any positive hours of work provided in the previous year. Also in this case, the results are generally in line with those for the Oaxaca-Blinder decomposition, IPW no W, IPW with W, and IPW MAR in Table 1. Concerning

[14] We note that when defining the female (rather than male) wages as reference (implying that $G = 1$ for women rather than men), the explained component increases and the unexplained component decreases substantially for IPW MAR. For the other decomposition methods, results are more homogeneous across reference group definitions.

IPW IV, it is worth noting that the total wage gap and the unexplained component are statistically significant in Table 10, with quite different point estimates when compared to Table 1, pointing again to the sensitivity of the IV approach in our data.

We conduct several sensitivity checks by gradually reducing the set of post-group characteristics X. Table 11 in Appendix 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 somewhat while the unexplained components increase. The exception is the IPW IV decomposition, in which all estimates are, however, again statistically insignificant.

Our next robustness check excludes not only the higher-order and interaction terms, but also all variables in X that reflect histories of variables over time, like years in marriage, years worked in the current occupation, etc. We point out that many of these variables reflecting developments over several years 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, not only the current occupation likely matters for human capital accumulation and the determination of the current wage, but also the 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 74% to 79% of the total gap across the first four methods. IPW IV yields even more and implausibly extreme estimates, which are, however, far from being statistically significant. Table 2 provides the results.

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 0.03 log points such that the explained component now accounts for only 26% or 25% of the total wage gap when omitting or including control variables W, respectively. For IPW MAR, the reduction in the explained component amounts to 0.02 such that the unexplained part now accounts for 79% of the total wage gap. The IPW IV estimator once again yields rather implausible point estimates that are not statistically significant at any conventional level.

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 well as its interaction with gender as additional instruments for selection S. Table 12 in Appendix shows that the estimates are very much in line with those of IPW IV in the main specification in Table 1. 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

Table 2 Robustness check: parsimonious set of X

s	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	0.023	0.000	77.5%	0	0.0%
IPW no W	0.298	0.018	0.000	0.077	0.027	0.005	25.9%	0.221	0.029	0.000	74.1%	1	0.0%
IPW with W	0.269	0.017	0.000	0.068	0.026	0.010	25.3%	0.201	0.029	0.000	74.7%	2	0.0%
IPW MAR	0.362	0.031	0.000	0.075	0.027	0.006	20.6%	0.288	0.041	0.000	79.4%	1	0.0%
IPW IV	0.129	0.152	0.398	–0.110	0.255	0.667	–85.3%	0.239	0.237	0.313	185.3%	799	12.0%

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

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.

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.

We found the total wage gap as well as the explained and unexplained components to differ importantly across some, but not all of the methods considered. For instance, controlling or not controlling for confounders in IPW (while ignoring sample selection) did not importantly affect the results, which might admittedly be due to omitting important confounders in our limited set of control variables. On the other hand, additionally tackling sample selection based on observables entails a non-negligibly larger total wage gap and unexplained component, thus reducing the relative importance of the explained component. Furthermore, the definition of the observed characteristics related to the explained component mattered: Including only current values of variables rather than both current values 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.

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
Observed characteristics 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

Table 3 (continued)

Variables	Male($G = 1$)	Female($G = 0$)	Difference	p-value
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
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

Table 3 (continued)

Variables	Male($G = 1$)	Female($G = 0$)	Difference	p-value
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
Mother's educ. <high school (ref.)	0.376	0.421	–0.045	0.000
Mother's educ. high school graduate	0.393	0.369	0.024	0.048
Mother's educ. some college	0.094	0.091	0.003	0.616
Mother's educ. college/more	0.076	0.071	0.005	0.411
Father born in U.S.	0.878	0.884	–0.006	0.410
Father's educ. <high school (ref.)	0.351	0.366	–0.015	0.201
Father's educ. high school graduate	0.291	0.297	–0.006	0.560
Father's educ. some college	0.087	0.076	0.011	0.105
Father's educ. college/more	0.131	0.117	0.014	0.085
Order of birth	3.195	3.259	–0.064	0.256

Table 3 (continued)

Variables	Male($G = 1$)	Female($G = 0$)	Difference	p-value
Age in 1979	17.501	17.611	–0.110	0.047
Selection indicator S				
Worked 1,000 hrs or more previous 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	.	.

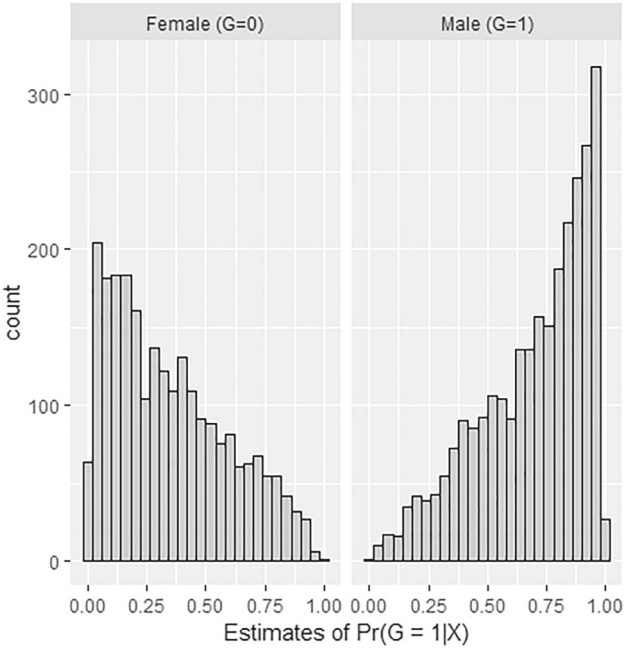


Fig. 5 Distribution of the estimated $\Pr(G = 1|X)$ by treatment states in seleted population

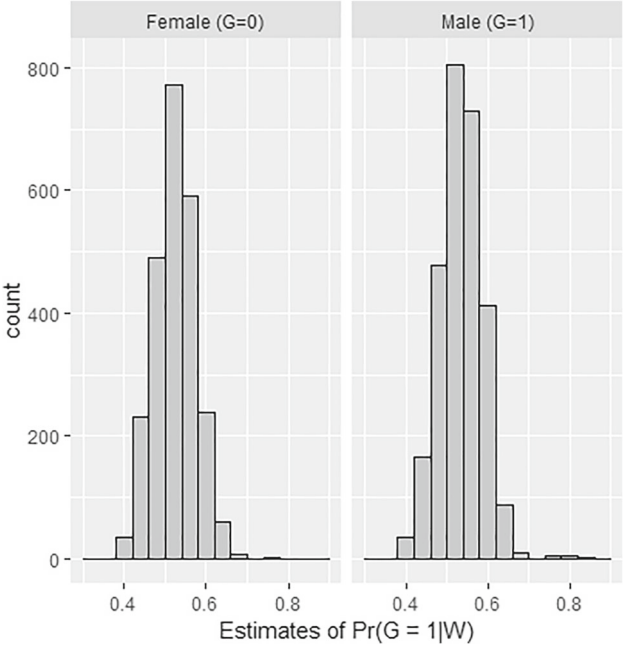


Fig. 6 Distribution of the estimated $\Pr(G = 1|W)$ by treatment states in seleted population

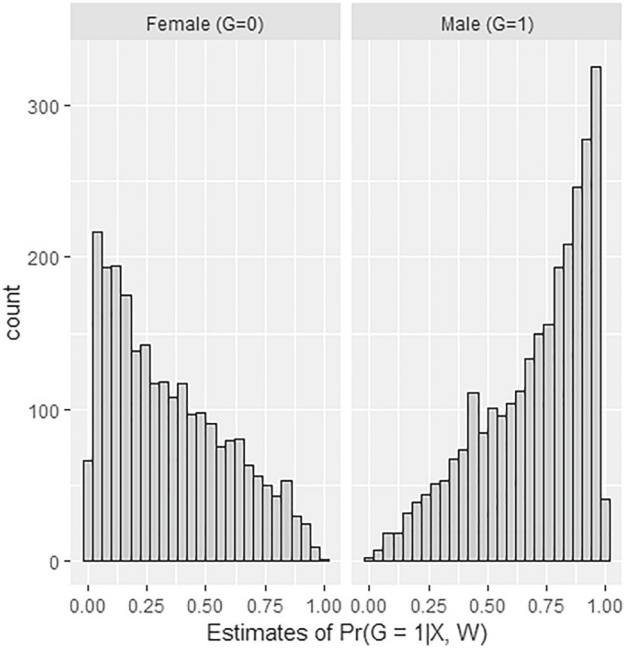


Fig. 7 Distribution of the estimated $\Pr(G = 1|X, W)$ by treatment states in seleted population

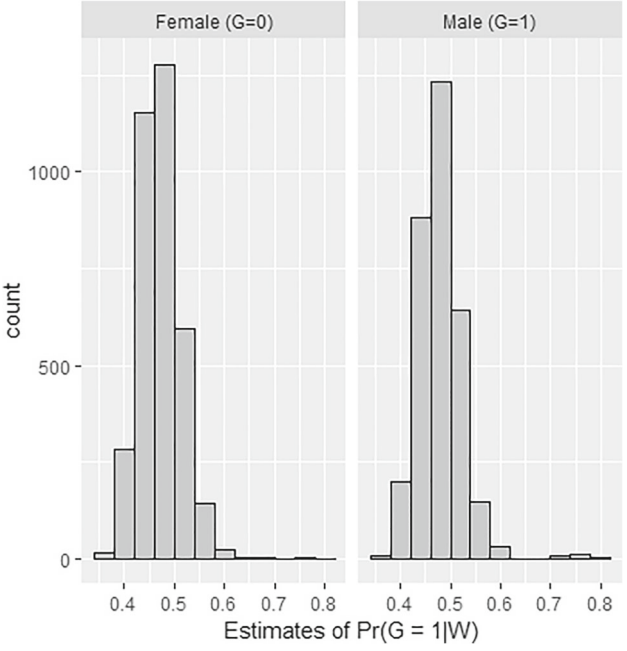


Fig. 8 Distribution of the estimated $\Pr(G = 1|W)$ by treatment states in total population

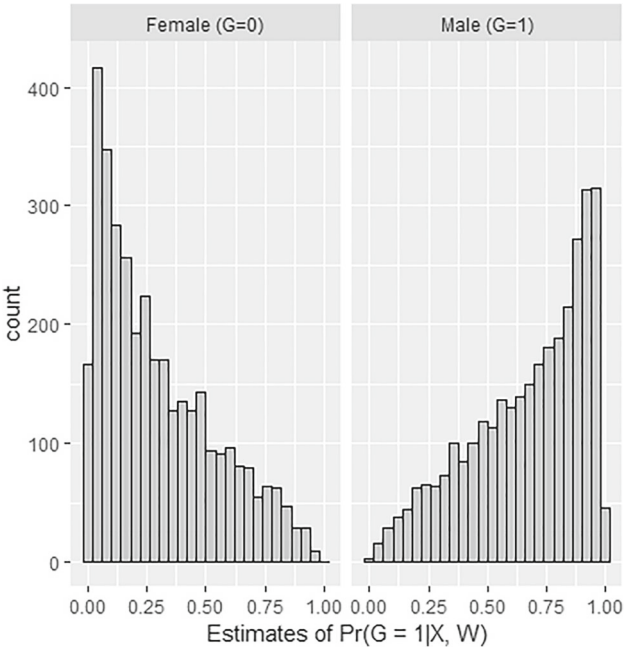
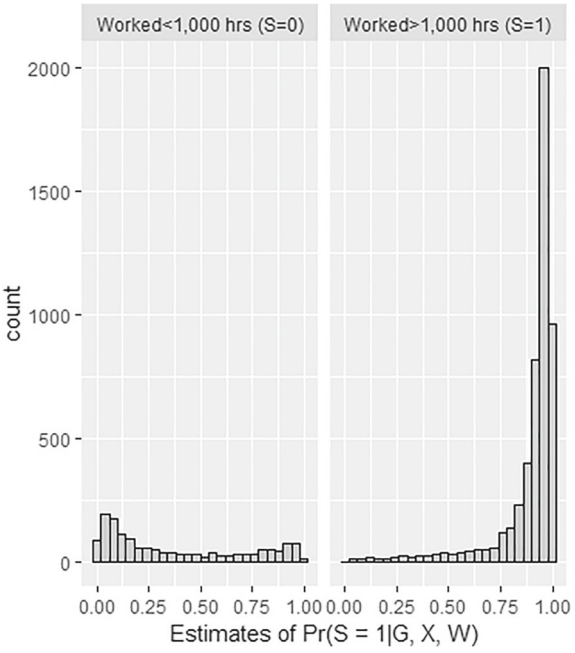


Fig. 9 Distribution of the estimated $\Pr(G = 1|X, W)$ by treatment states in total population

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



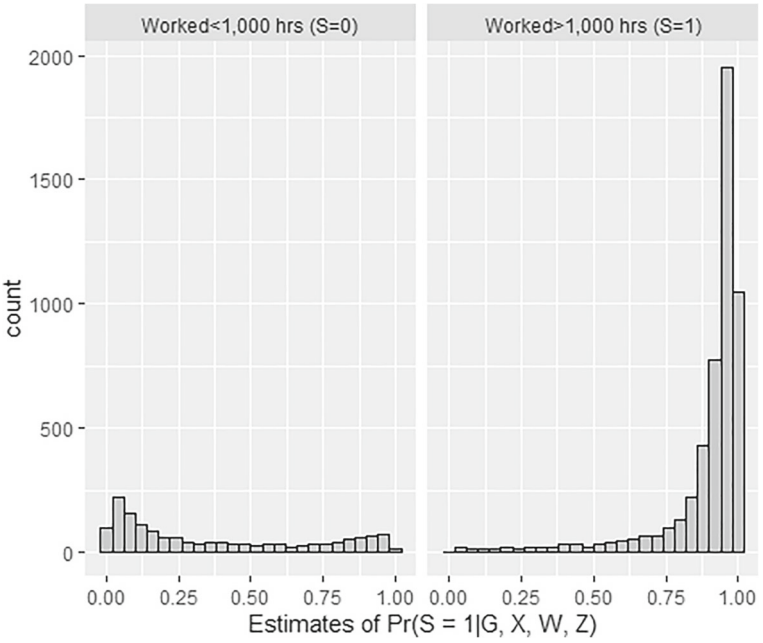


Fig. 11 Distribution of the estimated $p(Q) = \Pr(S = 1|G, X, W, Z)$ by selection states

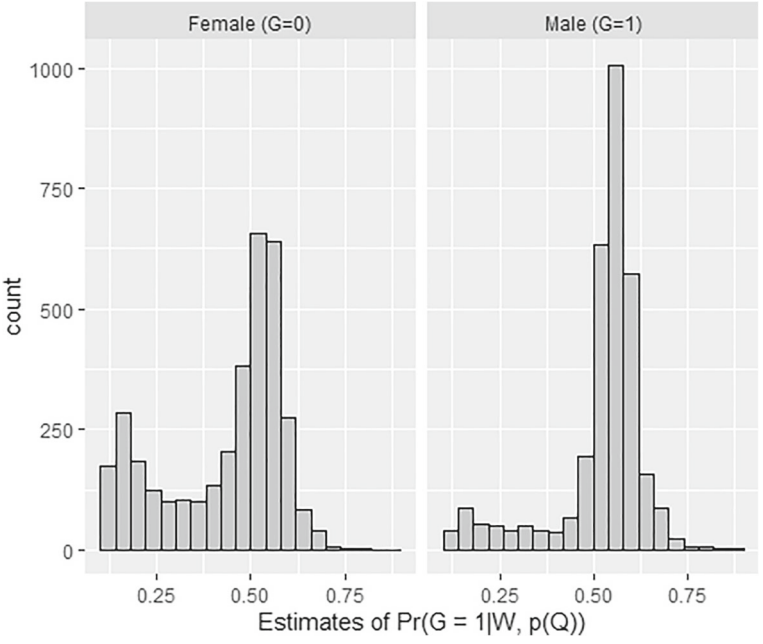


Fig. 12 Distribution of the estimated $\Pr(G = 1|W, p(Q))$ by treatment states in total population

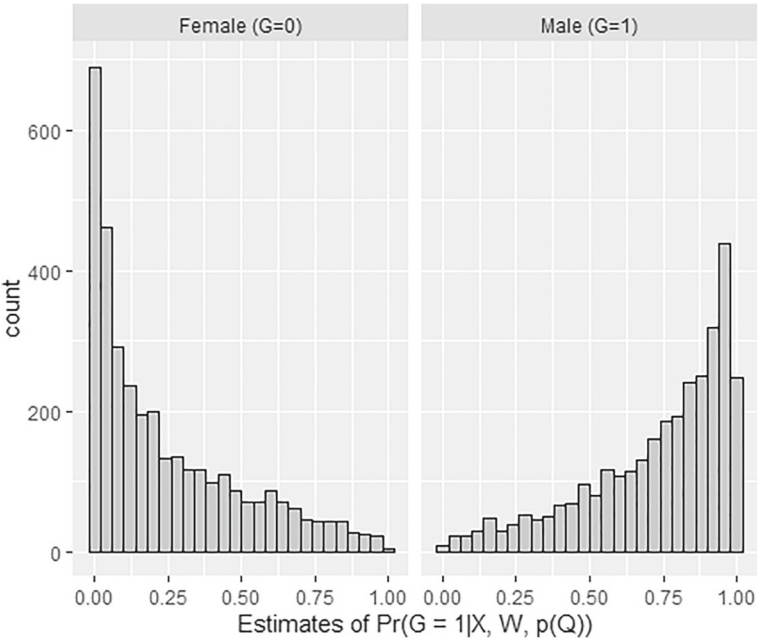


Fig. 13 Distribution of the estimated $\Pr(G = 1|X, W, p(Q))$ by treatment states in total population

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

	Female (G=0)			Male (G=1)		
	Min	Mean	Max	Min	Mean	Max
$\Pr(G = 1\|X)$	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

	Female (G=0)			Male (G=1)		
	Min	Mean	Max	Min	Mean	Max
$\Pr(G = 1\|W)$	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
$\Pr(G = 1\|W, p(Q))$	0.10424	0.43193	0.80992	0.10013	0.52245	0.89331
$\Pr(G = 1\|X, W, p(Q))$	0.00000	0.24461	0.99951	0.00063	0.72955	0.99998

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

	Did not work (S=0)			Worked (S=1)		
	Min	Mean	Max	Min	Mean	Max
$\Pr(S=1\|G, X, W)$	0.00327	0.36952	0.99272	0.02076	0.89392	0.99911
$\Pr(S=1\|G, X, W, Z)$	0.00351	0.36428	0.99259	0.02284	0.89543	0.99904

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
$\Pr(S=1\|G, X, W, Z)<0.01$	33	0.5
$\Pr(G=1\|W, p(Q))<0.01$	0	0.0
$\Pr(G=1\|W, p(Q))>0.99$	0	0.0
$\Pr(G=1\|X, W, p(Q))<0.01$	520	7.8

Table 8 Robustness check: No trimming

	Total gap in log wages			Explained (Indirect)				Unexplained (Direct)			
	est.	s.e.	p-val	est.	s.e.	p-val	% tot.	est.	s.e.	p-val	
IPW no W	0.299	0.019	0.000	0.109	0.046	0.018	36.4%	0.190	0.045	0.000	63.6%
IPW with W	0.269	0.017	0.000	0.093	0.045	0.040	34.5%	0.176	0.045	0.000	65.5%
IPW MAR	0.369	0.036	0.000	0.092	0.048	0.052	25.0%	0.277	0.056	0.000	75.0%
IPW IV	0.197	0.243	0.417	0.130	0.236	0.581	66.0%	0.067	0.246	0.785	34.0%

Notes: Standard errors and p-values are estimated based on 999 bootstrap replications

Table 9 Robustness check: *S* defined as working at least 1700 hours in the previous year

	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.	0.283	0.019	0.000	0.087	0.020	0.000	30.9%	0.195	0.022	0.000	69.1%	0	0.0%
IPW no *W*	0.278	0.018	0.000	0.127	0.044	0.004	45.5%	0.152	0.045	0.001	54.5%	16	0.3%
IPW with *W*	0.241	0.018	0.000	0.101	0.043	0.019	41.8%	0.140	0.044	0.001	58.2%	17	0.4%
IPW MAR	0.371	0.038	0.000	0.109	0.058	0.060	29.3%	0.262	0.069	0.000	70.7%	88	1.3%
IPW IV	0.073	0.225	0.745	-0.205	0.270	0.448	-280.8%	0.278	0.336	0.407	380.8%	1455	21.9%

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 10 Robustness check: *S* defined as working positive hours in the previous year

	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.	0.310	0.020	0.000	0.093	0.024	0.000	30.1%	0.217	0.027	0.000	69.9%	0	0.0%
IPW no *W*	0.303	0.019	0.000	0.114	0.046	0.012	37.8%	0.188	0.047	0.000	62.2%	43	0.8 %
IPW with *W*	0.281	0.018	0.000	0.110	0.044	0.012	39.1%	0.171	0.045	0.000	60.9%	45	0.8 %
IPW MAR	0.372	0.035	0.000	0.076	0.042	0.072	20.5%	0.296	0.053	0.000	79.5%	62	0.9%
IPW IV	0.241	0.076	0.002	0.004	0.123	0.972	1.8%	0.237	0.111	0.032	98.2%	154	2.3 %

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 11 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	0.021	0.000	28.1%	0.215	0.024	0.000	71.9%	0	0.0%
IPW no *W*	0.295	0.018	0.000	0.085	0.039	0.031	28.8%	0.210	0.040	0.000	71.2%	21	0.4%
IPW with *W*	0.265	0.017	0.000	0.067	0.036	0.058	25.4%	0.198	0.037	0.000	74.6%	22	0.4%
IPW MAR	0.375	0.033	0.000	0.053	0.032	0.098	14.1%	0.322	0.045	0.000	85.9%	44	0.7%
IPW IV	0.147	0.152	0.335	0.003	0.255	0.991	2.0%	0.144	0.237	0.543	98.0%	626	9.4%

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 12 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	0.136	0.152	0.372	–0.059	0.255	0.818	–43.2%	0.195	0.237	0.410	143.3%	556	8.4%

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 $p(q) < 0.01$

References

Angrist JD, Pischke JS (2009) Mostly harmless econometrics: an epiricist's companion. Princeton University Press

Arellano M, Bonhomme S (2010) Quantile selection models. unpublished manuscript

Azmat G, Petrongolo B (2014) Gender and the labor market: what have we learned from field and lab experiments? Labour Econ 30:32–40. https://doi.org/10.1016/j.labeco.2014.06.005

Barsky R, Bound J, Charles K, Lupton J (2002) Accounting for the black-white wealth gap: a nonparametric approach. J Am Stat Assoc 97:663–673

Bertrand M (2011) New perspectives on gender. In: Ashenfelter O, Card D, Bertrand M (eds). Elsevier, pp 1543–1590

Bičakova A (2014) Selection into labor force and gender unemployment gaps. CERGE-EI Working Paper 513

Blau F, Kahn L (2006) The US gender pay gap in the 1990s: slowing convergence. Industr Labor Relat Rev 60:45–66

Blinder A (1973) Wage discrimination: reduced form and structural estimates. J Human Resour 8:436–455

Brunello G, Fort M, Schneeweis N, Winter-Ebmer R (2016) The causal effect of education on health: what is the role of health behaviors? Health Econ 25:314–336

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, Fernandez-Val I, Melly B (2009) Inference on counterfactual distributions. CeMMAP working paper CWP09/09

Cobb-Clark DA (2016) Biology and gender in the labor market. IZA DP No. 10386

DiNardo J, Fortin N, Lemieux T (1996) Labor market institutions and the distribution of wages, 1973-1992: a semiparametric approach. Econometrica 64:1001–1044

Duncan OD (1967) Discrimination against negroes. Ann Am Acad Polits Soc Sci 371:85–103

Firpo S, Fortin NM, Lemieux T (2007) Decomposing wage distributions using recentered influence functions regressions. mimeo, University of British Columbia

Firpo S, Fortin NM, Lemieux T (2009) Unconditional quantile regressions. Econometrica 77:953–973

Fortin N, Lemieux T, Firpo S (2011) Chapter 1 - decomposition methods in economics. In: Ashenfelter O, Card D (eds), vol 4, Part A. Elsevier, pp 1–102. https://doi.org/10.1016/S0169-7218(11)00407-2, http://www.sciencedirect.com/science/article/pii/S0169721811004072

Frölich M (2007) Propensity score matching without conditional independence assumption—with an application to the gender wage gap in the United Kingdom. Econometr J 10:359–407

Garbarino E, Slonim R, Sydnor J (2011) Digit ratios (2D:4D) as predictors of risky decision making for both sexes. J Risk Uncertain 42(1):1–26. http://www.jstor.org/stable/23884160

García J, Hernández PJ, López-Nicolás A (2001) How wide is the gap? an investigation of gender wage differences using quantile regression. Empir Econ 26:149–167

Goraus K, Tyrowicz J, van der Velde L (2015) Which gender wage gap estimates to trust? a comparative analysis. Rev Income Wealth 63:118–146

Greiner DJ, Rubin DB (2011) Causal effects of perceived immutable characteristics. Rev Econ Stat 93:775–785

Heckman J (1979) Sample selection bias as a specification error. Econometrica 47:153–161
Heckman J, Pinto R, Savelyev P (2013) Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. Am Econ Rev 103:2052–2086
Heckman JJ (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Ann Econ Soc Meas 5:475–492
Hirano K, Imbens GW, Ridder G (2003) Effects using the estimated propensity score efficient estimation of average treatment. Econometrica 71:1161–1189
Huber M (2014) Identifying causal mechanisms (primarily) based on inverse probability weighting. J Appl Econ 29:920–943
Huber M (2015) Causal pitfalls in the decomposition of wage gaps. J Business Econ Stat 33:179–191
Huber M, Mellace G (2014) Testing exclusion restrictions and additive separability in sample selection models. Empir Econ 47:75–92
Huber M, Solovyeva A (2018) Evaluating direct and indirect effects under sample selection and outcome attrition. SES Working Paper 496, University of Fribourg
Huber M, Lechner M, Mellace G (2017) Why do tougher caseworkers increase employment? the role of program assignment as a causal mechanism. Rev Econ Stat 99:180–183
Imai K, Keele L, Yamamoto T (2010) Identification, inference and sensitivity analysis for causal mediation effects. Stat Sci 25:51–71
Juhn C, Murphy K, Pierce B (1993) Wage inequality and the rise in returns to skill. J Polit Econ 101:410–442
Keele L, Tingley D, Yamamoto T (2015) Identifying mechanisms behind policy interventions via causal mediation analysis. J Policy Anal Manage 34:937–963
Kunze A (2008) Gender wage gap studies: consistency and decomposition. Empir Econ 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. J Labor Econ 16:261–291
Maasoumi E, Wang L (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, Mata J (2005) Counterfactual decomposition of changes in wage distributions using quantile regression. J Appl Econ 20:445–465
Manning JT, Reimers S, Baron-Cohen S, Wheelwright S, Fink B (2010) Sexually dimorphic traits (digit ratio, body height, systemizing–empathizing scores) and gender segregation between occupations: evidence from the bbc internet study. Person Indiv Diff 49(5):511–515. https://doi.org/10.1016/j.paid.2010.05.015, http://www.sciencedirect.com/science/article/pii/S0191886910002552
Manski CF (1989) Anatomy of the selection problem. J Human Resour 24:343–360
Melly B (2005) Decomposition of differences in distribution using quantile regression. Labour Econ 12:577–590
Mora R (2008) A nonparametric decomposition of the mexican american average wage gap. J Appl Econ 23:463–485
Mulligan CB, Rubinstein Y (2008) Selection, investment, and women's relative wages over time. Q J Econ 123:1061–1110
Neuman S, Oaxaca R (2003) Gender versus ethnic wage differentials among professionals: Evidence from Israel. Annales d'Économie et de Statistique 71/72:267–292
Neuman S, Oaxaca R (2004) Wage equations: a methodological note wage decompositions with selectivity-corrected wage equations: a methodological note. J Econ Inequal 2:3–10
Newey WK (2007) Nonparametric continuous/discrete choice models. Int Econ Rev 48:1429–1439
Ñopo H (2008) Matching as a tool to decompose wage gaps. Rev Econ Stat 90:290–299
Nye J, Orel E (2015) The influence of prenatal hormones on occupational choice: 2D:4D evidence from Moscow. Person Indiv Diff 78(Supplement C):9–42. https://doi.org/10.1016/j.paid.2015.01.016, http://www.sciencedirect.com/science/article/pii/S0191886915000367
Oaxaca R (1973) Markets male-female wage differences in urban labour markets. Int Econ Rev 14:693–709
Olivetti C, Petrongolo B (2008) Unequal pay or unequal employment? As cross-country analysis of gender gaps. J Labor Econ 26:621–654

Pearl J (2001) Direct and indirect effects. In: Proceedings of the seventeenth conference on uncertainty in artificial intelligence. Morgan Kaufman, San Francisco, pp 411–420
Robins JM, Greenland S (1992) Identifiability and exchangeability for direct and indirect effects. Epidemiology 3:143–155
Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66:688–701
Rubin DB (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. Port Econ J 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