Universität St.Gallen

Causal pitfalls in the decomposition of wage gaps

Martin Huber

February 2014 Discussion Paper no. 2014-05

School of Economics and Political Science,
Department of Economics

University of St. Gallen

Editor: Martina Flockerzi
University of St.Gallen
School of Economics and Political Science
Department of Economics
Bodanstrasse 8
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35
Email seps@unisg.ch
Publisher: School of Economics and Political Science
Department of Economics
University of St.Gallen
Bodanstrasse 8
CH-9000 St. Gallen
Phone +41 71 224 23 25
Electronic Publication: Fax +41 71 224 31 35
http://www.seps.unisg.ch

Causal pitfalls in the decomposition of wage gaps[1]

Martin Huber

Author's address:

Martin Huber, Ph.D.
SEW-HSG
Varnbüelstr. 14
CH-9000 St. Gallen
Phone +41 71 224 2300
Fax +41 71 224 2302
Email martin.huber@unisg.ch
Website http://www.sew.unisg.ch/en/Assistenzprofessoren

[1] I have benefited from comments by Christina Felfe, Michael Lechner, Andreas Steinmayr, and three anonymous referees.

Abstract

The decomposition of gender or ethnic wage gaps into explained and unexplained components (often with the aim to assess labor market discrimination) has been a major research agenda in empirical labor economics. This paper demonstrates that conventional decompositions, no matter whether linear or non-parametric, are equivalent to assuming a (probably too) simplistic model of mediation (aimed at assessing causal mechanisms) and may therefore lack causal interpretability. The reason is that decompositions typically control for post-birth variables that lie on the causal pathway from gender/ethnicity (which are determined at or even before birth) to wage but neglect potential endogeneity that may arise from this approach. Based on the newer literature on mediation analysis, we therefore provide more attractive identifying assumptions and discuss non-parametric identification based on reweighting.

Keywords

Wage decomposition, causal mechanisms, mediation.

JEL Classification

J31, J71, C21, C14.

1 Introduction

The decomposition of empirically observed wage gaps across gender or ethnicity has continued to attract substantial attention in labor economics. The idea is to disentangle the total gap into an explained component that can be attributed to differences in (observed) labor market relevant characteristics such as education and occupation and an unexplained remainder, which is often interpreted as discrimination. In addition to the classical linear decomposition of Blinder (1973) and Oaxaca (1973), the use of more flexible non-parametric decomposition methods has been proposed for instance in DiNardo, Fortin, and Lemieux (1996), Barsky, Bound, Charles, and Lupton (2002), Frölich (2007), Mora (2008), and Nopo (2008). Furthermore, the literature has also moved from the mere assessment of mean gaps to decompositions at particular quantiles in the outcome distribution, see 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).

These advancements in the estimation of decompositions stand in stark contrast to the lack of a formal identification theory in many if not most studies, as also pointed out by Fortin, Lemieux, and Firpo (2011): "*In econometrics, the standard approach is to first discuss identification(...)and then introduce estimation procedures to recover the object we want to identify. In the decomposition literature, most papers jump directly to the estimation issues (i.e. discuss procedures) without first addressing the identification problem.*"

To close this gap, the main contribution of this paper is to shed light on the (plausibility of the) identifying assumptions required for disentangling explained and unexplained components. It will be demonstrated that conventional decompositions of gender and ethnic wage gaps can be equivalently expressed as a system of equations that corresponds to a (probably too) simplistic model for mediation analysis, which allows explicating the rather strong identifying assumptions underlying most of the literature. Mediation analysis, as outlined in the seminal paper of Baron and Kenny (1986), aims at disentangling the causal mechanisms through which an explanatory variable affects an outcome of interest, with mediators being intermediate outcomes lying on the causal pathway between the explanatory variable and the outcome. Applied to the context of wage decompositions, gender and ethnicity can be regarded as variables that stand at the beginning of any individual's causal chain affecting wage, because they are determined at or prior to birth. Education, occupation, work experience etc. are all mediators because they occur later in life and are thus potentially driven by gender and ethnicity, while the mediators themselves likely affect wage. If we accept this causal structure, which in terms of the time line of events in life appears to be the only reasonable choice, then the explained component of the wage gap can be shown to correspond to the "indirect" wage effect of gender or ethnicity that operates through these mediators. Conversely, the unexplained component equals the "direct" effect of gender or ethnicity on wage that either is inherently direct or operates through unobserved mediators (e.g. discrimination).

Disentangling direct and indirect effects requires conditioning on the mediators while also controlling for confounders jointly related with the outcome, the mediators, and/or the initial explanatory variable, see Judd and Kenny (1981) for an early discussion of the confounding problem in mediation. The identification issue arising in many decompositions is that they merely incorporate mediators but typically neglect potential confounders, which may jeopardize the causal interpretability of the explained and unexplained components. E.g., assume that even conditional on the initial variable gender or ethnicity, family background (such as parents' education) affects both the mediator education and the outcome variable wage, for instance through unobserved personality traits like self-esteem (see Heckman, Stixrud, and Urzua (2006)).[1] Then, using education in the decomposition without controlling for family background generally biases the explained and unexplained components in the wage decomposition. The reason is that education is itself already an intermediate outcome such that conditioning on it without accounting for confounders is likely to introduce bias, a problem thoroughly discussed in Rosenbaum (1984) and Robins and Greenland (1992), among many others.

[1] Analyzing Brazilian earnings data, Lam and Schoeni (1993) argue in a similar manner that family background is correlated with unobserved worker characteristics that affect earnings, while also driving education.

It is important to note that theoretical cases in which conventional decompositions bear a causal interpretation do exist, but may appear unrealistic. The first scenario is that there are no confounders of gender/ethnicity –henceforth referred to as group variables– and/or the mediators, which seems quite restrictive. The second scenario is a reversal of the causal chain of the group variable and the mediators. If the observed characteristics related to the explained component were determined prior to the group variable (rather than vice versa) and included all confounders of the group variable, the decomposition would satisfy the kind of conditional independence assumption frequently imposed in the treatment evaluation literature, see Rosenbaum and Rubin (1983) and Imbens (2004). Fortin, Lemieux, and Firpo (2011) point out that in this case, the unexplained component corresponds to the "treatment effect on the treated" or "non-treated" (depending on the reference group) of the treatment evaluation literature, while the explained component reflects the selection bias into the group variable. Unfortunately, this framework appears ill-suited for decompositions related to gender and ethnicity, which naturally stand at the beginning of any causal chain in an individual's life. Even potentially interesting group variables occurring later in life like unionization may at least partially affect variables we would consider important for the explained component, such as tenure with the current employer.

Therefore, as a second contribution this paper suggests the use of arguably more realistic identifying assumptions for wage decompositions (by borrowing from the newer literature on mediation analysis) that allow for confounding of the group variable and mediators and also discusses non-parametric identification. Given the satisfaction of the assumptions, the unexplained and explained components bear a clear causal interpretation. Identification relies on a sequential ignorability assumption as (among others) discussed in Imai, Keele, and Yamamoto (2010), implying that all confounders of the group variable and the mediators are observed. Using the approach suggested in Huber (2013), the explained and unexplained components are then non-parametrically identified by reweighting observations as a function of their propensities to be in a group (i) given the confounders and (ii) given the confounders and mediators. We provide a brief simulation study that conveys the identification issues in wage decompositions. As an empirical illustration, we use the National Longitudinal Survey of Youth 1979 (NLSY79) to estimate the decomposition of the ethnic wage gap among males in the year 2000. Our results are not robust to the choice of identifying assumptions, suggesting that conventional decompositions may be substantially biased. Albeit more general than the standard decomposition, it has to be stressed that the assumptions proposed in this paper are not innocuous either. Besides requiring the observability of all confounders, they also rule out that some confounders of the mediators are themselves a function of the group variable. Alternative identification strategies allowing for the latter case are therefore also briefly discussed and may be considered in future research on wage decompositions.

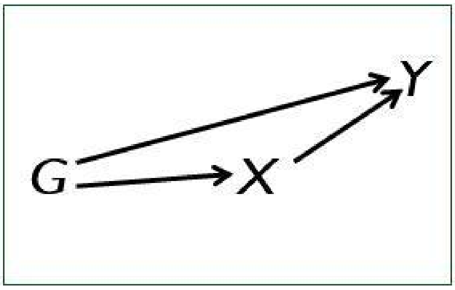
The remainder of this paper is organized as follows. Section 2 shows the equivalence of conventional decompositions and simplistic mediation models, given that the group variable precedes the observed characteristics (mediators). Secondly, it discusses arguably more realistic identifying assumptions and identification based on reweighting. Section 3 provides a brief simulation study, while Section 4 presents an application to the NLSY79. Section 5 concludes.

2 Models and identifying assumptions

2.1 Notation and definition of parameters

For reasons discussed in Section 1, we will assume throughout that the group variable precedes the observed characteristics we would like to adjust for in the decomposition. Let G denote the binary group variable (e.g. female/male or black/white), Y the outcome of interest (e.g. log wage) and X the vector of observed characteristics (e.g. age, education, work experience, occupation, industry). X_k corresponds to the k^{th} element in X with $k \in \{1, 2, .., K\}$ and K being the number of characteristics so that $X = [X_1, ..., X_K]$. A graphical representation of the causal mechanism considered is given in Figure 1, where the arrows represent a causal relation. G (e.g. gender) has an effect on the mediators X (e.g. education and occupation) which themselves affect Y (log wage). At the same time, G also influences Y "directly". However, it is important to note that this causal path need not be inherently direct, but may also or even exclusively include (unobserved) mediators not appearing in X. E.g., gender and ethnicity most likely affect the perception of individual traits by decision makers in the labor market (see the discussion in Greiner and Rubin (2011)), which in turn may entail discriminatory behavior that adds to the unexplained component in wage gaps.[2]

Figure 1: A graphical representation of the decomposition



To formally define the causal parameters of interest, we denote by $Y(g)$ and $X(g)$ the potential outcome and mediators when exogenously setting G to g, with $g \in \{1,0\}$ (see for instance Rubin (1974) for an introduction to the potential outcome framework). $E(Y(1)) - E(Y(0))$ therefore gives the total average causal effect of G on Y, represented by the sum of direct and indirect (i.e. operating through X) effects in Figure 1. $E(X(1)) - E(X(0))$, on the other hand, is 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. Furthermore, assume that the conditional mean of the potential outcome $Y(g)$ given X is characterized by the following linear model:

$$E(Y(g)|X) = c(g) + \sum_{k=1}^{K} X_k \beta_k(g), \qquad g \in \{1,0\}. \tag{1}$$

$c(g)$ is a constant and the vector $\beta(g) = [\beta_1(g), ..., \beta_K(g)]$ contains the (linear) effects of the

[2] Even if we followed the principle "No causation without manipulation" and were uncomfortable with investigating causal effects of immutable characteristics like gender and ethnicity, see Greiner and Rubin (2011), acknowledging that the causal relations in Figure 1 run through perceived treats may render wage decompositions nevertheless interesting. The reason is that in contrast to *factual* gender and ethnicity, the *perception* of these (or related) characteristics may in principal be (experimentally) manipulable, see for instance Bertrand and Mullainathan (2004).

mediators on the potential outcome, which (as implied by our notation) may differ across $g = 1, 0$ and therefore allow for group-mediator-interaction effects. The sum of elements in $\beta(g)$ corresponds to the 'second stage' of the indirect effect, namely the causal arrow from X to Y in Figure 1, which is permitted to depend on the group state. After having defined all ingredients of our causal framework, the total causal effect can be decomposed into the indirect effect explained by the mediators, denoted by ψ, and the unexplained direct effect, denoted by η, as follows:

$$E(Y(1)) - E(Y(0)) = \underbrace{\sum_{k=1}^{K}[E(X_k(1)) - E(X_k(0))]\beta_k(1)}_{\text{explained component } (\psi)} + \underbrace{(c(1) - c(0)) + \sum_{k=1}^{K} E(X_k(0))(\beta_k(1) - \beta_k(0))}_{\text{unexplained component } (\eta)}. \quad (2)$$

To see this, first note that $E(Y(1)) - E(Y(0)) = E[Y(1)|E(X(1))] - E[Y(0)|E(X(0))]$ under the linear model postulated in (1). Second, $E[Y(1)|E(X(1))] - E[Y(0)|E(X(0))] = \{E[Y(1)|E(X(1))] - E[Y(1)|E(X(0))]\} + \{E[Y(1)|E(X(0))] - E[Y(0)|E(X(0))]\}$ by subtracting and adding $E[Y(1)|E(X(0))]$. Third, again by equation (1), $E[Y(1)|E(X(1))] - E[Y(1)|E(X(0))] = \psi$ and $E[Y(1)|E(X(0))] - E[Y(0)|E(X(0))] = \eta$.

2.2 Linking conventional wage gap decompositions to mediation models

We demonstrate that the linear decomposition of Blinder (1973) and Oaxaca (1973) (i) can be equivalently expressed as a simplistic mediation model and (ii) fails to identify the causal paths in Figure 1 or equivalently, the true explained and unexplained components in (2) if G and/or X is confounded. We start by recalling the standard decomposition, where it is assumed that the observed outcome within a group is characterized by the following linear equation:

$$Y^g = c^g + \sum_{k=1}^{K} X_k^g \beta_k^g + \epsilon^g, \qquad g \in \{1, 0\}. \quad (3)$$

Y^g and X_k^g denote the conditional outcome and the conditional observed characteristics given $G = g$.[3] Likewise, c^g and $\beta^g = [\beta_1^g, ..., \beta_K^g]$ represent the constant and the slope coefficients, while

[3] Note the difference to equation (1), which is defined in terms of potential outcomes $Y(g)$ rather than conditional outcomes Y^g.

ϵ^g denotes the error term, again conditional on the group variable. It is typically assumed (see for instance Fortin, Lemieux, and Firpo (2011)) that

$$E(\epsilon^g|X^g) = E(\epsilon|X, G = g) = 0, \qquad g \in \{1, 0\}, \tag{4}$$

for which conditional independence of ϵ^g and X^g is a sufficient condition (while ϵ denotes the unconditional error).

We assume that we have an i.i.d. sample drawn from $\{Y, X, G\}$ consisting of n observations. Denote by $\bar{Y}^g$ and $\bar{X}^g$ the sample averages of Y and X conditional on $G = g$, with $\bar{X}^g = [\bar{X}^g_1, ..., \bar{X}^g_K]$. The standard decomposition of Blinder (1973) and Oaxaca (1973) is given by

$$\bar{Y}^1 - \bar{Y}^0 = \underbrace{\sum_{k=1}^{K}(\bar{X}^1_k - \bar{X}^0_k)\hat{\beta}^1_k}_{\text{explained component }(\hat{\psi})} + \underbrace{(\hat{c}^1 - \hat{c}^0) + \sum_{k=1}^{K}\bar{X}^0_k(\hat{\beta}^1_k - \hat{\beta}^0_k)}_{\text{unexplained component }(\hat{\eta})}. \tag{5}$$

$\hat{\beta}^1_k$, $\hat{\beta}^0_k$ denote the estimates of the coefficients β^1_k, β^0_k and $\hat{\eta}$, $\hat{\psi}$ those of the unexplained and explained components, respectively.

If G stands at the beginning of the causal mechanism affecting Y as displayed in Figure 1, one can easily translate the standard decomposition into a simplistic mediation model. The latter consists of a system of linear equations characterizing each mediator as a function of G and an error term ν and the outcome as a function of G, X, the interactions of G and X, and an error term ϵ:

$$X_k = c_{X_k} + G\alpha_k + \nu_k, \quad \text{for } k \in \{1, ..., K\}, \tag{6}$$

$$Y = c_Y + G\delta + \sum_{k=1}^{K} X_k\theta_k + \sum_{k=1}^{K} GX_k\gamma_k + \epsilon. \tag{7}$$

c_{X_k}, α_k, ν_k denote the constant, coefficient, and error term in the equation of the k^{th} element in X. If applicable, we denote by the respective parameter without subscript k the vector of K elements, e.g. $\alpha = [\alpha_1, \alpha_2, ..., \alpha_K]$. c_Y, δ, θ represent the constant and the coefficients on G and

X in the outcome equation, respectively, while γ denotes the coefficients on the interactions. It is obvious that a causal interpretation of the various coefficients and the identification of direct and indirect effects generally requires that ν is not associated with G, and ϵ is not associated with G and X. Therefore, this rather simple model rules out confounding of G and/or X.

We now demonstrate that the indirect effect of G on Y working through X in our linear mediation model (i.e. the effect of G on X times the effect of X on Y) corresponds to the explained component in a wage decomposition. Imai, Keele, and Yamamoto (2010) provide the expression of the indirect effect for a scalar mediator X_k when interaction between G and X_k is allowed for as in (7): $\alpha_k(\theta_k + G\gamma_k)$. It follows that the (overall) indirect effect through all mediators is the sum of the indirect effects through each mediator:

$$\text{(overall) indirect effect} = \sum_{k=1}^{K} \alpha_k(\theta_k + G\gamma_k). \tag{8}$$

Considering X_k, we note that in equation (6), which contains a single binary regressor, the constant corresponds to the average mediator conditional on $G = 0$, i.e. $c_{X_k} = E(X_k^0)$, while $\alpha_k = E(X_k^1) - E(X_k^0)$. Furthermore, any coefficient θ_k reflects the effect of X_k on Y conditional on $G = 0$ and is therefore equal to β_k^0 in (3). Correspondingly, the coefficient on the interaction term of G and X_k gives the difference in the effects of X_k given $G = 1$ and $G = 0$, i.e. $\gamma_k = (\beta_k^1 - \beta_k^0)$. It follows that the indirect effect given in (8) conditional on $G = 1$ is identical to the probability limit (plim) of the explained component $\hat{\psi}$ in (5): $\sum_{k=1}^{K} \alpha_k(\theta_k + G\gamma_k) = \sum_{k=1}^{K} [E(X_k^1) - E(X_k^0)](\beta_k^0 + (\beta_k^1 - \beta_k^0)) = \sum_{k=1}^{K} [E(X_k^1) - E(X_k^0)]\beta_k^1$.

Secondly, we show that the direct effect is equal to the unexplained component. Imai, Keele, and Yamamoto (2010) provide the expression for the direct effect under our linear mediation model:

$$\text{direct effect} = \delta + \sum_{k=1}^{K} \gamma_k(c_{X_k} + \alpha_k G). \tag{9}$$

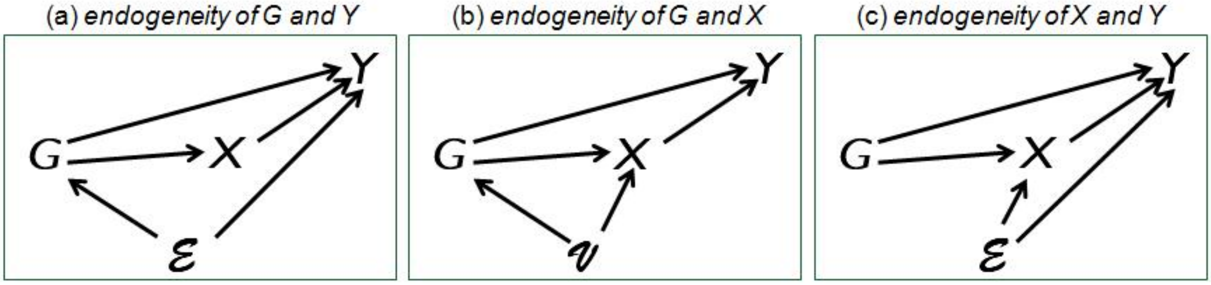
Note that δ reflects the part of the direct effect that is net of interactions with X. This obviously corresponds to the differences in the constants in equation (3) for $G = 1$ and $G = 0$, which would be an alternative way to identify the effect of switching G when holding X fixed. Therefore, $\delta = c^1 - c^0$. Conditional on $G = 0$, (9) simplifies to $\delta + \sum_{k=1}^{K} c_{X_k} \gamma_k$. By recalling that $c_{X_k} = E(X_k^0)$ and $\gamma_k = (\beta_k^1 - \beta_k^0)$, it follows that this corresponds to the plim of the unexplained component $\hat{\eta}$ in (5): $\delta + \sum_{k=1}^{K} \gamma_k (c_{X_k} + \alpha_k G) = (c^1 - c^0) + \sum_{k=1}^{K} E(X_k^0)(\beta_k^1 - \beta_k^0)$.

Linking the standard decomposition to the mediation model highlights that the estimates $\hat{\psi}$ and $\hat{\eta}$ in (5) only bear a causal interpretation and converge to the true explained and unexplained components ψ and η in (2) as n increases under the following conditions: (i) G is exogenous, i.e. not confounded by ν, and (ii) X is exogenous conditional on G, i.e. not confounded by ϵ such that equation (4) holds. Exogeneity of G implies that the group variable is as good as randomly assigned w.r.t. ϵ (which enters Y) as well as any unobservables affecting X. This would for instance be violated if family background (e.g. parents' education and socio-economic status) was unobserved, entered ϵ, and was also correlated with the group variable. Therefore, one might be particularly concerned if G represents ethnicity, while gender is arguably randomly assigned by nature. However, even in the latter case, one could think of possible violations of exogeneity if for instance gender bias and thus, the gender ratio (through the inclination to get more children conditional on the gender of the first child) differed systematically across socio-economic groups.

If G is confounded by ϵ, the conditional mean outcome $E(Y^g) = E(Y|G = g)$, which is the plim of $\bar{Y}^g$, does not correspond to the mean potential outcome $E(Y(g))$. Therefore, the left hand side of equation (5) does not converge to $E(Y(1)) - E(Y(0))$ such that not even the total causal effect of G on Y (the sum of the explained and unexplained components) is identified. Similarly, there might exist unobservables that jointly affect G and X, which could even overlap with those in ϵ (e.g. family background drives educational choices). Then, $E(X^g) = E(X|G = g)$, the plim of $\bar{X}^g$, does not correspond to the mean potential mediator state $E(X(g))$. It follows that the right hand side of (5) is asymptotically biased (i.e. does not correspond to the right hand side of (2)) even if the left hand side (i.e. the total effect) was identified.

We now consider the second issue that conditional exogeneity of X given G does not hold,

Figure 2: Scenarios in which identification fails


(a) endogeneity of G and Y
G
X
Y
ε
(b) endogeneity of G and X
G
X
Y
ν
(c) endogeneity of X and Y
G
X
Y
ε

implying that the effect of X on Y is confounded by ϵ even conditional on G. As in standard OLS models, this implies that the coefficients β^g and c^g in equation (3) differ from the parameters $\beta(g)$ and $c(g)$ in the true causal model (1). Therefore, the right hand side of (5) does not converge to the true decomposition (2) even if G was exogenous. Figure 2 illustrates scenarios in which identification fails. In (a), ϵ jointly affects G and Y whereas in (b), the unobserved term ν jointly affects G and X. Under both (a) and (b), exogeneity of G does not hold. In (c), ϵ jointly affects X and Y (even conditional on G) such that (4) is violated. In applications, several issues might occur at the same time. Note that identification also fails if the respective unobserved terms do not directly influence G and/or X, but are correlated with further unobservables that affect the group variable or the mediator, respectively. Furthermore, ϵ and ν might be correlated or even overlap.

Summing up, formulating the standard decomposition in terms of a mediation model facilitates understanding the assumptions required for causal inference, which are obviously not very attractive: c_X and α and thus, $E(X(0))$ and $E(X(1)) - E(X(0))$ are only identified if G is not confounded by ν. Likewise, δ, θ, and γ and thus, $c(1) - c(0)$, $\beta(0)$, and $(\beta(1) - \beta(0))$ are only identified if ϵ does not confound G and/or X. Otherwise, the unexplained and explained components in (5) are asymptotically biased and not to be causally interpreted. In addition, the standard decomposition imposes linearity and homogenous effects within group states. Only if all these conditions do hold, the explained component corresponds to the indirect effect in a mediation model for $G = 1$, e.g. being male. It then reflects the wages differentials between males and females due to group-induced differences in the mediators $(X(1), X(0))$ assessed at the males' rate of returns to the mediators (e.g. the returns to education and occupation). In contrast, the unexplained component corresponds to the direct effect for $G = 0$, e.g. being female. It is the (potentially discriminatory) effect of gender on wage when holding the mediators fixed at their values among females, $X(0)$. As well acknowledged in the mediation literature, the direct and indirect effects defined on opposite group states add up to the total effect of G, namely $E(Y(1)) - E(Y(0))$, the left hand side of (2).

The question remains whether standard wage decompositions are useful at all in the arguably more realistic case of confounding, where the average observed differences $E(Y^1) - E(Y^0)$ and $E(X^1) - E(X^0)$ deviate from the average causal effects $E(Y(1) - Y(0))$ and $E(X(1) - X(0))$. While observed differences are descriptive about the status quo, they appear less suitable for deriving policy conclusions, which usually rely on causal inference based on comparing counterfactual states of the world. Basing wage decompositions on observed differences without controlling for confounders therefore gives explained and unexplained components that are hard to interpret. As elsewhere in econometrics, it is not obvious what to make of numbers obtained by conditioning on a battery of endogenous variables (in our case the mediators and the group variable). In this light, the terminology 'explained component' may even appear misleading, because neither do endogenous group states causally explain observed differences in mediators, nor do endogenous group states and mediators causally explain observed differences in outcomes. One should also be cautious with interpreting the unexplained component as labor market discrimination, as often seen in the literature, because this would be a causal claim, too: namely, that a part of the supposedly causal effect of G on wage which materializes in the observed wage gap is not driven by observed differences in the mediators, but by other (and unobserved) causal mechanisms like discrimination. As for the explained component, endogeneity jeopardizes the identification of the unexplained component based on standard decompositions. The latter may therefore not be very meaningful for deriving policy recommendations, as also argued by Kunze (2008).

2.3 An alternative set of identifying assumptions

We suggest the use of a different set of assumptions that extends standard decompositions in two dimensions: Firstly, we allow for confounding of the group variable and the mediators, as long as it is related to observed covariates, which we denote by W. Secondly, we ease the linearity assumptions, which may entail misleading results due to functional form restrictions and inadequate extrapolation (see Barsky, Bound, Charles, and Lupton (2002)). Instead, we consider (at least for identification) a fully non-parametric model that maximizes flexibility in terms of model specification. To this end, we introduce a more general notation for the explained and unexplained components that comes from the non-parametric mediation literature (e.g. Pearl (2001) and Robins (2003)). First of all, note that instead of defining a potential outcome as a function of the group variable only, we may equivalently define it as a function of the group variable and the potential mediators on the causal pathway from the group to the outcome. In fact, $Y(g) = Y(g, X(g))$. This change in notation makes explicit that the potential outcome is affected by the group variable both directly and indirectly via $X(g)$. It allows us to rewrite the total effect of G on Y (the left hand side of (2)) as $E(Y(1)) - E(Y(0)) = E[Y(1, X(1))] - E[Y(0, X(0))]$. More importantly, the latter expression can (by subtracting and adding $E[Y(1, X(0))]$ on the right hand side) be disentangled into the indirect effect (or explained component) due to varying $X(1)$ and $X(0)$ while keeping the group fixed at $G = 1$ and the direct effect (or unexplained component) of G while keeping the mediators fixed at $X(0)$:

$$E[Y(1, X(1))] - E[Y(0, X(0))] = \underbrace{E[Y(1, X(1))] - E[Y(1, X(0))]}_{\psi} + \underbrace{E[Y(1, X(0))] - E[Y(0, X(0))]}_{\eta}. \quad (10)$$

See also the discussion of these two components in Flores and Flores-Lagunes (2009), who derive an equivalent decomposition in their equation (2).

Imai, Keele, and Yamamoto (2010) provide a sequential ignorability assumption under which $E[Y(1, X(1))]$, $E[Y(0, X(0))]$, and $E[Y(1, X(0))]$ are identified.[4] Identifying the latter parameter

[4] Equivalent or closely related assumptions have been considered in Pearl (2001), Flores and Flores-Lagunes (2009), Imai, Keele, and Yamamoto (2010), and Tchetgen Tchetgen and Shpitser (2011), among others.

is particularly challenging because $Y(1, X(0))$ is never observed.

Assumption 1 (sequential ignorability):

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

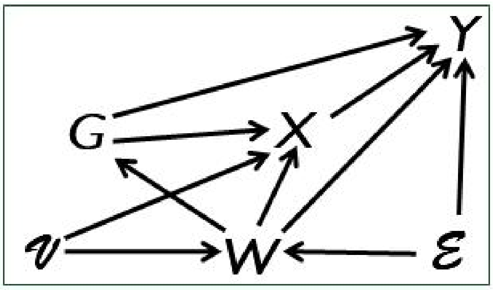
(b) $Y(g', x) \perp X | G = g, W = w$ for all $g', g \in \{0, 1\}$ and x, w in the support of X, W,

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

By Assumption 1(a), G is conditionally independent of the mediators and of any unobservable factors jointly affecting the group status on the one hand and the mediator and/or the outcome on the other hand, given the observed confounders W. Assumption 1(b) imposes conditional independence of the mediator given the confounders and the group status. That is, conditional on G and W, the effect of the mediator on the outcome is assumed to be unconfounded. Assumption 1(c) is a common support restriction requiring that the conditional probability to be in a particular group given X and W (henceforth referred to as propensity score) is larger than zero. Note that by Bayes' theorem, Assumption 1(c) equivalently implies that $\Pr(X = x | G = g, W = w) > 0$ (or in the case of X being continuous, that the conditional density of X given G and W is larger than zero: $f_{X|G,W}(x, g, w) > 0$). That is, conditional on W, the mediator state must not be a deterministic function of the group, otherwise identification is infeasible due to a lack of comparable units in terms of X across groups. Furthermore, from Assumption 1(c) it also follows that $\Pr(G = g | W = w) > 0$ for all $g \in \{0, 1\}$ and w in the support of W, which must hold to find comparable units in terms of the confounders W. Figure 3 displays a causal framework that is in line with sequential ignorability.

The following linear model satisfies the sequential ignorability assumption if ν is conditionally independent of G given W and ϵ is conditionally independent of G and X given W. Note that this parametric specification does not require the common support assumption 1(c) due to the possibility to extrapolate beyond observed data, which, however, comes at the cost of considerably

Figure 3: A causal framework in which sequential ignorability holds



stronger functional form assumptions than needed for non-parametric identification:

$$X_k = c_{X_k} + G\alpha_k + \sum_{j=1}^{J} W_j \omega_{jk} + \nu_k, \quad \text{for } k \in \{1, ..., K\} \tag{11}$$

$$Y = c_Y + G\delta + \sum_{k=1}^{K} X_k \theta_k + \sum_{k=1}^{K} G X_k \gamma_k + \sum_{j=1}^{J} W_j \lambda_{jk} + \epsilon, \tag{12}$$

Given that the conditional independence assumptions are satisfied, we can in principle allow for a much more flexible non-parametric model when identifying the explained and unexplained components:

$$X_k = \chi_k(G, W, \nu), \quad \text{for } k \in \{1, ..., K\}. \tag{13}$$

$$Y = \phi(G, X, W, \epsilon), \tag{14}$$

ϕ and χ are general functions that remain unspecified by the researcher, so that linearity and homogenous effects within group states need not be assumed as it is the case in the standard decomposition. Note that the potential outcome notation can be readily translated into this non-parametric model: $E[Y(g, X(g'))] = E[\phi(g, X(g'), W, \epsilon)] = E[\phi(g, \chi_k(g', W, \nu), W, \epsilon)]$ for g, $g' \in \{1, 0\}$.

Under Assumption 1, Propositions 1 and 2 in Huber (2013) for the identification of direct and indirect effects can be used to non-parametrically identify the explained and unexplained components based on reweighting observations by the inverse of the propensity scores $\Pr(G = 1|X,W)$ and $\Pr(G = 1|W)$:

$$\begin{aligned}\psi &= E\left[\frac{Y\cdot G}{\Pr(G=1|W)}\right] - E\left[\frac{Y\cdot G}{\Pr(G=1|X,W)}\cdot\frac{1-\Pr(G=1|X,W)}{1-\Pr(G=1|W)}\right], && (15)\\ \eta &= E\left[\frac{Y\cdot G}{\Pr(G=1|X,W)}\cdot\frac{1-\Pr(G=1|X,W)}{1-\Pr(G=1|W)}\right] - E\left[\frac{Y\cdot(1-G)}{1-\Pr(G=1|W)}\right]. && (16)\end{aligned}$$

The proof of this result is provided in Appendix A.1. $E\left[\frac{Y\cdot G}{\Pr(G=1|W)}\right]$ and $E\left[\frac{Y\cdot(1-G)}{1-\Pr(G=1|W)}\right]$ correspond to the mean potential outcomes $E(Y(1))$ and $E(Y(0))$, because weighting by the inverse of $\Pr(G = 1|W)$ and $1-\Pr(G = 1|W)$, respectively, adjusts the distribution of confounders W given $G = 1, 0$ to that in the total population. $E\left[\frac{Y\cdot G}{\Pr(G=1|X,W)}\cdot\frac{1-\Pr(G=1|X,W)}{1-\Pr(G=1|W)}\right]$, on the other hand, identifies $E[Y(1, X(0))]$, because the law of iterated expectations and Bayes' theorem imply that $\frac{1-\Pr(G=1|X,W)}{\Pr(G=1|X,W)\cdot(1-\Pr(G=1|W))}$ reweights W, X given $G = 1$ to the distribution of $W, X(0)$ in the total population.

The attractiveness of expressions (15) and (16) is that they can be straightforwardly estimated by their sample counterparts with (parametric or non-parametric) plug-in estimators for the propensity scores $\Pr(G = 1|X,W)$ and $\Pr(G = 1|W)$. However, estimation based on weighting also has its drawbacks: If the common support assumption 1(c) is close to being violated, estimation may be unstable due to an explosion of the variance, see for instance Khan and Tamer (2010). Furthermore, weighting may be less robust to propensity score misspecification than other classes of estimators, as documented for instance in Kang and Schafer (2007) and Waernbaum (2012). Alternative non- or semi-paratemetric estimators of direct and indirect effects include the regression estimator of Imai, Keele, and Yamamoto (2010) and 'multiply robust' estimation either based on the efficient influence function, see Tchetgen Tchetgen and Shpitser (2011), or on targeted maximum likelihood, see Zheng and van der Laan (2012).

Note that if G and X were not confounded by W, Assumption 1(a) could be replaced by $\{Y(g', x), X(g)\}\perp G$ (random assignment of G) and 1(b) by $Y(g', x)\perp X|G = g$. This would correspond to a non-parametric version of the model described in (6) and (7). It is easy to show that in this case the identification results (15) and (16) simplify to

$$\psi = E\left[\frac{Y\cdot G}{\Pr(G=1)}\right] - E\left[\frac{Y\cdot G}{\Pr(G=1|X)}\cdot\frac{1-\Pr(G=1|X)}{1-\Pr(G=1)}\right], \tag{17}$$

$$\eta = E\left[\frac{Y\cdot G}{\Pr(G=1|X)}\cdot\frac{1-\Pr(G=1|X)}{1-\Pr(G=1)}\right] - E\left[\frac{Y\cdot(1-G)}{1-\Pr(G=1)}\right]. \tag{18}$$

Interestingly, (18) looks identical to the identification of the average treatment effect on the non-treated in Hirano, Imbens, and Ridder (2003) based on weighting, however, under the conceptually different framework of a conditionally independent treatment given observed pre-treatment (or pre-group) variables X. The crucial difference is that in Hirano, Imbens, and Ridder (2003), conditioning on X controls for selection into the treatment (or group), whereas here, with X being a post-group mediator, conditioning on X controls for the indirect effect via X in order to identify the unexplained component. Obviously, this is only feasible if G and X are not confounded. In this case, any non-parametric method for the average treatment effect on the non-treated under conditionally independent treatment assignment consistently estimates η, including matching, see Rubin (1974), and doubly robust estimation, see Rothe and Firpo (2013). However, ruling out any confounding appears restrictive in empirical applications.

Before concluding this section, it has to be pointed out that even though Assumption 1 improves on the identifying assumptions implied by the standard decomposition by permitting observed confounders of G and X, it is still quite restrictive. In particular, it does not allow for post-group confounders of X which are affected by G, which would make identification more cumbersome. As discussed in the empirical application in Section 4, it may, however, appear plausible that some variables influencing both the mediators and the outcome are themselves influenced by the group state (e.g. the development of personality traits as a function of gender/ethnicity). We therefore briefly review alternative identification strategies suggested in Robins and Richardson (2010), Albert and Nelson (2011), Tchetgen Tchetgen and VanderWeele (2012), Imai and Yamamoto (2013), and Huber (2013) which do allow for post-group confounders that are a function of G based on various sets of assumptions. A crucial difference to Assumption 1 is that in addition to particular sequential ignorability assumptions, further restrictions on the group-mediator, confounder-mediator, or mediator-outcome relation are required for identification. Imai and Yamamoto (2013), for instance, assume that the individual group-mediator-interaction effects are homogeneous (i.e. the same for all individuals with the same values in X). Assumption 5 in Huber (2013) restricts the average group-mediator-interaction effects to be homogeneous conditional on the pre- and post-group confounders and assumes the outcome to be linear in the mediators. Alternatively, Albert and Nelson (2011) and Robins and Richardson (2010) acknowledge that the parameters of interest are identified if the potential values of the post-group confounders under $G = 1$ and $G = 0$ are either independent of each other or the functional form of their dependence is known (which appears, however, unrealistic in applications). Finally, Tchetgen Tchetgen and VanderWeele (2012) show that identification is feasible (under particular ignorability conditions) if all post-group confounders are binary and monotonic in G or if there are no average interaction effects of the post-group confounders and the mediators on the outcome (which is testable in the data).[5]

[5] As elsewhere in empirical economics, it may not appear plausible that all (pre- or post-group) confounders are observed. If credible instrumental variables for the group and/or the mediators are available, they may provide an alternative source for identifying direct and indirect effects, see for instance the non-parametric approaches in Imai, Tingley, and Yamamoto (2012) and Yamamoto (2013). If no instruments are at hand and confounders are not observed, partial identification based on deriving upper and lower bounds on the direct and indirect effects might nevertheless be informative, see for instance Cai, Kuroki, Pearl, and Tian (2008) and Flores and Flores-Lagunes (2010). However, most of these instrument- and bounds-based approaches were developed for the case of a scalar mediator and may not easily generalize to the framework of multiple mediators as usually encountered in wage decompositions.

3 Simulation

This section presents a brief simulation study in which the following data generating process (DGP) is considered:

$$\begin{aligned} G &= I\{\alpha W + \xi > 0\}, \\ X &= 0.5G + \beta W + \nu, \\ Y &= G + X + GX + \beta W + \gamma XW + \epsilon, \\ W, \xi, \nu, \epsilon &\sim N(0,1), \text{ independently of each other.} \end{aligned}$$

ξ, ν, ϵ are the error terms, while W is a potential confounder. For $\beta \neq 0$, the mediator is confounded, for $\alpha, \beta \neq 0$ both the group and the mediator are confounded. Finally, if $\gamma \neq 0$, then there exists an interaction of X and W in the outcome equation. The explained component is $\psi = E[0.5(1 + GX)|G = 1] = 0.5 + E[X|G = 1] = 1$, while the unexplained component is $\eta = E[1 + (0 + 0.5G)|G = 0] = 1$. We simulate 1000 times for various choices of α, β, γ and two sample sizes ($n = 500, 2000$), considering three different estimators: the standard decomposition ("decomposition") of Blinder (1973) and Oaxaca (1973) based on (5), OLS estimation of direct and indirect effects ("mediation.ols"), which controls for the confounder W based on the linear equations (11) and (12), and finally, semi-parametric inverse probability weighting (IPW) based on the sample analogs of (15) and (16), respectively ("mediation.ipw"). The propensity scores $\Pr(G = 1|X, W)$ and $\Pr(G = 1|W)$ entering the IPW formulae are estimated by the covariate-balancing propensity score method of Imai and Ratkovic (2014), which models group assignment while at the same time optimizing the covariate balance using an empirical likelihood approach.[6]

Table 1 presents the bias, standard deviation (s.d.), and root mean squared error (RMSE) of the various estimators. For $\alpha = 0, \beta = 0, \gamma = 0$, W does neither confound X, nor G and all

[6] IPW based on probit models for the propensity scores was also considered in the simulations, but the results are not reported because it behaved very similarly to IPW based on the empirical likelihood approach of Imai and Ratkovic (2014). Note that semi-parametric IPW estimation using parametric propensity score models is $\sqrt{n}$-consistent, which can be shown in a sequential GMM framework (see Newey (1984)) in which propensity score estimation represents the first step and the IPW decomposition the second step.

Table 1: Simulations

	n=500						n=2000					
	explained comp. ψ			*unexplained comp.* η			*explained comp.* ψ			*unexplained comp.* η		
	bias	s.d.	RMSE	bias	s.d.	RMSE	bias	s.d.	RMSE	bias	s.d.	RMSE
	$\alpha = 0, \beta = 0, \gamma = 0$						$\alpha = 0, \beta = 0, \gamma = 0$					
decomposition	-0.003	0.182	0.182	0.001	0.118	0.118	0.001	0.091	0.091	-0.000	0.055	0.055
mediation.ols	-0.003	0.182	0.182	0.001	0.118	0.118	0.001	0.091	0.091	-0.000	0.055	0.055
mediation.ipw	-0.004	0.198	0.198	0.002	0.143	0.143	0.001	0.099	0.099	0.000	0.067	0.067
	$\alpha = 0, \beta = 1, \gamma = 0$						$\alpha = 0, \beta = 0.5, \gamma = 0$					
decomposition	0.246	0.319	0.403	-0.250	0.145	0.289	0.250	0.166	0.300	-0.252	0.070	0.261
mediation.ols	-0.003	0.181	0.181	0.001	0.117	0.117	0.001	0.091	0.091	-0.000	0.055	0.055
mediation.ipw	-0.008	0.211	0.211	0.006	0.167	0.167	-0.001	0.105	0.105	0.001	0.080	0.080
	$\alpha = 0.25, \beta = 1, \gamma = 0$						$\alpha = 0.5, \beta = 0.5, \gamma = 0$					
decomposition	1.208	0.320	1.250	-0.240	0.148	0.282	1.204	0.159	1.215	-0.242	0.073	0.253
mediation.ols	-0.002	0.187	0.187	0.001	0.117	0.117	-0.001	0.092	0.092	0.000	0.056	0.056
mediation.ipw	-0.010	0.231	0.231	0.014	0.187	0.187	-0.003	0.109	0.109	0.002	0.085	0.085
	$\alpha = 0.25, \beta = 1, \gamma = 0.5$						$\alpha = 0.5, \beta = 0.5, \gamma = 0.5$					
decomposition	1.449	0.388	1.500	-0.431	0.194	0.472	1.440	0.191	1.453	-0.429	0.096	0.439
mediation.ols	0.102	0.214	0.237	-0.100	0.160	0.188	0.101	0.106	0.146	-0.100	0.075	0.125
mediation.ipw	-0.011	0.222	0.222	0.016	0.172	0.173	-0.004	0.105	0.105	0.005	0.077	0.078

Note: "decomposition" refers to estimation based on (5), "mediation.ols" to OLS estimation of (11) and (12), and "mediation.ipw" to weighting based on (15) and (16), where $\Pr(G = g|X, W)$ and $\Pr(G = g|W)$ are estimated by the Imai and Ratkovic (2014) procedure.

methods are consistent and close to being unbiased. Not unexpectedly, semi-parametric IPW has a somewhat higher standard deviation than the parametric methods due to imposing weaker assumptions. Setting $\beta = 1$ in the second panel of Table 1 introduces confounding of the mediator. Now, the standard decomposition is biased and has a substantially higher RMSE than the other two methods which control for W. In the third panel, $\alpha = 0.25, \beta = 1, \gamma = 0$, which implies that also G is confounded. This further increases the bias of the standard decomposition, while the other estimators are again almost unbiased. In the fourth panel, $\alpha = 0.25, \beta = 1, \gamma = 0.5$, such that there exists an interaction of W and X. As the OLS estimator does not account for the latter (but only for W), it is somewhat biased due to this misspecification. In contrast, IPW remains close to being unbiased due to its higher flexibility w.r.t. structural form assumptions and now dominates the other methods in terms of bias and RMSE. Summing up, the simulations demonstrate the non-robustness of the standard decomposition to confounding and violations of functional form restrictions as well as the potential gains coming from the use of more flexible estimation methods that are based on less rigid identifying assumptions.

4 Empirical application

In this section, we provide an empirical illustration in which we investigate the robustness of the results when applying various decomposition methods to the ethnic wage gap among males participating in the National Longitudinal Survey of Youth 1979 (NLSY79). The NLSY79 was designed to represent the youth population in the US in 1979 and consists of individuals that were between 14 and 22 years old in that year. It was repeated annually through 1994 and every other year ever after. The survey contains very rich information on labor market relevant characteristics such as education, detailed work experience, occupational choices, and many others. Our analysis focusses on males only, for whom the literature typically finds larger ethnic wage differentials than for females, see for instance Bayard, Hellerstein, Neumark, and Troske (1999) and O´Neill and O´Neill (2005). To be precise, we confine our sample to non-Hispanic males in order to estimate the decomposition for blacks ($G = 0$) and whites ($G = 1$).

The outcome of interest (Y) is the log hourly wage in the year 2000 of the then 35 to 43 years old survey participants, which is observed for 2,571 cases or 48% of the initial non-Hispanic male sample, thereof 851 blacks and 1,720 whites. Similarly to what is standard in the literature, our variables X characterizing the explained component include age, education, labor market history, tenure with the current employer, industry, type of occupation, whether living in an urban area, and region. Table 2 provides descriptive statistics on these variables, namely the respective mean values for blacks and whites, the mean differences, and the t-values (t-val) and p-values (p-val) based on two-sample t-tests. We see that the groups differ significantly in terms of labor market experience, job characteristics, and region. As mentioned before, conventional decompositions would merely make use of X without conditioning on potential confounders of X and/or G. Here, we also consider controlling for pre-group variables W that reflect family background and could potentially confound the group variable and the mediators. To be specific, W contains mother's and father's levels of education as well as dummies indicating whether the respondent or her/his mother were born in a foreign country.

Table 2: Means and mean differences in mediators across blacks and whites

	blacks ($G = 0$)	whites ($G = 1$)	difference	t-val	p-val
age	20.054	20.053	0.001	-0.007	0.994
education	12.910	13.609	-0.700	7.479	0.000
weeks ever worked	87.765	95.258	-7.493	4.843	0.000
tenure at primary employer in days	233.174	313.003	-79.829	7.155	0.000
industry: mining	0.005	0.008	-0.003	1.077	0.282
industry: construction	0.099	0.120	-0.021	1.635	0.102
industry: manufacturing	0.213	0.234	-0.022	1.245	0.213
industry: transport	0.143	0.098	0.045	-3.222	0.001
industry: sales	0.129	0.128	0.001	-0.096	0.923
industry: finance	0.022	0.050	-0.028	3.790	0.000
industry: business	0.092	0.084	0.007	-0.615	0.538
industry: personal services	0.021	0.010	0.011	-1.939	0.053
industry: entertainment	0.007	0.018	-0.011	2.549	0.011
industry: professional services	0.109	0.112	-0.002	0.179	0.858
industry: public admin.	0.067	0.052	0.015	-1.448	0.148
occupation: manager, professionals	0.161	0.335	-0.174	10.238	0.000
occupation: support services	0.143	0.170	-0.027	1.792	0.073
occupation: farming	0.025	0.029	-0.004	0.657	0.511
occupation: production	0.170	0.223	-0.052	3.200	0.001
occupation: operators	0.340	0.169	0.170	-9.167	0.000
in urban area	0.736	0.583	0.152	-7.924	0.000
region: Northeast	0.127	0.157	-0.030	2.088	0.037
region: West	0.082	0.152	-0.070	5.473	0.000

Table 3 provides the estimates of the total wage gap as well as the explained and unexplained components using four different decomposition methods along with standard errors (s.e.) and p-values (p-val) based on 999 bootstrap replications. It also gives the explained and unexplained components as percentages of the total gap (% of total) to assess their relative importance. The total gap in average log wages of blacks and whites is 0.434 log points. The standard decomposition of Blinder (1973) and Oaxaca (1973) ("decomposition") suggests that 0.299 log points or 68.9% of the total gap are explained by differences in X while 0.135 log points or 31.1% are unexplained. Both estimates are highly significant. We also consider the in terms of functional form assumptions more flexible IPW estimator ("ipw") based on (17) and (18), which, however, does not control for W either. As in the simulations, $\Pr(G = 1|X)$ is estimated by the empirical likelihood approach of Imai and Ratkovic (2014). The propensity score specification is provided in Appendix A.2. As this semi-parametric method does not extrapolate to sparse data regions, its applicability hinges on a sufficiently high overlap of the propensity scores across groups, also known as common support. Appendix A.3 provides density estimates of the within-group propensity scores using the logspline command in R. The common support appears decent, so that for any individual comparable observations w.r.t. the propensity score can be found in the respective other group. We see that despite the more flexible functional form assumptions, the results based on weighting (without controlling for W) are very similar to those of the standard decomposition.

Table 3: Black-White wage gap decomposition based on NLSY79

	total gap in log wages			*explained component* $\hat{\psi}$				*unexplained component* $\hat{\eta}$			
	estimate	s.e.	p-val	estimate	s.e.	p-val	% of total	estimate	s.e.	p-val	% of total
decomposition	0.434	0.040	0.000	0.299	0.035	0.000	68.9	0.135	0.047	0.004	31.1
ipw	0.434	0.044	0.000	0.306	0.041	0.000	70.4	0.128	0.057	0.024	29.6
mediation.ols	0.288	0.048	0.000	0.176	0.033	0.000	61.3	0.112	0.053	0.036	38.7
mediation.ipw	0.280	0.046	0.000	0.171	0.060	0.004	60.9	0.109	0.069	0.114	39.1

Note: "decomposition" refers to estimation based on (5), "ipw" to weighting based on (17) and (18) not controlling for W, "mediation.ols" to OLS estimation of (11) and (12) which controls for W, and "mediation.ipw" to weighting based on (15) and (16) which also controls for W. All standard errors (s.e.) are based on 999 bootstrap replications.

We also consider two estimators controlling for family background characteristics W, namely OLS estimation of direct and indirect effects ("mediation.ols") based on equations (11) and (12), and IPW ("mediation.ipw") based on (15) and (16). The propensity score specification of $\Pr(G = 1|X, W)$ is provided in Appendix A.2 and the common support, which is again quite satisfactory, is displayed in Appendix A.3.[7] Note that the total gap now corresponds to the difference in log wages of blacks and whites after making both groups comparable in W. It is considerably smaller than the (initial) wage gap ignoring confounding by W (0.434), no matter whether estimation is based on OLS (0.288) or IPW (0.280). This suggests that the total impact of ethnicity on wages is overestimated when not controlling for family background. The same applies to the magnitudes of explained components, which drop substantially when controlling for W (from roughly 0.3 to

[7] We also investigated whether the Imai and Ratkovic (2014)-based estimate of $\Pr(G = 1|X, W)$ does indeed optimally balance the distributions of X, W across groups as one would expect from empirical likelihood estimation. In the original data prior to IPW weighting, the mean and maximum absolute differences in X, W across groups amount to 1.698 and 79.828, respectively. The standardized mean and maximum absolute differences (standardized by the standard deviations of the respective variables) are 0.142 and 0.592. After IPW weighting, any of the measures is 0.000, so that X, W have been balanced successfully.

less than 0.18 log points). The unexplained components, in contrast, decrease only moderately to 0.11 log points. Therefore, their relative importance measured as percentage of the total gap increases somewhat to 39%. As before, the OLS and IPW estimates are very similar, with the latter being somewhat nosier. In particular, the unexplained component based on IPW is not significantly different from zero at the 10% level, while that based on OLS is significant at the 5% level.

Our results suggest that the decomposition estimates are not robust to ignoring potential pre-group confounders, but may entail overestimation of the absolute values of the explained and unexplained components as well as the total wage gap. However, even the estimates based on controlling for W should not be taken at face value. Firstly, in addition to the family background variables considered, there might exist further pre-group confounders that were omitted in the estimation. Potential examples include parents' personality traits, parents' health, and neighborhood conditions, which might differ systemically across G or X and at the same time affect labor market success later in life, entailing a violation of Assumption 1. Secondly, as a further issue not considered in our analysis, there may also exist post-group confounders of X that are themselves potentially affected by G. This case is ruled out in Assumption 1, but clearly a plausible scenario, given that the mediators are measured with a substantial time lag after the determination of G (at birth).

As an example, consider the possibility that G affects the development of personality traits (e.g. through a systematically different exposure to discrimination, peer groups, etc. while growing up), which themselves influence X (e.g. schooling and occupation) as well as potential wages. Further possible sources of post-group confounding are attrition and selection into employment, as roughly half of the initial sample had dropped out of the survey by 2000 or did not report any employment.[8] Therefore, controlling for pre-birth confounders does likely not suffice for tackling mediator endogeneity. At the very least, our analysis and the sensitivity of the results to different sets of assumptions demonstrate the importance of considering the problem of confounding in

[8] We refer to MaCurdy, Mroz, and Gritz (1998) for a detailed analysis of the attrition patterns in the NLSY79.

wage decompositions, an issue apparently neglected in much of the literature. Future research may discuss the implementation and attractiveness of alternative identification strategies that also allow for post-group confounders (see for instance the work of Robins and Richardson (2010), Albert and Nelson (2011), Imai and Yamamoto (2013), Huber (2013), and Tchetgen Tchetgen and VanderWeele (2012) mentioned in Section 2.3) in the context of wage decompositions.

5 Conclusion

This paper has made explicit the identifying assumptions underlying conventional decompositions of gender or ethnic wage gaps, which continue to receive much attention in the empirical labor literature, by translating the decomposition into an equivalent model for mediation analysis. Inspecting the latter immediately shows that conventional decompositions do not control for confounders of the group variable (e.g. gender or ethnicity) and/or the variables characterizing the explained component (e.g. education and occupation), if the latter are determined after the group variable. This appears to be the standard case, as gender or ethnicity are determined at or prior to birth and therefore precede mediators like education or profession. For this reason, we have suggested the use of an alternative set of identifying assumptions that assumes exogeneity of the group variable and the mediators only conditional on observed confounders. Then, the unexplained and explained components of the decomposition can be non-parametrically identified by using a simple weighting expression that reweights observations by the inverse of the conditional propensity to belong to a particular group given the mediators and confounders. Finally, we have provided a simulation study as well as an empirical application to data from the National Longitudinal Survey of Youth 1979, in which we have also pointed to approaches permitting to further relax our identifying assumptions.

A Appendix

A.1 Proof of equations (15) and (16) under Assumption 1

The following proof is closely related to Huber (2013). To prove equations (15) and (16) under Assumption 1, we need to show that $E[Y(g, X(g))]$ for $g \in \{1, 0\}$ and $E[Y(1, X(0))]$ are identified. Starting with the former, note that

$$\begin{aligned} E[Y(g, X(g))] &= \int E[Y(g, X(g))|W = w] dF_W(w) \\ &= \int E[Y|G = g, W = w] dF_W(w) \\ &= E\left[E\left[\frac{Y \cdot I\{G = g\}}{\Pr(G = g|W)}\middle| W = w\right]\right] \\ &= E\left[\frac{Y \cdot I\{G = g\}}{\Pr(G = g|W)}\right]. \end{aligned} \tag{A.1}$$

The first equality follows from the law of iterated expectations and from replacing the expectation by an integral, the second from Assumption 1(a), the third from basic probability theory and from replacing the integral by the expectation, and the last from the law of iterated expectations.

Concerning the latter,

$$\begin{aligned} & E[Y(1, X(0))] \\ = & \int\int E[Y(1, x)|X(0) = x, W = w] dF_{X(0)|W=w}(x) dF_W(w) \\ = & \int\int E[Y(1, x)|G = 1, X = x, W = w] dF_{X|G=0,W=w}(x) dF_W(w) \\ = & \int\int E[Y|G = 1, X = x, W = w] \cdot \frac{\Pr(G = 0|X, W)}{\Pr(G = 0|W)} dF_{X|W=w}(x) dF_W(w) \\ = & E\left[E\left[E\left[\frac{Y \cdot G}{\Pr(G = 1|X, W)}\middle| X = x, W = w\right] \cdot \frac{1 - \Pr(G = 1|X, W)}{1 - \Pr(G = 1|W)}\middle| W = w\right]\right] \\ = & E\left[\frac{Y \cdot G}{\Pr(G = 1|X, W)} \cdot \frac{1 - \Pr(G = 1|X, W)}{1 - \Pr(G = 1|W)}\right]. \end{aligned} \tag{A.2}$$

The first equality follows from the law of iterated expectations and from replacing the outer expectations by integrals, the second from Assumptions 1(a) and 1(b), the third from Bayes' theorem, the fourth from basic probability theory and from replacing the integrals by expectations, and the last from the law of iterated expectations.

Therefore, the explained component $\psi = E[Y(1, X(1))] - E[Y(1, X(0))]$ is equal to

$$E\left[\frac{Y \cdot G}{\Pr(G = 1|W)}\right] - E\left[\frac{Y \cdot G}{\Pr(G = 1|X, W)} \cdot \frac{1 - \Pr(G = 1|X, W)}{1 - \Pr(G = 1|W)}\right], \tag{A.3}$$

while the unexplained component $\eta = E[Y(1, X(0))] - E[Y(0, X(0))]$ is equal to

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

A.2 Propensity score specifications

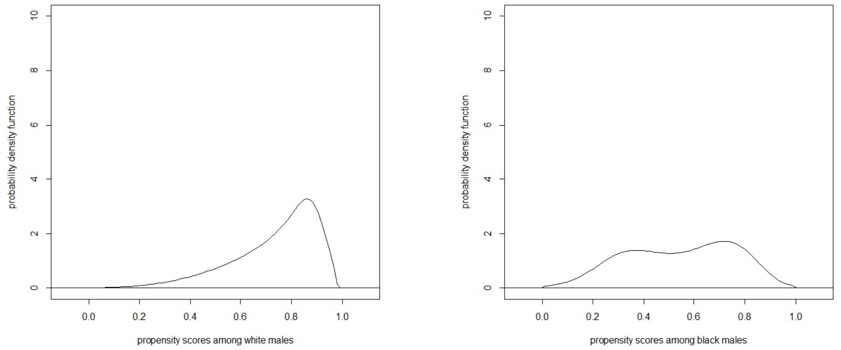
Table 4: Propensity score specifications based on the Imai and Ratkovic (2014) procedure

	$\Pr(G=1\|X,W)$				$\Pr(G=1\|X)$			
	coef.	s.e.	z-val	p-val	coef.	s.e.	z-val	p-val
X								
age	-0.047	0.019	-2.418	0.016	-0.051	0.023	-2.245	0.025
education: 12 yrs	-0.421	0.289	-1.455	0.146	-0.061	0.194	-0.314	0.754
education: 13 yrs	-0.815	0.373	-2.185	0.029	-0.074	0.329	-0.224	0.823
education: 14 yrs	-0.500	0.419	-1.192	0.233	0.170	0.370	0.460	0.645
education: 15 yrs	-1.349	0.431	-3.133	0.002	-0.588	0.376	-1.564	0.118
education: 16 yrs	-0.634	0.375	-1.689	0.091	0.329	0.226	1.458	0.145
education: 17 yrs	-0.820	0.544	-1.508	0.131	0.195	0.565	0.344	0.731
education: 18 yrs or more	-0.321	0.447	-0.717	0.473	0.767	0.353	2.174	0.030
weeks ever worked	0.005	0.002	2.546	0.011	0.003	0.003	1.040	0.298
weeks ever worked missing	-0.687	0.533	-1.289	0.197	-0.599	0.442	-1.356	0.175
years ever worked	0.122	0.018	6.947	0.000	0.137	0.017	8.211	0.000
tenure at primary employer	0.000	0.000	1.155	0.248	0.000	0.000	1.875	0.061
tenure missing	0.508	0.402	1.263	0.207	0.367	0.389	0.943	0.346
worked before	0.055	0.176	0.313	0.754	0.147	0.223	0.660	0.509
industry: mining	0.033	0.691	0.048	0.961	-0.275	0.524	-0.525	0.599
industry: construction	-0.166	0.223	-0.746	0.456	-0.254	0.229	-1.106	0.269
industry: manufact. durables	-0.488	0.253	-1.930	0.054	-0.586	0.251	-2.339	0.019
industry: manufact. non-dur.	-0.310	0.203	-1.523	0.128	-0.285	0.229	-1.244	0.213
industry: transport	-0.829	0.201	-4.130	0.000	-0.907	0.190	-4.780	0.000
industry: wholesale	-0.468	0.540	-0.866	0.386	-0.197	0.398	-0.494	0.621
industry: retail	-0.479	0.204	-2.347	0.019	-0.579	0.247	-2.340	0.019
industry: finance	-0.171	0.389	-0.440	0.660	-0.293	0.274	-1.068	0.286
industry: business	-0.399	0.211	-1.893	0.058	-0.400	0.210	-1.904	0.057
industry: personal services	-0.891	0.436	-2.044	0.041	-1.009	0.547	-1.843	0.065
industry: entertainment	1.112	0.523	2.127	0.033	0.651	0.402	1.618	0.106
industry: professional services	-0.651	0.256	-2.542	0.011	-0.816	0.222	-3.671	0.000
industry: public admin.	-0.757	0.382	-1.981	0.048	-0.831	0.488	-1.703	0.089
profession: manager, professionals	1.102	0.217	5.074	0.000	1.025	0.222	4.623	0.000
profession: support services	0.649	0.228	2.842	0.004	0.739	0.248	2.981	0.003
profession: farming	0.580	0.415	1.398	0.162	0.498	0.402	1.238	0.216
profession: production	0.834	0.226	3.686	0.000	0.768	0.213	3.603	0.000
profession: operators	0.043	0.217	0.199	0.842	-0.077	0.207	-0.372	0.710
in urban area	-0.992	0.141	-7.013	0.000	-0.863	0.152	-5.665	0.000
urban area missing	0.301	1.105	0.273	0.785	0.557	0.701	0.794	0.427
region: Northeast	0.282	0.152	1.856	0.063	0.363	0.248	1.466	0.143
region: West	0.776	0.196	3.959	0.000	0.974	0.255	3.817	0.000
W								
foreign born	-0.463	0.555	-0.834	0.405				
mother foreign born	0.910	0.392	2.323	0.020				
mothers educ.: 12	0.652	0.181	3.601	0.000				
mothers educ.: 13	0.469	0.456	1.029	0.304				
mothers educ.: 14	0.554	0.478	1.159	0.247				
mothers educ.: 15	0.470	0.515	0.912	0.362				
mothers educ.: 16	-0.053	0.379	-0.141	0.888				
mothers educ.: 17	0.226	0.901	0.251	0.802				
mothers educ.: 18 or more	-0.498	0.685	-0.728	0.467				
mothers educ. missing	-0.155	0.372	-0.416	0.678				
fathers educ.: 12	0.414	0.216	1.918	0.055				
fathers educ.: 13	1.701	0.490	3.472	0.001				
fathers educ.: 14	0.992	0.427	2.324	0.020				
fathers educ.: 15	1.372	0.531	2.584	0.010				
fathers educ.: 16	1.516	0.301	5.041	0.000				
fathers educ.: 17	0.519	0.522	0.995	0.320				
fathers educ.: 18 or more	2.132	0.403	5.294	0.000				
fathers educ. missing	-1.042	0.226	-4.617	0.000				
constant	-0.410	0.419	-0.977	0.328	-0.356	0.642	-0.555	0.579

Note: The variables in X are measured in 1998 and those in W in 1979.

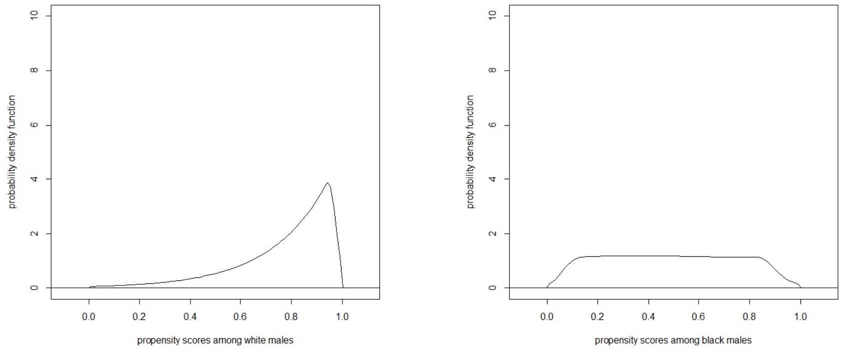
A.3 Propensity score distributions

Figure 4: Density estimates of the estimates of $\Pr(G = 1|X)$ across groups



Note: Density estimation is based on splines using the (default setting of the) logspline command for R. The lower and upper bounds of the support of the propensity scores are set to 0 and 1.

Figure 5: Density estimates of the estimates of $\Pr(G = 1|X, W)$ across groups



Note: Density estimation is based on splines using the (default setting of the) logspline command for R. The lower and upper bounds of the support of the propensity scores are set to 0 and 1.

References

Albert, J. M., and S. Nelson (2011): "Generalized causal mediation analysis," *Biometrics*, 67, 1028–1038.

Baron, R. M., and D. A. Kenny (1986): "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology*, 51, 1173–1182.

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.

Bayard, K., J. Hellerstein, D. Neumark, and K. Troske (1999): "Why are racial and ethnic wage gaps larger for men than for women? Exploring the role of segregation using the new worker-establishment characteristic database," *National Bureau of Economic Research, Working Paper 6997.*

Bertrand, M., and S. Mullainathan (2004): "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *The American Economic Review*, 94, 991–1013.

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

Cai, Z., M. Kuroki, J. Pearl, and J. Tian (2008): "Bounds on Direct Effects in the Presence of Confounded Intermediate Variables," *Biometrics*, 64, 695–701.

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

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.

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.

Flores, C. A., and A. Flores-Lagunes (2009): "Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness," *IZA DP No. 4237.*

——— (2010): "Nonparametric Partial Identification of Causal Net and Mechanism Average Treatment Effects," *mimeo, University of Florida.*

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 assumption-with an application to the gender wage gap in the United Kingdom," *Econometrics Journal*, 10, 359–407.

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. J., J. Stixrud, and S. Urzua (2006): "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior," *NBER Working Paper No. 12006.*

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. (2013): "Identifying causal mechanisms (primarily) based on inverse probability weighting," *forthcoming in the Journal of Applied Econometrics.*

Imai, K., L. Keele, and T. Yamamoto (2010): "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71.

Imai, K., and M. Ratkovic (2014): "Covariate balancing propensity score," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 243–263.

Imai, K., D. Tingley, and T. Yamamoto (2012): "Experimental Designs for Identifying Causal Mechanisms," *forthcoming in the Journal of the Royal Statistical Society, Series A.*

Imai, K., and T. Yamamoto (2013): "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments," *Political Analysis*, 21, 141–171.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29.

Judd, C. M., and D. A. Kenny (1981): "Process Analysis: Estimating Mediation in Treatment Evaluations," *Evaluation Review*, 5, 602–619.

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

Kang, J. D. Y., and J. L. Schafer (2007): "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, 22, 523–539.

Khan, S., and E. Tamer (2010): "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78, 2021–2042.

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

Lam, D., and R. F. Schoeni (1993): "Effects of Family Background on Earnings and Returns to Schooling: Evidence from Brazil," *Journal of Political Economy*, 101, 710–740.

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

MaCurdy, T., T. Mroz, and R. M. Gritz (1998): "An Evaluation of the National Longitudinal Survey on Youth," *The Journal of Human Resources*, 33, 345–436.

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.

Newey, W. K. (1984): "A method of moments interpretation of sequential estimators," *Economics Letters*, 14, 201–206.

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

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

O´Neill, J. E., and D. M. O´Neill (2005): "What Do Wage Differentials Tell Us about Labor Market Discrimination?," *National Bureau of Economic Research, Working Paper 11240.*

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. (2003): "Semantics of causal DAG models and the identification of direct and indirect effects," in *In Highly Structured Stochastic Systems*, ed. by P. Green, N. Hjort, and S. Richardson, pp. 70–81, Oxford. Oxford University Press.

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

Robins, J. M., and T. Richardson (2010): "Alternative graphical causal models and the identification of direct effects," in *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, ed. by P. Shrout, K. Keyes, and K. Omstein. Oxford University Press.

Rosenbaum, P. (1984): "The consequences of adjustment for a concomitant variable that has been affected by the treatment," *Journal of Royal Statistical Society, Series A*, 147, 656–666.

Rosenbaum, P., and D. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Rothe, C., and S. Firpo (2013): "Semiparametric Estimation and Inference Using Doubly Robust Moment Conditions," *IZA Discussion Paper No. 7564.*

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

Tchetgen Tchetgen, E. J., and I. Shpitser (2011): "Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis," *Harvard University Biostatistics Working Paper 130.*

Tchetgen Tchetgen, E. J., and T. J. VanderWeele (2012): "On Identification of Natural Direct Effects when a Confounder of the Mediator is Directly Affected by Exposure," *forthcoming in Epidemiology.*

WAERNBAUM, I. (2012): "Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation," *Statistics in Medicine*, 31, 1572–1581.

YAMAMOTO, T. (2013): "Identification and Estimation of Causal Mediation Effects with Treatment Noncompliance," *unpublished manuscript, MIT Department of Political Science.*

ZHENG, W., AND M. J. VAN DER LAAN (2012): "Targeted Maximum Likelihood Estimation of Natural Direct Effects," *The International Journal of Biostatistics*, 8, 1–40, Article 3.