

Research Article

Christiane Didden*

Targeting mediating mechanisms of social disparities with an interventional effects framework, applied to the gender pay gap in Western Germany

<https://doi.org/10.1515/jci-2023-0015>

received March 28, 2023; accepted November 20, 2024

Abstract: The Oaxaca-Blinder (OB) decomposition is a widely used method to explain social disparities. However, assigning causal meaning to its estimated components requires strong assumptions that often lack explicit justification. This article emphasizes the importance of clearly defined estimands and their identification when targeting mediating mechanisms of social disparities. Three approaches are distinguished on the basis of their scientific questions and assumptions: a mediation approach and two interventional approaches. The OB decomposition and Monte Carlo simulation-based g-computation are discussed for estimation in relation to these approaches. The latter method is used in an interventional effects analysis of the observed gender pay gap in Western Germany, using data from the 2017 German Socio-Economic Panel. Ten mediators are considered, including indicators of human capital and job characteristics. Key findings indicate that the gender pay gap in log hourly wages could be reduced by up to 86% if these mediators were equally distributed between women and men. Substantial reductions could be achieved by aligning full-time employment and work experience.

Keywords: mediation, interventional effects, gender pay gap, social inequality

MSC 2020: 62P25

1 Traditional and modern approaches to decomposing social disparities

Social disparities exist in various domains, including wealth, education, the labor market, health, and political participation [1]. A prominent example is the gender pay gap, which typically indicates that women, on average, earn less than men. Gender differences in wage-relevant factors, such as education, career paths, work experience, and working hours, are typically used to explain this phenomenon [2–6]. A common method to evaluate the contribution of these explanatory factors to wage disparities between two groups, also applied by official statistical offices [7], is the Oaxaca-Blinder (OB) decomposition method [8,9]. Central to this method is the twofold decomposition for linear outcome models. A linear regression model with wage as the outcome is fitted for each group with the explanatory factors as regressors. Based on these models, the marginal wage disparity is then decomposed into two components: the “explained” part, which relates to the differences between the two groups in the means of the explanatory variables, and the “unexplained” part, which relates to the differences between the two groups in the associations between the explanatory variables and wage.

* **Corresponding author: Christiane Didden**, Department of Sociology, LMU Munich, Munich, Germany,
e-mail: Christiane.Didden@soziologie.uni-muenchen.de

This relatively straightforward decomposition, implemented in common software packages [10,11], is not restricted to wage gaps, but can be applied to various outcome disparities (see, e.g., [12]).

Decomposition analyses that aim to provide explanations for disparities beyond mere statistical descriptions suggest a focus on causal relationships. Petersen and van der Laan [13] and Lundberg et al. [14] emphasized the importance of guiding a causal analysis with a clearly defined target parameter. The steps involved in a causal analysis using observational data, as outlined by Petersen and van der Laan [13], can be categorized into three stages: the conceptual, the statistical, and the interpretation stage. The conceptual stage involves several key steps. First, a causal model is established to summarize assumptions about the data-generating process. The observed data are linked to this model, followed by a formulation of a precise research question. This question is then translated into a formal target parameter, known as the causal estimand. Finally, the identifiability of the parameter is evaluated; if it is identified, it can be expressed as a function of the observed data. The subsequent statistical stage involves specifying the statistical estimation problem, selecting the estimation method, and performing the estimation. The final stage focuses on selecting an appropriate interpretation of the estimated parameter. According to this roadmap, the conceptual stage underpins the analysis and guides the subsequent steps, not the formulation of a statistical model. However, researchers have noted that despite this crucial role, conceptual foundations including the definition of a causal estimand are often neglected in social science applications [14], particularly in decomposition analyses, including the OB decomposition [15,16]. This article addresses this concern by explicitly emphasizing conceptual aspects in research targeting mediating mechanisms of social disparities, incorporating recent methodological findings [17–22]. Central to this article is a nonparametric causal model adaptable to a wide range of social contexts. It includes a time-fixed exposure (or treatment), a single outcome, multiple interdependent mediators, and different types of confounders (see directed acyclic graph [DAG] in Figure 1 (a)). The exposure variable pertains to a social characteristic that may be ascribed, such as gender or race, or may be acquired over the course of an

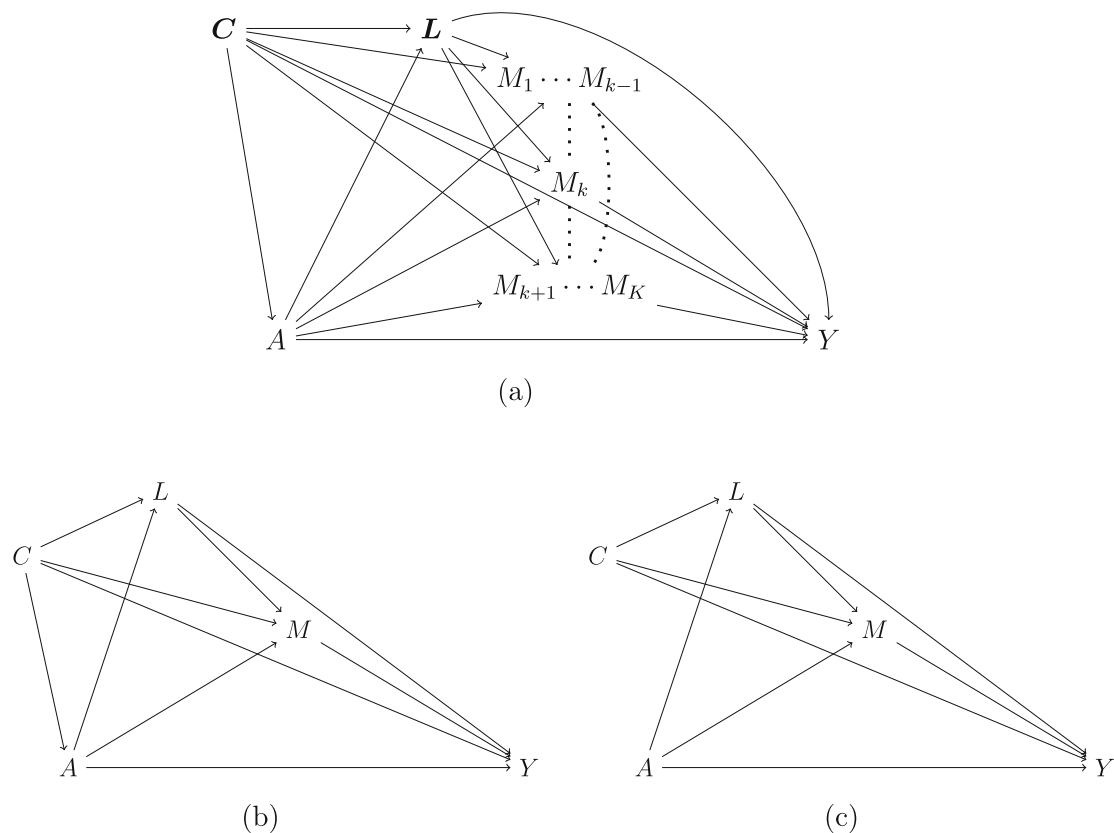


Figure 1: Structural assumptions: (a) with multiple mediators $M_1, \dots, M_K$, (b) and (c) with a single mediator M . Dotted lines among mediators indicate unknown structural dependencies among them.

individual's lifetime, such as educational attainment or union membership. It is assumed to affect the outcome both directly and indirectly through intermediate factors, i.e., mediators. Drawing on recent classifications of effect types [23] and methodological insights [22], three approaches are distinguished. While all three approaches target mediating mechanisms of social disparities, they differ in terms of their scientific questions, estimands, and the strength of the required identification assumptions. The first approach (Approach 1) aims to conduct a causal mediation analysis to explain how the exposure affects the outcome by examining direct and indirect pathways (see refs [17,19,22], among others). The second approach (Approach 2) aims to evaluate the effects of hypothetical mediator manipulations on the exposure-induced outcome disparity [20]. The third approach (Approach 3) aims to evaluate the effects of manipulations of explanatory factors, such as potential mediators, on the actual observed outcome disparity [18,24].

Approaches 1–3 will be illustrated by nonparametrically defined causal estimands. These quantities do not depend on specific model assumptions or parameters, unlike the effects in traditional parametric methods in mediation [25,26] (see refs [27,28] for limitations of these methods). Specifically, causal effects are defined on the additive scale as contrasts between two potentially counterfactual mean outcomes: one that would be observed in response to a specific intervention and another that would be observed without this intervention (or in response to a different intervention) [29]. In this context, the term “intervention” is employed in a hypothetical manner to describe a potential manipulation of a variable, without specifying the details of how this intervention could be implemented or by whom. With regard to the mediator interventions under consideration, this concept aligns with the ill-defined type of interventions discussed in the study of Moreno-Betancur et al. [20]. As noted by Nguyen et al. [23], any contrast of outcomes under two different intervention conditions that could be conceivable for hypothetical future studies falls under the class of interventional effects. This article focuses on a specific type of interventional effect, namely, that characterized by interventions on the distribution of hypothesized mediators. This type of interventional effect has received substantial attention in methods research on mediation analysis (see, e.g., refs [17,19,22,30,31]). Applications can be found in the field of epidemiology, in particular (see, e.g., refs [32–34]).

Causal implementations of the twofold OB decomposition and the associated identification assumptions have been discussed in previous studies with different foci. Fortin et al. [16] focused on causal inferences more generally, with a particular emphasis on the interpretation of the unexplained component. Huber [15] examined the method within the context of mediation analysis. Jackson and VanderWeele [18] linked the twofold OB method to an interventional approach for observed disparities. This article contributes by examining the applicability of the twofold OB decomposition for linear outcome models as an estimator in the context of Approaches 1–3. For simplicity, causal models with a single mediator are used for this purpose. Unlike the causal diagrams in Huber [15], Jackson and VanderWeele [18], these models include a confounder of the mediator–outcome relationship that is affected by the exposure and may interact with the mediator in its effect on the outcome.

Finally, this article applies Approach 3 to the gender pay gap, using data from the 2017 German Socio-Economic Panel (SOEP) [35,36] on households located in the region of former West Germany, hereafter referred to as Western Germany. Ten potential mediators are considered, including job characteristics as well as measures of education and labor market experience, both key components of human capital [37]. The rationale for using a g-computation approach for estimation is outlined, and the steps involved in the Monte Carlo simulation-based g-computation, hereafter referred to as MC g-computation, are described.

The structure of this article is as follows: it begins with an introduction to counterfactual notation (Section 2) and a general review of interventional effects defined by stochastic mediator interventions (Section 3). Approaches 1–3 are then outlined and illustrated with specific causal estimands (Sections 3.1–3.3). Section 4 addresses the use of the twofold OB decomposition for linear outcome models in the context of Approaches 1–3. Section 5 illustrates the application of Approach 3 to the gender pay gap in Western Germany. The approach's adaptability is demonstrated through the consideration of different types of mediator interventions. The article concludes with a discussion in Section 6.

2 Counterfactual notation

Let A denote the exposure and Y the outcome. Let M denote an intermediate variable, which is assumed to mediate the effect of A on Y . Let $\mathbf{C}$ denote a set of baseline covariates, including exposure–mediator, exposure–outcome as well as mediator–outcome confounders that are not affected by the exposure. Additionally, let $\mathbf{L}$ denote a set of intermediate mediator–outcome confounders that are affected by the exposure; therefore, they may also act as mediators.

Let Y_a denote the counterfactual outcome Y under an intervention that sets the exposure to the value a . The expected counterfactual outcome under $A = a$, denoted by $E[Y_a]$, generally differs from $E[Y|A = a]$, the expected outcome at the actual exposure level a . This difference arises because $E[Y_a]$ represents what the expected outcome would be in a hypothetical scenario where the exposure value is a for all individuals in the population of interest, while $E[Y|A = a]$ represents the expected outcome among individuals who are actually exposed to a . Central to mediation are counterfactuals of the form Y_{am} , defined by interventions on the exposure and on the mediator. According to the composition assumption, Y_a can be expressed as Y_{aM_a} , which is the counterfactual outcome under $A = a$ and the mediator value naturally arising under $A = a$ [38]. Additionally, using the notation in the study of Loh et al. [39], let $\tilde{M}_{a|\mathbf{C}}$ be a random draw from the counterfactual distribution of M under $A = a$, given $\mathbf{C}$, denoted by $P_{M_{a|\mathbf{C}}}(m|\mathbf{C})$. Furthermore, let $\tilde{M}|A = a, \mathbf{C}$ denote a random draw from the distribution of M among those with exposure value a and covariates $\mathbf{C}$, given by $P_{M|A=a, \mathbf{C}}(m|A = a, \mathbf{C})$.¹ More generally, let $\tilde{M}$ denote a random draw from a user-specified distribution of M .

The focus of Sections 3.1–3.3 and Sections 4–5 is on a binary exposure that takes values in $\{0, 1\}$. In this context, $E[Y_1]$ refers to the expected counterfactual outcome *under exposure*, and $E[Y_0]$ to the expected counterfactual outcome *under control*. $E[Y|A = 1]$ refers to the expected outcome among the actual observed exposure group (the exposed), and $E[Y|A = 0]$ to the expected outcome among the actual observed control group (the unexposed).

3 Interventional effects for decomposing social disparities

Mediation is commonly understood as a phenomenon that occurs when the exposure affects a mediator M and the resulting change in M goes on to affect the outcome [22,40]. The central measure for the indirect effect through M is the natural indirect effect (NIE) through M , usually defined in the mediation literature by $\text{NIE}_M = E[Y_{aM_a} - Y_{aM_{a^*}}]^2$, with a referring to the exposure level and a^* to the control level [42]. The effect of the exposure on the outcome that is not mediated by M is known as the natural direct effect (NDE), defined as $\text{NDE}_M = E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}]$. Both the NIE_M and NDE_M are measures of the underlying mechanisms through which the exposure affects the outcome. Notably, they add up to the total effect of the exposure (TE), also known as average treatment effect, given by $\text{TE} = E[Y_{aM_a} - Y_{a^*M_{a^*}}]$ [42]. In this article, the TE is also referred to as the *exposure-induced disparity*. Natural effects are defined starting at the individual level, with average NIEs being population means of the individual (in)direct effects [23,43]. However, they involve cross-world counterfactual outcomes in the form of $Y_{aM_{a^*}}$ that are empirically inaccessible because it is not possible to bring an individual's exposure to level a and the mediator for that individual to its natural level under the counterfactual level a^* . Therefore, the identification of NIEs requires strong assumptions, including the cross-world independence assumption $Y_{am} \perp\!\!\!\perp M_{a^*}|\mathbf{C}$, ($a \neq a^*$). This assumption does often not hold, as it is known to be violated in the presence of confounders of the mediator–outcome relationship affected by the exposure, referred to as *exposure-induced confounders* [44]. Even if a well-controlled randomized experiment were feasible, it cannot eliminate exposure-induced confounding of the mediator–outcome relationship [22].

¹ $P_{M|A, \mathbf{C}}(m|A, \mathbf{C}) = P(M = m|A = a, \mathbf{C})$, and $P_{M_{a|\mathbf{C}}}(m|\mathbf{C}) = P(M_a = m|\mathbf{C})$.

² Robins and Greenland [41] differ between the “total indirect effect”, given by $E[Y_{aM_a} - Y_{aM_{a^*}}]$, and the “pure indirect effect”, given by $E[Y_{a^*M_a} - Y_{a^*M_{a^*}}]$.

In response to this problem, interventional (in)direct effects, also described as “randomized interventional analog[ues] of the notions of natural direct and indirect effects” [17], have been examined as alternatives to natural effects (see refs [17,19,22,31,45], among others). Interventional (in)direct effects are defined by weighted expectations of counterfactual outcomes, weighted according to counterfactual mediator distributions representing specific stochastic mediator interventions [43]. In particular, the interventional analogue to NIE_M , as proposed by VanderWeele et al. [17], is given by $IIE_{M|C} = E[Y_{a\tilde{M}_{a|C}} - Y_{a\tilde{M}_{a^*|C}}]$ (see the study of VanderWeele and Tchetgen Tchetgen [31] for an alternative interventional analogue³). This effect is defined by the contrast between (1) the counterfactual outcome expected under $A = a$ and the mediator being randomly drawn from its counterfactual distribution under $A = a$, given C , and (2) the counterfactual outcome expected under $A = a$ and the mediator being randomly drawn from its counterfactual distribution under $A = a^*$, given C . It follows that the contrast between these two expected counterfactual outcomes is due to a distributional shift in the mediator – from the distribution under the exposure level a to that under the level a^* , given C . These counterfactual *mediator intervention distributions* are marginal with respect to exposure-induced confounders L , ignoring the dependence between M and L . Notably, interventional (in)direct effects do not rely on cross-world independence assumptions, which is crucial in contexts with multiple interdependent mediators, where these assumptions are typically violated [31,45–47].

Nguyen et al. [23] note that discussing interventional (in)direct effects at the individual level is not meaningful because they depend on mediator values randomly drawn from counterfactual distributions in subpopulations defined by the covariate strata in C . Moreno-Betancur and Carlin [43] argue that such “population-level interventions” interventions could be conceptualized for hypothetical future trials. Therefore, interventional (in)direct effects may serve as “more natural initial target estimands for mediation” than natural (in)direct effects, which rely on infeasible “individual-level interventions”. Miles [22], however, emphasized the importance of the individual level in mediation analysis, arguing that any true effect measure for the mediated effect through M ought to take its null value when no individual-level indirect effect through M exists in the population of interest. Certain indirect effect measure criteria, namely, the sharp(er) null criterion and the monotonicity criterion, ought to be satisfied when measuring mediation and the direction of the mediated effect. To ensure that interventional indirect effects meet these criteria and can serve as substitutes for NIEs – which satisfy these criteria – restrictive conditions must hold beyond those required for identification. Otherwise, interventional indirect effects may lack true mediational meaning, potentially showing non-zero values even in the absence of individual-level indirect effects. Furthermore, they may misidentify the direction of the mediated effect, even if it were uniform (or null) across all individuals [22]. However, in the absence of mediational meaning, interventional effects – defined by interventions on hypothesized mediators – can still offer valuable insights into the extent to which social disparities could change if mediating mechanisms were altered [18,20,23].

Drawing from recent literature on mediation and interventional effects [18–20,22–24], the following sections distinguish three approaches to mediating mechanisms of social disparities. Interventional effects are integral to all three approaches. Causal mediation is the focus of Approach 1. In this context, interventional (in)direct effects are discussed as alternative effect measures when natural (in)direct effects are not identified [17,19,22,31,43,47]. Approaches 2 and 3 are classified as pure interventional effects approaches [23]. Approach 2 evaluates the impact of hypothetical mediator manipulations on the exposure-induced disparity [20]. In contrast to Approach 1, it does not aim to measure mediation. Approach 3 focuses on the actual observed disparity and evaluates the impact of manipulations of explanatory factors – such as potential mediators, the focus of this article – on this disparity [18,24]. In contrast to Approach 1, it does not aim to measure mediation, and, in contrast to Approaches 1 and 2, it does not involve specifying causal effects of the exposure.

³ This alternative interventional analogue to NIE_M is $E[Y_a - Y_{a\tilde{M}_{a^*|C}}]$, which compares the expected counterfactual outcome under $A = a$ with the expected outcome when the mediator is randomly drawn from its counterfactual distribution under $A = a^*$, given C . The corresponding analogue to NDE_M , given by $E[Y_{a\tilde{M}_{a^*|C}} - Y_{a^*}]$, however, captures both the effect of changing the exposure from a^* to a and the effect of having the mediator set to its natural level under a^* compared to a random draw from the mediator under a^* [31].

Approaches 1–3 are illustrated using the DAG in Figure 1 (a), which involves multiple interdependent mediators $M_1, \dots, M_K$. The indices assigned to the mediators represent a working order, which may not align with the true causal order. The focus will be on interventional effects, as proposed by Vansteelandt and Daniel [19], Moreno-Betancur et al. [20], and Jackson and VanderWeele [18], that do not rely on the correct specification of the structural dependence among mediators. This acknowledges the practical difficulties associated with such specifications. The DAG in Figure 1 (a) explicitly distinguishes between the mediators of interest and other upstream intermediates (i.e., those causally prior to the mediators), unlike the DAGs in previous studies [18–20]. This distinction may be relevant in practical applications. An example is provided in Section 5, where L involves socio-demographic covariates that may act as exposure-induced confounders.

For the effect definitions, the following additional notation is used: $\mathbf{M}$ denotes the set of the mediators $M_1, \dots, M_K$, i.e., $\mathbf{M} = (M_1, \dots, M_K)$. Furthermore, $\tilde{\mathbf{M}}$ denotes the vector of random draws $\tilde{M}_1, \dots, \tilde{M}_K$ from a user-specified joint intervention distribution. $\mathbf{I}$ denotes a subset of $\mathbf{M}$ such that $\mathbf{I} \subseteq \mathbf{M}$. The remaining mediators in $\mathbf{M}$ are summarized in the vector $\mathbf{R}$ such that $\mathbf{I} \cup \mathbf{R} = \mathbf{M}$. For any subset of mediators $\mathbf{I}$ (or $\mathbf{R}$), $\tilde{\mathbf{I}}$ ($\tilde{\mathbf{R}}$) denotes the vector of random draws from a user-specified joint intervention distribution of the variables in $\mathbf{I}$ ($\mathbf{R}$). Unless otherwise stated, this article refers to average effects. Effects for the covariate strata in $\mathbf{C}$ can be obtained by conditioning on $\mathbf{C}$. Positivity of exposure conditions⁴ and of relevant mediator values in the support of the intervention distributions⁵, as well as consistency, which links counterfactual to observed values,⁶ are assumed throughout. As a reminder, the following sections focus on a binary exposure that takes values in $\{0, 1\}$.

3.1 Approach 1: Causal mediation analysis of the exposure-induced disparity

Approach 1, the causal mediation approach, aims to decompose the exposure-induced disparity into direct and indirect effects through which the exposure affects the outcome. The targeted estimands are typically natural effects and path-specific effects [28,41,42,44]. However, these effects require strong assumptions that may not hold in scenarios with multiple mediators [48]. For example, the NIE through a subset of mediators $\mathbf{I}$, $\mathbf{I} \subseteq \mathbf{M}$, is generally not identified under the structural assumptions in the model in Figure 1 (a). This is due to exposure-induced confounding through L and potentially through other mediators in $\mathbf{R}$ that are upstream of $\mathbf{I}$.

For settings with multiple interdependent mediators, Vansteelandt and Daniel [19] introduced a decomposition of the total effect, $TE = E[Y_{1M_1} - Y_{0M_0}]$, into interventional path-specific (in)direct effects. This approach does not require cross-world independence assumptions or assumptions about the structural dependence of the mediators. Consequently, it is also applicable when the causal ordering of mediators is unknown or when mediators share unmeasured common causes. To illustrate this method and the challenges that arise when pursuing a conventional mediational interpretation, the following question will be addressed:

What is the average indirect effect of A on Y through the mediators in $\mathbf{I}$, capturing all of the exposure effect that is mediated by $\mathbf{I}$, but not by causal descendants of $\mathbf{I}$ in the graph? (Question 1)

The method proposed by Vansteelandt and Daniel [19] suggests that Question 1 can be approached using an interventional indirect effect, defined by the contrast between two expected counterfactual outcomes under exposure resulting from two different stochastic mediator interventions. The first intervention is defined by setting the values of the mediators in $\mathbf{I}$ to random draws from their counterfactual joint distribution under exposure, given $\mathbf{C}$, denoted by $\tilde{\mathbf{I}}_{1|\mathbf{C}}$. Similarly, the values of the mediators in $\mathbf{R}$ are assigned random draws from their counterfactual joint distribution under exposure, given $\mathbf{C}$, denoted by $\tilde{\mathbf{R}}_{1|\mathbf{C}}$. In summary, the mediators

⁴ $P(A = a'|\mathbf{C}) > 0$ for all $a' \in \{a, a^*\}$

⁵ For example, if the intervention distribution of the mediators is specified to be $P_{M_a|\mathbf{C}}(\mathbf{m}|\mathbf{C})$, positivity of relevant mediator values $\mathbf{m}$ refers to $P(\mathbf{M} = \mathbf{m}|A = a, \mathbf{C} = \mathbf{c}) > 0$ for all $\mathbf{m}$ in the support of the intervention distribution where $P(A = a, \mathbf{C} = \mathbf{c}) > 0$ [21].

⁶ $P(Y_{am} = Y) = 1$ if $A = a$ and $\mathbf{M} = \mathbf{m}$, and $P(\mathbf{M}_a = \mathbf{M}) = 1$ if $A = a$.

are set to a random draw from the distribution $P_{I|C}(\mathbf{i}|\mathbf{C}) \times P_{R|C}(\mathbf{r}|\mathbf{C})$. The second intervention is defined by setting the values of the mediators in $\mathbf{I}$ to random draws from their counterfactual joint distribution under control, given $\mathbf{C}$, denoted by $\tilde{\mathbf{I}}_{0|C}$. The distribution of the mediators in $\mathbf{R}$ is held constant at their counterfactual joint distribution under exposure, given $\mathbf{C}$. In summary, the mediators are set to random draws from the distribution $P_{I_0|C}(\mathbf{i}|\mathbf{C}) \times P_{R|C}(\mathbf{r}|\mathbf{C})$. These interventions define an interventional indirect effect through $\mathbf{I}$ as follows:

$$\text{IIE}_{I|C} = E[Y_{\tilde{\mathbf{I}}_{0|C}\tilde{\mathbf{R}}_{1|C}} - Y_{\tilde{\mathbf{I}}_{0|C}\tilde{\mathbf{R}}_{0|C}}]. \quad (1)$$

Specifying the intervention distributions of both $\mathbf{I}$ and $\mathbf{R}$ marginally with respect to $\mathbf{L}$ enables a decomposition of the TE into interventional effects, similar to that proposed by Vansteelandt and Daniel [19] (where $\mathbf{I} = \mathbf{M}_2$, $\mathbf{R} = \mathbf{M}_1$, and $\mathbf{L} = \emptyset$). This decomposition consists of $\text{IIE}_{I|C}$, as well as an interventional indirect effect through $\mathbf{R}$, an interventional indirect effect through the interdependence of the mediators, and an interventional effect of A on Y not through $\mathbf{I}$ and $\mathbf{R}$.⁷ For all possible subsets of $M_1, \dots, M_K$ in $\mathbf{I}$ and $\mathbf{R}$, these components are nonparametrically identified under the following conditional independence assumptions [19], interpreted as assumptions of no unmeasured confounding:

- A1 $Y_{am_1 \dots m_K} \perp\!\!\!\perp A|\mathbf{C}$: no unmeasured confounding of the relationship between A and Y conditional on $\mathbf{C}$,
- A2 $Y_{am_1 \dots m_K} \perp\!\!\!\perp (M_1, \dots, M_K) \setminus \{A = a, \mathbf{C}, \mathbf{L}\}$: no unmeasured confounding of the relationship between the mediators $M_1, \dots, M_K$ and Y conditional on $A = a, \mathbf{C}, \mathbf{L}$,
- A3 $(M_{1a}, \dots, M_{Ka}) \perp\!\!\!\perp A|\mathbf{C}$: no unmeasured confounding of the relationship between A and $M_1, \dots, M_K$ conditional on $\mathbf{C}$.

If A1, A2, and A3 hold, $\text{IIE}_{I|C}$ is nonparametrically identified by a function of the observed data (see [19] and Appendix, Section C), given by

$$\sum_{\mathbf{c}} \sum_{\mathbf{l}} \sum_{m_1, \dots, m_K} E[Y|A = 1, \mathbf{c}, \mathbf{l}, \mathbf{m}](P(\mathbf{i}|A = 1, \mathbf{c}) - P(\mathbf{i}|A = 0, \mathbf{c}))P(\mathbf{r}|A = 1, \mathbf{c})P(\mathbf{l}|A = 1, \mathbf{c})P(\mathbf{c}). \quad (2)$$

A1, A2, and A3 hold under the structural assumptions in Figure 1 (a). They also hold when $\mathbf{L}$ and $\mathbf{M}$ share unmeasured common causes, or when $\mathbf{L}$ and $\mathbf{Y}$ share unmeasured common causes (see Figure A1 in the Appendix) [17,19]. However, these conditions do not ensure that $\text{IIE}_{I|C}$ satisfies the indirect effect measure criteria outlined by Miles [22]. These criteria consist of the sharp null, sharper null, and the monotonicity criteria. The sharp(er) null criterion asserts that an indirect effect measure is equal to the null value when no individual indirect effect exists within the population of interest. The monotonicity criterion asserts that an indirect effect measure is not only capable of detecting the presence of an indirect effect, but is also able to correctly identify the direction of the mediated effect at least for some subjects. Based on a single-mediator model, Miles [22] showed that the interventional indirect effect $\text{IIE}_{M|C} = E[Y_{1\tilde{\mathbf{M}}_{1|C}} - Y_{1\tilde{\mathbf{M}}_{0|C}}]$ can fail to satisfy these criteria. Scenarios in which this happens, may be rare. For instance, the absence of overlap between the groups of individuals for whom the exposure affects the mediator and the groups of individuals for whom the mediator affects the outcome would be indicative of such an occurrence. Nonetheless, additional assumptions are necessary to exclude such scenarios and ensure a valid mediational interpretation of $\text{IIE}_{M|C}$. Either of the following two conditions is sufficient for $\text{IIE}_{M|C}$ to satisfy the indirect effect measure criteria [22]:

- A4 There are no exposure-induced confounders of the mediator and the outcome, or
- A5 There is no mean interaction between exposure-induced confounders and the mediator on the outcome on the additive scale.

Miles' findings suggest that the mediator can be either a single variable or a vector of variables. Thus, assumptions A4 and A5 can be mapped to $\text{IIE}_{I|C}$ with $\mathbf{I} = \mathbf{M}$. To assess the plausibility of A4 or A5 with respect to a proper mediator subset $\mathbf{I}$, assumptions must be made about which factors in $\mathbf{R}$ precede mediators in $\mathbf{I}$ and

⁷ Specifically, these four components are: 1) $\text{IIE}_{I|C}$, 2) an interventional indirect effect through $\mathbf{R}$, given by $\text{IIE}_{R|C} = E[Y_{\tilde{\mathbf{I}}_{0|C}\tilde{\mathbf{R}}_{1|C}} - Y_{\tilde{\mathbf{I}}_{0|C}\tilde{\mathbf{R}}_{0|C}}]$, 3) an interventional direct effect not through $\mathbf{I}$ and $\mathbf{R}$, given by $\text{IDE}_{M|C} = E[Y_{1\tilde{\mathbf{M}}_{0|C}} - Y_{0\tilde{\mathbf{M}}_{0|C}}]$, 4) the difference between the sum of these three interventional effects and the TE, which, in turn, constitutes a separate interventional indirect effect resulting from the effect of the exposure on the dependence between mediators, given $\mathbf{C}$ [19,39].

act as exposure-induced confounders in the $I - Y$ relationship. The indirect effect measure criteria ensure that $IIE_{I|C}$ can be interpreted as an analogue of a natural path-specific indirect effect from A to Y via I . This effect captures all pathways from A to I (direct and via upstream intermediates), and from I to Y , excluding pathways passing through causal descendants of I in the graph. This interpretation holds regardless of whether the underlying causal structure among the mediators is known [19].

Tchetgen Tchetgen and VanderWeele [49] showed that the NIE through a mediator is nonparametrically identified, even in the presence of an exposure-induced confounder, provided that assumption A5 holds. Therefore, under assumptions A4 or A5, the interventional indirect effect analogue offers no practical advantage over the NIE, as both share the same identification formula [22]. If neither assumption holds, alternative indirect effects could be considered that meet the indirect effect measure criteria (see Appendix, Section B). However, these alternatives do not adequately address Question 1.

To conclude, without additional assumptions beyond those required for its identification, $IIE_{I|C}$ may fail to adequately address Question 1. These considerations with respect to Question 1 and $IIE_{I|C}$ demonstrate that mediation analysis in contexts with interdependent mediators is challenging, as it relies on heavy assumptions. Although interventional (in)direct effects were presented as a pragmatic approach to mediation in such complex settings [17,19], the results of Miles [22] raise doubts about whether this approach can truly facilitate mediation analyses. Nevertheless, even without a conventional mediational interpretation, interventional (in)direct effects may still provide valuable insights. For instance, $IIE_{I|C}$ provides a measure of the impact of differing counterfactual distributions of I under exposure and under control (given C) on the outcome. It can be non-zero only if the exposure alters the distribution of I , and this change in distribution affects the outcome [31]. Such a measure may capture mediational concepts relevant from a population-level perspective [43]. See Miles [22] for alternative causal interpretations of interventional indirect effects.

3.2 Approach 2: Evaluating the impact of manipulations of mediating mechanisms on the exposure-induced disparity

Moreno-Betancur et al. [20] described an interventional approach that evaluates the effects of mediator interventions, conceivable for hypothetical target trials, on the exposure-induced disparity. Based on this work, Approach 2 is defined, in which interventional effects are of intrinsic interest, rather than pragmatic alternatives. In contrast to Approach 1, Approach 2 does not aim to measure the causal mechanisms (i.e., the direct and indirect effects) through which the exposure affects the outcome. Consequently, indirect effect measure criteria are not crucial for Approach 2.

As shown by Moreno-Betancur et al. [20], various intervention strategies can be considered in a multi-mediator setting. For example, by evaluating interventions on each mediator separately, it is possible to determine which intervention would most reduce the exposure-induced disparity. A sequential intervention strategy entails the modification of mediators in a predetermined sequence. The reduction in disparity can be evaluated following each modification. Generally, Approach 2 allows for the flexible specification of mediator interventions tailored to the scientific question at hand. As such, it is possible to fix mediators at certain values, resulting in degenerate mediator distributions. Consequently, controlled direct effects (CDEs) and the portion eliminated ($TE - CDE$) [41,42] can be categorized within Approach 2 [20,21]. CDEs are defined through interventions where the mediators under exposure and under control are set to constant values. For example, the average CDE when setting the mediators in M to the values m is defined by $CDE_m = E[Y_{1m} - Y_{0m}]$. Interventional direct effects can be viewed as the average of CDEs corresponding to various levels of the mediator, averaged according to a specified mediator distribution [17]. They can be employed in Approach 2 to measure the exposure-induced disparity that would remain if mediator distributions under exposure were equal to those under control.

The following estimand, which illustrates the approach, involves a similar intervention on the mediators as defined above for $IIE_{I|C}$ (equation (1)). However, the scientific question differs: the aim is not to measure an

indirect effect, but to evaluate the reduction in the exposure-induced disparity resulting from a specific stochastic mediator intervention. Specifically, the question is:

What is the average reduction in the exposure-induced disparity (TE) achieved by setting the joint distribution of the mediators in $\mathbf{I}$ under exposure equal to what it would be in the counterfactual world under control, given $\mathbf{C}$, while keeping the joint distribution of the other mediators constant as it would be under exposure, given $\mathbf{C}$ and $\mathbf{L}$ under exposure? (Question 2)

The mediator intervention corresponding to Question 2 involves setting the mediators in $\mathbf{I}$ and $\mathbf{R}$ under exposure to a random draw from the distribution $P_{I_0|C}(i|\mathbf{C}) \times P_{R_1|C,L_1}(\mathbf{r}|\mathbf{C}, \mathbf{L}_1)$, where the counterfactual $\mathbf{L}_1$ refers to $\mathbf{L}$ under exposure. This intervention results in a shift of the joint distribution of $\mathbf{I}$ under exposure to match the counterfactual distribution under control, given $\mathbf{C}$, thereby severing the dependence of $\mathbf{I}$ on $\mathbf{L}$ and $\mathbf{R}$. Meanwhile, the joint distribution of $\mathbf{R}$ remains unchanged under exposure, preserving the dependence among $\mathbf{L}$ and $\mathbf{R}$. The reduction in the TE resulting from this intervention is an interventional effect that refers to the change in the counterfactual outcome under exposure. Formally, it is given by

$$IE_{I|C} = E[Y_{1M_1} - Y_{I_0|C} \bar{R}_{1|C,L_1}]. \quad (3)$$

The residual exposure-induced disparity, given the world under control remains unchanged, is given by $RE_{I|C} = E[Y_{I_0|C} \bar{R}_{1|C,L_1} - Y_{0M_0}]$. To identify $IE_{I|C}$ and $RE_{I|C}$, assumptions A1–A3 are required [20,21].⁸ A1 and A2 ensure that the effects of exposure and mediator interventions on the outcome are identified. A3 ensures that the counterfactual mediator distributions under different exposure conditions (given $\mathbf{C}$) are identified. If A1–A3 hold, $IE_{I|C}$ is nonparametrically identified for each possible mediator subset of $M_1, \dots, M_K$ in $\mathbf{I}$ by (see Appendix, Section C, and Moreno-Betancur et al. [20], where $\mathbf{I} = M_k$ ($k = 1, \dots, K$) and $\mathbf{L} = \emptyset$):

$$\sum_c \sum_{\mathbf{l}} \sum_{m_1, \dots, m_K} E[Y|A = 1, \mathbf{c}, \mathbf{l}, \mathbf{m}](P(\mathbf{m}|\mathbf{A} = 1, \mathbf{l}, \mathbf{c}) - P(\mathbf{i}|\mathbf{A} = 0, \mathbf{c})P(\mathbf{r}|\mathbf{A} = 1, \mathbf{l}, \mathbf{c}))P(\mathbf{l}|\mathbf{A} = 1, \mathbf{c})P(\mathbf{c}). \quad (4)$$

3.3 Approach 3: Evaluating the impact of manipulations of explanatory mechanisms on the actual observed disparity

VanderWeele and Robinson [24] and Jackson and VanderWeele [18] outlined a framework for decomposing actual observed disparities, referred to as Approach 3. This approach evaluates the effects of altering the distributions of explanatory variables, such as potential mediators or variables that lie on a backdoor path from the exposure to the outcome, on the exposure–outcome association. Such evaluations help identify intervention targets aimed at reducing actual observed social disparities [18]. Unlike Approaches 1 and 2, Approach 3 does not aim to identify causal effects of the exposure. This makes it a viable option in cases where there are unmeasured confounders of the exposure–outcome or exposure–mediator relationships, or when interventions on the exposure would be ill-defined (e.g., because the exposure is non-manipulable). Furthermore, the method is adaptable to multiple-mediator settings and, like Approach 2, offers flexibility in specifying intervention strategies and distributions.

The following question, which serves to illustrate the approach, involves interventions on hypothesized mediators that may resemble those defined above for $IE_{I|C}$ (equation (3)). The key difference lies in the nature of the intervention distributions. Unlike before, the question refers to an observed intervention distribution rather than a counterfactual one. Specifically, the question is:

What is the average reduction in the actual observed outcome disparity, $E[Y|A = 1] - E[Y|A = 0]$, achieved by setting the joint distribution of the mediators in $\mathbf{I}$ among the exposed group equal to that among the unexposed group, given $\mathbf{C}$, while keeping the joint distribution of the other mediators constant, given $\mathbf{C}$ and $\mathbf{L}$? (Question 3)

⁸ As stated by Nguyen et al. [23], conditioning on L_a requires consistency of L_a and the conditional independence assumption $(L_a, M_{1a}, \dots, M_{Ka}) \perp\!\!\!\perp A|\mathbf{C}, a \in \{0, 1\}$. This independence is implied by assumption A3: $(M_{1a}, \dots, M_{Ka}) \perp\!\!\!\perp A|\mathbf{C}$, since L_a is a cause of M_a . Therefore, $\mathbf{C}$ must also include confounders of the A – $\mathbf{L}$ relationship.

The intervention associated with Question 3 sets the mediators in $\mathbf{I}$ and $\mathbf{R}$ among the exposed to a random draw from the distribution $P_{\mathbf{I}|A=0, \mathbf{C}}(\mathbf{i}|A = 0, \mathbf{C}) \times P_{\mathbf{R}|A=1, \mathbf{C}, \mathbf{L}}(\mathbf{r}|A = 1, \mathbf{C}, \mathbf{L})$. As a consequence, the joint distribution of $\mathbf{I}$ among the exposed is aligned with that among the unexposed, given $\mathbf{C}$, thereby removing the dependence of $\mathbf{I}$ on $\mathbf{L}$ and $\mathbf{R}$. The resulting reduction in the observed disparity, attributable to a change in the mean of Y among the exposed, is formally given by

$$\text{IE}_{\mathbf{I}|\mathbf{C}(\text{obs})} = E[Y|A = 1] - E[Y_{\bar{\mathbf{I}}|A=0, \mathbf{C}\bar{\mathbf{R}}|A=1, \mathbf{C}, \mathbf{L}}|A = 1]. \quad (5)$$

The residual disparity, given the world among the unexposed remains unchanged, is given by $\text{RE}_{\mathbf{I}|\mathbf{C}(\text{obs})} = E[Y_{\bar{\mathbf{I}}|A=0, \mathbf{C}\bar{\mathbf{R}}|A=1, \mathbf{C}, \mathbf{L}}|A = 1] - E[Y|A = 0]$, which is a non-causal estimand. The counterfactual $E[Y_{\bar{\mathbf{I}}|A=0, \mathbf{C}\bar{\mathbf{R}}|A=1, \mathbf{C}, \mathbf{L}}|A = 1]$ is identified if there is no unmeasured confounding of the relationship between $M_1, \dots, M_K$ and Y conditional on $A = 1, \mathbf{C}, \mathbf{L}$. In particular, $\text{IE}_{\mathbf{I}|\mathbf{C}(\text{obs})}$ is then nonparametrically identified by (see Appendix, Section C, and [18] for a related estimand):

$$E[Y|A = 1] - \sum_{\mathbf{c}} \sum_{\mathbf{l}} \sum_{m_1, \dots, m_K} E[Y|A = 1, \mathbf{c}, \mathbf{l}, \mathbf{m}]P(\mathbf{i}|A = 0, \mathbf{c})P(\mathbf{r}|A = 1, \mathbf{l}, \mathbf{c})P(\mathbf{l}|A = 1, \mathbf{c})P(\mathbf{c}|A = 1). \quad (6)$$

Based on the structural assumptions in Figure 1 (a), $\text{IE}_{\mathbf{I}|\mathbf{C}(\text{obs})}$ is nonparametrically identified. Notably, it would also be identified under weaker assumptions, such as when A and Y , or A and $\mathbf{M}$, share unmeasured common causes (see Figure A2 in the Appendix).

Jackson and VanderWeele [18] linked their proposed interventional framework to the twofold OB decomposition method. Building on their findings, the next section explores using the twofold OB decomposition for linear outcome models within Approaches 1–3.

4 Twofold OB decomposition for linear outcome models

This section addresses the applicability of the twofold OB decomposition for linear outcome models as an estimator within the three approaches described in Section 3. It is guided by pre-specified nonparametric causal estimands, structural assumptions, and modeling assumptions. For illustrative purposes, the focus will be on settings with a single mediator and a single exposure-induced confounder, as illustrated in the models presented in Figure 1 (b) and (c).

4.1 Specification and identification of causal estimands

Consider the following scientific questions and corresponding causal estimands:

- (1) *What is the average reduction in the actual observed outcome disparity achieved by setting M among the exposed equal to its marginal mean among the unexposed?* (Question 4)

According to the interventional effects classification in Section 3, Question 4 represents a specific causal estimand within Approach 3, given by

$$\text{IE}_{M(\text{obs})} = E[Y - Y_{E[M|A=0]}|A = 1]. \quad (7)$$

The remaining exposure–outcome association is given by $\text{RE}_{M(\text{obs})} = E[Y_{E[M|A=0]}|A = 1] - E[Y|A = 0]$. If the conditional independence assumption $Y_m \perp\!\!\!\perp M|A = 1, \mathbf{C}, \mathbf{L}$ holds as in the models in Figure 1 (b) and (c), $E[Y_{E[M|A=0]}|A = 1]$ is nonparametrically identified by (see [18] for a related estimand):

$$\sum_{\mathbf{c}} \sum_{\mathbf{l}} E[Y|A = 1, M = E[M|A = 0], \mathbf{l}, \mathbf{c}]P(\mathbf{l}|A = 1, \mathbf{c})P(\mathbf{c}|A = 1). \quad (8)$$

- (2) *What is the average reduction in the exposure-induced disparity achieved by setting M under exposure equal to the counterfactual expectation of M under control?* (Question 5)

Question 5 represents a specific causal estimand within Approach 2, given by

$$\text{IE}_M = E[Y_1 - Y_{1E[M_0]}]. \quad (9)$$

The remaining exposure-induced disparity is defined by $RE_M = E[Y_{E[M_0]} - Y_0]$. IE_M and RE_M are nonparametrically identified under the structural assumptions of the DAGs in Figure 1 (b) and (c) (see [21] for a similar context). Specifically, under those in Figure 1 (c), which imply $Y_a \perp\!\!\!\perp A$, $Y_{am} \perp\!\!\!\perp M|A = a, C, L$, and $M_a \perp\!\!\!\perp A$, $E[Y_{E[M_0]}]$ is nonparametrically identified by the formula in equation (8).⁹

- (3) *What is the average indirect effect of A on Y through M , capturing all pathways from the exposure to the outcome through M ?* (Question 6)

Question 6 represents a causal estimand within Approach 1, defined here as the NIE through M given by

$$NIE_M = E[Y_{1M_1} - Y_{1M_0}]. \quad (10)$$

In general, however, NIE_M is not identified under the structural assumptions of the DAGs in Figure 1 (b) and (c), as they involve an exposure-induced confounder L . As an alternative, one may consider the interventional indirect effect, which is defined by setting the marginal distribution of M under exposure equal to that under control. Formally, this interventional indirect effect is given by

$$IIE_M = E[Y_{1\tilde{M}_1} - Y_{1\tilde{M}_0}], \quad (11)$$

with $\tilde{M}_a$, $a \in \{0, 1\}$, denoting a random draw from $P_{M_a}(m) = E_C[P_{M_a|C}(m|C)]$ [21]. The corresponding interventional effect not through M is given by $IDE_M = E[Y_{1\tilde{M}_0} - Y_{0\tilde{M}_0}]$.

IIE_M and IDE_M are nonparametrically identified under the structural assumptions of the DAGs in Figure 1 (b) and (c). Specifically, under the DAG in Figure 1 (c), IIE_M is identified by (see identification results in [21], also with respect to the DAG in Figure 1 (b)):

$$\sum_c \sum_l \sum_m E[Y|A = 1, m, l, c](P(m|A = 1) - P(m|A = 0))P(l|A = 1, c)P(c). \quad (12)$$

If $L = \emptyset$ (A4), or if there is no mean interaction between L and M on Y on the additive scale (A5) [22], IIE_M satisfies the indirect effect measure criteria, and thus serves as an indirect effect analogue to NIE_M .

As a side note, Vanderweele and Vansteelandt [50] demonstrated with respect to natural (in)direct effects that, if the outcome is linear in the mediator, it suffices to correctly specify the mediator's expectation rather than its entire distribution (for binary outcomes, see refs [51,52]). Extending these results to interventional effects suggests that, under linearity of the outcome in the mediator, $IE_{M(\text{obs})}$ (equation (7)) can be interpreted as the change in the expected outcome among the exposed when the distribution of M is aligned with its marginal distribution among the unexposed. Furthermore, IE_M (equation (9)) can be interpreted as the change in the counterfactual outcome under exposure when aligning the distribution of M under exposure with its marginal distribution under control.

4.2 Estimation using a variant of the twofold OB decomposition for linear outcome models

Consider the following linear outcome models for the exposed group ($A = 1$) and the unexposed group ($A = 0$):

$$E[Y|A = 1, M = m, L = l, C = c] = \alpha_0 + \alpha_1 m + \alpha_2 l + \alpha_3 ml + \alpha_4 c, \quad (13)$$

$$E[Y|A = 0, M = m, L = l, C = c] = \omega_0 + \omega_1 m + \omega_2 l + \omega_3 ml + \omega_4 c. \quad (14)$$

The marginal mean of Y in the exposure group is obtained by

$$E[Y|A = 1] = \alpha_0 + \alpha_1 E[M|A = 1] + \alpha_2 E[L|A = 1] + \alpha_3 E[ML|A = 1] + \alpha_4 E[C|A = 1]. \quad (15)$$

Equally, the marginal mean of Y in the unexposed group is obtained by

$$E[Y|A = 0] = \omega_0 + \omega_1 E[M|A = 0] + \omega_2 E[L|A = 0] + \omega_3 E[ML|A = 0] + \omega_4 E[C|A = 0]. \quad (16)$$

⁹ Given that A is assumed to be independent of C (see DAG in Figure 1 (c)), it follows that $P(c|A = 1) = P(c|A = 0) = P(c)$.

Using the standard twofold OB decomposition technique [8,9], the marginal disparity $E[Y|A = 1] - E[Y|A = 0]$ can be decomposed into two components:

$$E[Y|A = 1] - E[Y|A = 0] =$$

$$\left. \begin{aligned} &\alpha_1(E[M|A = 1] - E[M|A = 0]) + \alpha_2(E[L|A = 1] - E[L|A = 0]) + \\ &\alpha_3(E[ML|A = 1] - E[ML|A = 0]) + \alpha_4(E[C|A = 1] - E[C|A = 0]) + \end{aligned} \right\} \text{ explained,}$$

$$\left. \begin{aligned} &(\alpha_0 - \omega_0) + (\alpha_1 - \omega_1)E[M|A = 0] + (\alpha_2 - \omega_2)E[L|A = 0] + \\ &(\alpha_3 - \omega_3)E[ML|A = 0] + (\alpha_4 - \omega_4)E[C|A = 0], \end{aligned} \right\} \text{ unexplained,}$$

where the explained part refers to the part associated with (weighted) differences in the marginal means of the explanatory variables, and the unexplained part refers to the part associated with (weighted) differences in the coefficients. These two components can be further decomposed into the contributions of individual covariates, with the interaction between M and L being treated as a distinct covariate. When the focus is specifically on group differences in M , while taking into account the interaction between M and L , the following two components may be of interest:

$$OB_M = E[Y|A = 1] - E[Y|A = 1, M = E[M|A = 0], L = E[L|A = 1], C = E[C|A = 1]], \quad (18)$$

the component associated with differing means of M , while accounting for the interaction of M with L and holding L and C fixed at their expected values in the exposed group, and

$$OB_{RE} = E[Y|A = 1, M = E[M|A = 0], L = E[L|A = 1], C = E[C|A = 1]] - E[Y|A = 0], \quad (19)$$

the component not associated with differing means of M , but with differences in coefficients and differences in the expected values of L and C , as well as the interaction between M and L . Together, OB_M and OB_{RE} add up to the marginal disparity. This twofold decomposition, which is considered here as a variation of the twofold OB method, can be performed as follows:

$$E[Y|A = 1] - E[Y|A = 0] =$$

$$\left. \begin{aligned} &\alpha_1(E[M|A = 1] - E[M|A = 0]) + \\ &\alpha_3(E[ML|A = 1] - E[M|A = 0]E[L|A = 1]) + \end{aligned} \right\} OB_M$$

$$\left. \begin{aligned} &\alpha_0 - \omega_0 + (\alpha_1 - \omega_1)E[M|A = 0] + \alpha_2E[L|A = 1] - \omega_2E[L|A = 0] + \\ &\alpha_3(E[M|A = 0]E[L|A = 1]) - \omega_3(E[ML|A = 0]) + \\ &\alpha_4E[C|A = 1] - \omega_4E[C|A = 0]. \end{aligned} \right\} OB_{RE}.$$

Without further assumptions, OB_M and OB_{RE} are statistical quantities that measure differences between conditional expected means. Table 1 summarizes the nonparametric and parametric assumptions, which are

Table 1: Cumulative assumptions required for using OB_M to estimate $IE_{M(\text{obs})}$, IE_M , and $IE_M = NIE_M$

Assumptions	OB_M	OB_{RE}
Consistency	$IE_{M(\text{obs})} =$	$RE_{M(\text{obs})} =$
Positivity	$E[Y - Y_{E[M A=0]} A = 1]$	$E[Y_{E[M A=0]} A = 1] - E[Y A = 0]$
$Y_m \perp\!\!\!\perp M A = a, C, L$		(non causal)
Linear outcome model		
Correct model specification		
$Y_{am} \perp\!\!\!\perp M A = a, C, L$	$IE_M = E[Y_1 - Y_{1E[M_0]}]$	$RE_M = E[Y_{1E[M_0]} - Y_0]$
$Y_{am} \perp\!\!\!\perp A$		
$M_a \perp\!\!\!\perp A$		
A4 ($L = \emptyset$) [15,22], or	$IE_M = NIE_M =$	$IDE_M = NDE_M =$
A5 (no mean interaction	$E[Y_{1M_1} - Y_{1M_0}]$	$E[Y_{1M_0} - Y_{0M_0}]$
between L and M) [22]		

For completeness, estimands corresponding to OB_{RE} are also listed.

required to use OB_M to estimate the causal estimands $IE_{M(\text{obs})}$ (equation (7)), IE_M (equation (9)), and $IIE_M = NIE_M$ (equations (10) and (11)). They are cumulative, meaning each row's assumption assumes the previous rows' assumptions also hold. For completeness, the estimands corresponding to OB_{RE} are also listed.

As detailed in Table 1, the implementation of the proposed decomposition (equation 20), within Approach 3 requires the fewest assumptions. Conversely, its application within Approaches 1 and 2 entails considerably stronger conditions, including the exogeneity of the exposure. Jackson and VanderWeele [18] suggested applying the OB method to the strata of baseline covariates (i.e., conducting conditional OB decompositions), as this technique can have important implications for causal inference. For example, when the exposure shares a common cause C with the mediator or the outcome, as in Figure 1 (b), $(Y_a, M_a) \perp\!\!\!\perp A$ does not hold, but $(Y_a, M_a) \perp\!\!\!\perp A|C = c$ does. In conclusion, given the varying degrees of strength of the assumptions involved, it can be stated that OB decompositions (and their variations) are most applicable to Approach 3.

Correct model specification remains crucial for unbiased estimation in all approaches. In settings with complex relationships, such as various (possibly higher-order) interactions or nonlinearities, standard OB decompositions (or their variations) may prove inadequate for estimating the targeted estimand or may be impractical. In such cases, other estimation methods that allow for flexible modeling, such as g-computation [19], or semiparametric and nonparametric multiply-robust methods [53,54], may be more appropriate. Section 5.3 outlines the g-computation approach employed in the interventional effects analysis of the gender pay gap.

5 Application of the interventional effects framework to the gender pay gap

5.1 Preliminary considerations on the choice of approach

In studying the gender pay gap, Approach 1 would be of interest in examining the direct and indirect effects of gender on wages, while Approach 2 would be used to evaluate the reduction of gender-induced wage disparity through relevant mediator interventions. Both approaches would entail the construction of causal effects of gender, implying the notion of hypothetical manipulations of gender. However, these concepts become somewhat ambiguous without a precise definition of gender or of which aspect of gender is in view (e.g., social roles, norms, perceptions, behaviors, or physical attributes). There is a debate over whether valid causal inference is possible with respect to ill-defined or non-manipulable exposures such as gender or race [55–59]. Some authors suggest shifting the focus from actual traits to perceived traits, which facilitates experimental manipulation [60–62]. For example, researchers can obscure the visible attributes of an applicant's gender by employing methods such as placing a screen between candidates and decision-makers [63]. The illustrative model presented in Figure 2 for the effect of gender on wage involves gender perception. The model posits that gender affects wage through gender perception, suggesting that wage decisions may be

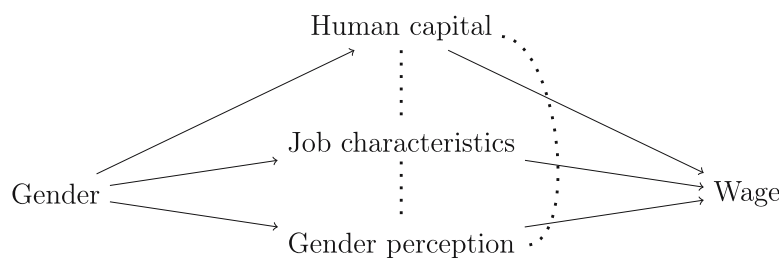


Figure 2: Conceptual overview of the factors assumed to mediate the gender pay gap, including “Gender perception”. Dotted lines indicate that there may be dependencies among these factors. This figure is illustrative and does not represent a formal DAG.

influenced by whether an individual is perceived as a man or a woman. The meaning of “gender perception” will not be further elaborated upon here; instead, the focus will be on analytical techniques when assuming a model like that in Figure 2. In this model, the pathway from gender to wage via gender perception could be addressed using the concept of CDEs [42]. This approach would set all mediators – except gender perception – at equal levels for both women and men. Consequently, any remaining wage disparity could be attributed to a direct effect of gender through gender perception. However, this method relies on challenging conditions. Apart from conditional independence assumptions [40], the adjusted set of mediators must be exhaustive to accurately separate the Gender → Gender Perception → Wage pathway from other pathways.

Approach 3 circumvents the pitfalls of imprecise causal attributions to gender or gender categories as measured in a survey. The approach can be employed to evaluate the contribution of differences between men and women in the distributions of wage-relevant factors to the actual observed pay gap, without the need to define causal effects of gender. By simulating relevant hypothetical scenarios (e.g., where mediators are equally distributed across men and women), intervention targets can be identified to reduce the gap. Approach 3 is illustrated in the following sections through an empirical analysis of the 2017 gender pay gap in Western Germany. Its comprehensive nature is shown by defining different mediator intervention distributions, and by directing interventions at women alone (Aims 1a–b), as well as at both women and men (Aims 2a–b), thereby addressing considerations regarding hypothetical changes in the working lives of both groups. For a comparison of Approach 3 with Approaches 1 and 2 with respect to the empirical analysis, see Section E.2 of the Appendix. This section of the Appendix summarizes key conditions under which the estimated effects represent specific quantities within Approaches 1–3, suggesting possible interpretations.

5.2 Data and methods

This section illustrates a cross-sectional application of Approach 3 to the gender pay gap in the labor market of Western Germany, using data from the 2017 SOEP, a large-scale, representative longitudinal survey of private households [35,36]. The analysis examined how, under certain assumptions, hypothetical manipulations of wage-relevant factors could have narrowed the observed gender pay gap between women and men. The focus was on ten potential mediators, including indicators of education, labor market experience, and characteristics of employment. These variables, which are presented with descriptive statistics in Table 2, include concepts that have been considered as relevant explanatory factors in gender pay gap analyses [4–6,64]. In general, categorical mediators with more than two categories were summarized into binary variables. Continuous mediators with highly skewed distributions were dichotomized to ease interpretation and to ensure that the parametric assumptions for estimation were met. The covariate set included age in years (C), an indicator for the presence of at least one child under the age of 18 in the household (L_1), and an indicator for a direct migration background (L_2) (Table 2). Based on prior research, these covariates were considered relevant to various labor market outcomes [65–70]. Gross hourly wages were calculated from gross labor income and weekly working hours at the time of the survey, as generated by the SOEP [71]. As suggested by Brenke and Müller [72], the calculation considered both compensated and uncompensated overtime, along with actual and contracted hours. Due to the skewed distribution of gross hourly wages, the natural logarithm of gross hourly wages was used as the primary outcome. Assumptions about the data-generating process are illustrated in Figure A3 in the Appendix.

The study population comprised 11,924 individuals after the exclusion of individuals with no earnings or less than one euro, those under the age of 18 or above the retirement age of 67, pensioners, persons in education, and individuals with missing values. Women (6,182, 51.8%) were treated as the exposure group ($A = 1$), while men (5,742) were treated as the control group ($A = 0$). Women earned an average of 16.96 euros per hour, around 5 euros less than men (21.94 euros). The marginal gender pay gap in this sample, defined as the percentage by which women earned less than men on average, was 22.7%. This number is in line with the findings of the German Federal Statistical Office, which reported a gender pay gap of 22% in Western Germany in 2017 [74]. The marginal gender pay gap in log gross hourly wages in the sample was 8.9%.

Table 2: Assumed mediators of the gender pay gap, covariates, and the outcome, with empirical means for both binary and numerical variables, and interquartile ranges (IQR) for numerical variables, presented for the total sample, women, and men

<i>M</i>	Mediators listed according to the working order	Total (11,924) Mean (IQR)	Women (6,182) Mean (IQR)	Men (5,742) Mean (IQR)
<i>M</i> ₁	“≥ 12 years of education” (yes = 1)	0.50	0.51	0.48
<i>M</i> ₂	College degree or higher (“≥ College degree,” yes = 1)	0.25	0.23	0.27
<i>M</i> ₃	Current job requiring at least a college degree (“≥ College degree job,” yes = 1)	0.26	0.22	0.30
<i>M</i> ₄	≥ the median of 6.7 years of employment with company (“≥ 6.7 years in company,” yes = 1)	0.50	0.46	0.54
<i>M</i> ₅	Job in on of the following industries: services, trade, banking, insurance (“Female-dominated industry,” yes = 1) ¹	0.62	0.78	0.45
<i>M</i> ₆	Leading function with managerial responsibility: senior civil servants, white-collar employees with highly qualified work or comprehensive management responsibility (“Leading position,” yes = 1)	0.20	0.12	0.27
<i>M</i> ₇	“Full-time employment” (yes = 1)	0.64	0.40	0.90
<i>M</i> ₈	“Flexible working hours” (yes = 1)	0.41	0.37	0.49
<i>M</i> ₉	Work experience in years ² relative to age (“Work experience”)	0.36 (0.21; 0.51)	0.29 (0.17; 0.40)	0.42 (0.30; 0.57)
<i>M</i> ₁₀	Standard International Occupational Prestige Scale, based on the International Standard Classification of Occupations 1988 (“Job prestige (SIOPS)”)	44.56 (33; 53)	44.08 (33; 53)	45.08 (34; 56)
C, L				
Covariates				
<i>C</i>	Age in years	44.33 (36; 53)	44.21 (36; 53)	44.47 (36; 53)
<i>L</i> ₁	Child in household (yes = 1)	0.48	0.47	0.48
<i>L</i> ₂	Direct migration background (yes = 1)	0.21	0.20	0.22
<i>Y</i>	Outcome: Log gross hourly wage	2.80 (2.45; 3.14)	2.67 (2.34; 3.02)	2.93 (2.60; 3.29)

1 *M*₅ is based on the 1 Digit Industry Code of Individual as generated by the SOEP [73].
2 Full-time and part-time work experience in years was generated by the SOEP using combined monthly and annual employment data. From these measures, a measure for work experience was created, with one year of part-time experience counted as 0.5 years of full-time experience.

Two strategies were considered regarding the target group for mediator interventions. The first targeted women (Aims 1a–b), while the second targeted both women and men (Aims 2a–b). Since the education gap between women and men has narrowed significantly in recent decades [5], the mediator interventions in Aims 1a–b focused on aligning job characteristics and work experience. In order to reflect the relevance of age (C) and education (M_1, M_2) to career paths, the intervention distributions were conditioned on these factors. They were not conditioned on having a child (L_1) and migration background (L_2) to avoid restricting women to the intervention distributions prevalent in the strata defined by these covariates among men. Specifically, the research questions are:

- Aim 1a What reduction in the marginal gender pay gap in log gross hourly wages could be achieved if women's job characteristics and work experience ($I = M_3, \dots, M_{10}$) were aligned with those of men within levels of age and educational background, while keeping the joint distribution of the other mediators in women constant as observed, given C and L ?
- Aim 1b Which intervention would yield the largest reduction when setting the distribution of a single mediator M_j ($j = 3, \dots, 10$) in women equal to that in men, given age and educational background, while keeping the joint distribution of the other mediators in women constant as observed, given C and L ?

Aims 2a–b addressed the potential reductions in the gender pay gap if a policy targeting both women and men equalized the distributions of the hypothesized mediators. Specifically, the research questions are:

- Aim 2a What reduction in the marginal gender pay gap in log gross hourly wages could be achieved by aligning the joint distribution of $M_1, \dots, M_{10}$ among women and among men with that of the total sample population?
- Aim 2b Which intervention would yield the largest reduction when setting the distribution of a single mediator M_k ($k = 1, \dots, 10$) among women and among men equal to that in the total sample population, while keeping the joint distribution of the other mediators in women and men constant as observed, given C and L ?

For an overview, Table 3 lists the intervention distributions of the manipulated mediators, the estimands, and their corresponding interpretations for all study aims.

The estimation of the expected counterfactual outcomes is based on the assumptions of positivity, consistency, and the absence of unmeasured mediator–outcome confounders. Notably, unmeasured common causes of the mediators (e.g., individual skills) should not bias the results. See Appendix, Section E.2, for a summary of the required assumptions.

5.3 MC g-computation for estimation

This analysis employed MC g-computation, following the method outlined by Vansteelandt and Daniel [19] for interventional (in)direct effects. The OB decomposition described in Section 4 and a parametric approach based on simple linear outcome models, as outlined by Jackson and VanderWeele [18], were deemed unsuitable for estimating the target estimands of Aims 1a–b and 2a–b. This is because wage models likely involve complex relationships, including various interaction terms and nonlinear effects. Furthermore, the OB decomposition does not address the targeted estimands. The intervention distributions for Aims 1a–b are conditional on covariates and, as a result, are not equivalent to marginal means. Aims 2a–b address interventions for both the exposure and control groups, rather than solely for the exposed.

G-computation relies on correct model specification; however, it offers flexibility in modeling. Specifically, MC g-computation employs a simulation approach to generate counterfactual outcomes. This is achieved by drawing values for the mediator and outcome variables under the specified interventions for each individual from fitted statistical models, thereby creating counterfactual datasets. The algorithm for estimating $E[Y_{\bar{I}\bar{R}}|A=a, C, L|A=a]$, $a \in \{0, 1\}$, is outlined in Table 4. It adapts the estimator described by Vansteelandt and Daniel [19] in one main aspect: both the counterfactual outcome and the mediator intervention distributions are conditional on A , meaning that no interventions are applied to the exposure. For interventional effects

involving exposure interventions, Table A1 in the Appendix outlines the steps for estimating $E[Y_{\tilde{a}|c}|\tilde{c}\tilde{a}|c]$, which is central to IIE_{IC} described in equation (1).

In the concrete analysis, the joint distribution of mediators was specified using factorization based on the working order in Table 2. Logit models were fitted for binary variables and linear models for continuous variable. All possible two-way interaction terms among the regressors have been considered in the model-building process for the outcome model and the joint mediator distribution. The models were chosen depending on Akaike's model selection criterion. Vansteelandt and Daniel [19] suggested drawing several million times from the fitted models to ensure that the results are free of Monte Carlo error to the number of decimal places given. This analysis achieved stable results at 300 draws per unit. The algorithm was therefore implemented with 300 Monte Carlo runs (i.e., $Z = 300$ in Table 4). 95% confidence intervals (CIs) were obtained using the nonparametric bootstrap, with 1,000 bootstrap samples (i.e., $B = 1,000$ in Table 4).

A side note on the log-transformed outcomes: The g-computation method generates counterfactual datasets with outcomes for each individual, which can be converted back to the original scale (as done, e.g., in [75]). Nevertheless, as the log transformation of wages is standard in econometric literature and gender pay gap research [2,7–9,76], this article presents the results based on log hourly wages in order to enhance comparability.

5.4 Results

5.4.1 Aims 1a–b

Aim 1a: The marginal gender pay gap in log gross hourly wages (8.9%, 95% CI: 8.2; 9.4) was estimated to be reduced by 71.2% (CI: 63.4; 80.2), when setting the joint distribution of job characteristics and work experience ($M_3, \dots, M_{10}$) given age and educational background among women equal to that among men, while holding the joint distribution of the other intermediates constant. This is attributed to an estimated increase of 6.9% (CI: 6.2; 7.7) in the average log wages of women. As a result, a gender pay gap of 2.5% (CI: 1.7; 3.4) would remain. **Aim 1b:** When prioritizing a single-mediator intervention, the marginal gender pay gap in log gross hourly wages would be most reduced, by 26.5% (CI: 20.9; 32.6), if women had the same distribution of work experience as men within strata of age and educational background. This is attributed to an estimated increase of 2.6% (CI: 2.0; 3.1) in the average log wages of women. As a result, a gender pay gap in log gross hourly wages of 6.5% (CI: 5.7; 7.2) would remain.

Table 3: Manipulated mediators, intervention distributions, estimands, and their interpretations for Aims 1a–b and 2a–b

Mediators in I	Intervention distribution of I	Causal estimand	Interpretation
Aim 1a: $M_3, \dots, M_{10}$ Aim 1b: $M_j, j = 3, \dots, 10$	$P_{I A=0,C,M_1,M_2}(\mathbf{i} A = 0, C, M_1, M_2)$	$E[Y - Y_{\tilde{I}\tilde{R}} A=1,C,L]A = 1]$	Expected change in Y in women following intervention (observed - counterfactual)
Aim 2a: $\mathbf{M} = M_1, \dots, M_{10}$ Aim 2b: $M_k, k = 1, \dots, 10$	$P_I(\mathbf{i})$	$A = E[Y - Y_{\tilde{I}\tilde{R}} A=1,C,L]A = 1]$ $B = E[Y_{\tilde{I}\tilde{R}} A=0,C,L - Y A = 0]$ $C = A + B$	A: Expected change in Y in women following intervention (observed - counterfactual) B: Expected change in Y in men following intervention (counterfactual - observed) C: Overall reduction in the gender pay gap

Relative wage changes for women and men were calculated by dividing the difference between observed and counterfactual (post-intervention) wages by the observed wage. For all aims, percentage reductions in the gender pay gap were calculated by multiplying the effects by $100/(E[Y|A = 1] - E[Y|A = 0])$.

5.4.2 Aims 2a–b

Aim 2a: The marginal gender pay gap in log gross hourly was estimated to be reduced by 85.6% (CI: 69.4; 104.0), if the joint distribution of all mediators $M_1, \dots, M_{10}$ in women and in men were set equal to that in the overall sample. About 59.4% (CI: 49.2; 68.6) of this reduction is due to a decrease in log wages in men (4.5% decrease, CI: 3.2; 5.8), while the remaining 40.6% (CI: 31.4; 50.8) is due to an increase in log wages in women (3.4% increase, CI: 2.5; 4.3). As a result, a gender pay gap of 1.3% (CI: -0.4; 2.8) would remain.

Aim 2b: When prioritizing a single-mediator intervention, the marginal gender pay gap in log gross hourly wages would be most reduced, by 29.8% (CI: 20.2; 40), if the distribution of full-time employment in women and in men matched that in the overall sample. About 90% (CI: 80.7; 98.2) of this reduction is due to a decline in log wages in men (2.4% decrease, CI: 1.6; 3.2), while the remaining 10% (CI: 1.8; 19.3) is due to a slight increase in women's log wages (0.3% increase, CI: 0.1; 0.5). As a result, the remaining gender pay gap would be 6.4% (CI: 5.3; 7.4). The second largest reduction in the pay gap (21.5%, CI: 9.2; 35.3) could be achieved by aligning the distribution of work experience with that of the overall sample, mainly due to an increase in women's wages. In contrast, aligning educational backgrounds, which would lead to wage losses for women and gains for men, would widen the gap (see Figure 3 and Table A5 in the Appendix for detailed results).

6 Discussion

This article has outlined three distinct approaches to the mediating mechanisms of social disparities, each illustrated by representative target estimands. Approach 1 focuses on causal mediation to investigate the mechanisms underlying a social disparity. Approaches 2 and 3 assess the effects of hypothetical manipulations of mediators: Approach 2 on the exposure-induced disparity, and Approach 3 on the actual observed disparity. As such, both approaches are policy-relevant for identifying potential intervention targets to reduce social disparities [18,20]. Approaches 1 and 2 entail the construction of causal effects of the social exposure, which imply hypothetical interventions on the exposure. However, when the exposure is a complex or non-manipulable construct, such as gender or race, defining meaningful interventions on the exposure becomes challenging. In such instances, Approach 3, which does not entail the manipulation of the exposure, may offer more clearly defined target quantities. With regard to other exposure variables, such as membership of an educational group or social organization causal effects of the exposure may be less challenging to conceptualize, thereby rendering Approaches 1 or 2 viable options. Regarding the strength of the required conditions,

Table 4: Proposed algorithm to estimate $E[Y_{\bar{R}|A=a,C,L}|A=a]$

Algorithm to estimate $E[Y_{\bar{R}|A=a,C,L}|A=a]$, $a \in \{0, 1\}$

- | | |
|--------|---|
| Step 1 | Specify and fit models for:
1) the distribution of I ; for Aims 1a–b: conditional on $A = 0, C, M_1$, and M_2 , for Aims 2a–b: the distribution of I in the overall sample (Model 1)
2) the distribution of R conditional on $A = a, C$, and L (Model 2),
3) the outcome Y conditional on $A = a, C, L$, and M (Model 3).
For each individual with $A = a$: |
| Step 2 | Set the distribution of C and L to the respective empirical distributions in $A = a$. |
| Step 3 | Set the variables in I to a random draw from Model 1 fitted in Step 1. |
| Step 4 | Set the variables in R to a random draw from Model 2 fitted in Step 1. |
| Step 5 | Draw the outcome from Model 3 fitted in Step 1 based on the updated data generated through Steps 2–4. |
| Step 6 | Estimate $E[Y_{\bar{R} A=a,C,L} A=a]$ by calculating the mean of the simulated outcomes in Step 5. |
| Step 7 | Repeat Steps 3–6 Z times with different seeds and calculate the mean of the Z estimates. |
| Step 8 | To obtain the nonparametric bootstrap standard error and percentile confidence intervals, repeat Steps 1–7 B times on the bootstrapped data. |
-

Aim 2b: % Reductions in gender pay gap achieved by single mediator interventions

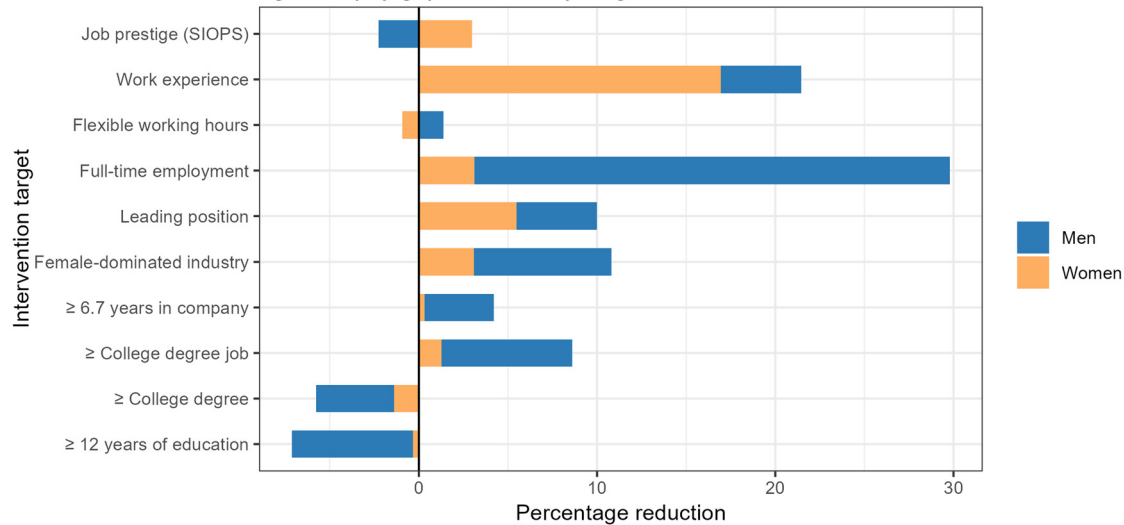


Figure 3: Percentage reductions in the marginal gender pay gap in log gross hourly wages achieved by setting the distribution of a single mediator among women and among men equal to that observed in the overall sample (Aim 2b), with portions attributable to changes in log wages for women and men. Positive values indicate a reduction in the pay gap, while negative values indicate an increase. For men, positive percentages reflect decreased average log wages, while for women they indicate increases; the opposite is true for negative values.

Approach 1 is the most demanding. In the context of mediation analysis, the key measures for causal mechanisms are typically natural (in)direct effects [23,42]. However, these measures rely on strong identification assumptions, including cross-world assumptions, which are often violated in multiple-mediator settings. Alternative indirect effect measures, including interventional indirect effects, may fail to capture the pathways under scrutiny [22]. A novel approach to mediation has recently been proposed by Díaz [77], showing promise for future research on the mechanisms of social disparities. This approach enables the identification and estimation of the strength of causal mediating mechanisms, even in the presence of exposure-induced mediator–outcome confounders. The proposed path-specific effects are based on non-agency interventions, a useful concept for non-manipulable exposures. Moreover, these effects are zero when no mediating mechanism exists, thus satisfying the path-specific sharp null criteria.

This article has placed a variant of the prominent twofold OB decomposition for linear outcome models within Approaches 1–3, concluding that causal implementations of the OB method are limited to estimating specific estimands that require strong identification assumptions. Moreover, when the model for the outcome (e.g., wage) involves a complex functional form, including nonlinearities or higher-order interactions, other estimation methods that allow for flexible modeling, such as MC g-computation, are considered more suitable.

MC g-computation has been employed in Approach 3 to analyze the gender pay gap in Western Germany, using data from the 2017 SOEP. The focus has been on hypothetical interventions targeting potential mediators of the gender pay gap. Specifically, the analysis has examined potential reductions in the observed gap by altering conditional mediator distributions. When interpreting the results, it is important to note that the findings relate to changes in the gender pay gap in log gross hourly wages. Due to the nonlinear transformation, these results do not directly translate to changes in the gap in actual hourly wages.

The results for the first study aim suggest that approximately 71% of the gender pay gap could be reduced if women had the same age- and education-specific distribution of job characteristics and work experience as men. In particular, a substantial portion of the gap could be reduced if women had the same distribution of work experience as men (given age and education). This may reflect wage penalties women face due to career interruptions, such as those related to parenthood and family care [68,78]. These findings are consistent with those of previous studies (see [5,6] for an overview of existing literature). However, there is considerable variation in the literature on the gender pay gap, particularly with regard to the set of variables, data sources,

time periods, regions, and analytical approaches. This variability complicates direct comparisons. For example, a decomposition analysis conducted by the German Federal Statistical Office, based on the 2018 Structure of Earnings Survey, attributed approximately 71% of the gender pay gap in log gross hourly wages to gender differences in the marginal means of wage-determining factors [7]. This analysis used different data and variables than this article, covered all of Germany, employed the OB method, and did not adopt a clear causal approach, despite being labeled a “cause analysis”.

The results of the second study aim differ from those of the first, highlighting the importance of how estimands are defined. Specifically, the findings suggest that the gender pay gap could be reduced by approximately 86% if the joint distribution of all considered mediators among women and among men were to align with that of the total sample population. The single-mediator analysis indicates that the gender pay gap would be substantially smaller if the proportion of men and women in full-time employment matched that of the full sample, while the joint distributions of other wage-relevant factors remained unchanged. Although women would experience a larger relative increase in full-time employment compared to the decrease for men as a result of this alignment, the relative wage gain for women would be modest, while the wage loss for men would be comparatively large. This could be due to a part-time wage penalty for men, as found in previous studies [79–83]. One possible explanation for a relatively high part-time penalty for men is that part-time work is less common among men than among women, making it more noticeable. This may lead to lower wages, as employers could perceive men’s part-time work as a signal of reduced work commitment [80].

Further research is needed to gain more nuanced insights into the gender-specific effects of mediator interventions on wages. One potential avenue of research would be to examine interventions in men and women with comparable covariates, such as equalizing full-time employment within the same occupations, while accounting for industry-specific wage structures (e.g., mandatory wage scales or the role of wage negotiations [84–87]). Defining a reasonable and comprehensive set of covariates on which the stochastic assignment of mediator values is conditioned is also crucial to achieve realistic interventions at the individual level. Without a meaningful selection of covariates, there is a risk of assigning implausible mediator values to individuals.

Approach 3 has the fewest prerequisites among the approaches discussed in Section 3. However, the identification of relevant counterfactual outcomes still relies on several key conditions, in particular: the absence of unmeasured confounders in mediator–outcome relationships, and assumptions of positivity and consistency. In the context of the gender pay gap, unmeasured confounders may include wage determinants that are not included in the analysis. Positivity violations occur when women or men have a zero probability of receiving a mediator value in the support of the intervention distribution, while consistency may be compromised if mediator interventions are not clearly defined. Additionally, real shifts in mediator distributions (e.g., more men working part-time) could affect working conditions for both genders in complex ways that are not captured by the causal model. This complicates the estimation of the real-world impact of mediator changes on the gender pay gap. Beyond the empirical example, this article concludes that the interventional effects framework – as discussed in various contexts, including mediation – is a useful tool in social inequality research. In particular, the identification of targets for reducing social inequalities is a useful application of the framework.

Acknowledgement: The author is grateful for the reviewer’s valuable comments, which improved the manuscript, and to Josef Brüderl and Michael Schomaker for their feedback on the early draft of the manuscript.

Funding information: The author states no funding involved.

Author contribution: The author confirms the sole responsibility for the conception of the study, presented results, and prepared the manuscript.

Conflict of interest: The author states no conflict of interest.

Ethical approval: The author used secondary data from the German Socio-Economic Panel (SOEP). The SOEP places the highest priority on protecting the confidentiality of respondents’ data by ensuring strict adherence

to European and German data protection regulations, consistent with the principles of the Helsinki Declaration.

Data availability statement: The data that support the findings of this study are available from the German Institute for Economic Research (DIW Berlin) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available from the author upon reasonable request and with permission of the DIW Berlin.

References

- [1] Neckerman K. Social inequality. New York: Russell Sage Foundation; 2004.
- [2] Schmitt L, Auspurg K. A stall only on the surface? Working hours and the persistence of the gender wage gap in Western Germany 1985–2014. *Europ Sociol Rev.* 2022 Feb;38(5):754–69.
- [3] Bobbitt-Zeher D. The gender income gap and the role of education. *Sociol Educ.* 2007;80(1):1–22.
- [4] Goldin C. A grand gender convergence: its last chapter. *Amer Econ Rev.* 2014;104(4):1091–119.
- [5] Blau FD, Kahn LM. The Gender Wage Gap: Extent, Trends, and Explanations. *J Econ Literature.* 2017;55(3):789–865.
- [6] Kunze A. The gender wage gap in developed countries. *The Oxford handbook of women and the economy.* New York: Oxford University Press; 2018. Vol. 4. p. 369–94.
- [7] Mischler F. A cause analysis based on the structure of earnings survey 2018. *WISTA - Wirtschaft und Statistik.* 2021;73(4):110–25.
- [8] Oaxaca RL. Male-female wage differentials in urban labor markets. *Int Econ Rev.* 1973;14(3):693–709.
- [9] Blinder AS. Wage discrimination: reduced form and structural estimates. *J Human Resources.* 1973;8(4):436–55.
- [10] Jann B. The Blinder-Oaxaca decomposition for linear regression models. *Stata J.* 2008;8(4):453–79.
- [11] Hlavac M. Oaxaca: Blinder-Oaxaca decomposition in R. R package version 0.1.4. 2018. <https://cran.r-project.org/web/packages/oaxaca/index.html>.
- [12] Sen B. Using the Oaxaca-Blinder decomposition as an empirical tool to analyze racial disparities in obesity. *Obesity.* 2014;22(7):1750–5.
- [13] Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology.* 2014;25(3):418–26.
- [14] Lundberg I, Johnson R, Stewart BM. What is your estimand? Defining the target quantity connects statistical evidence to theory. *Amer Sociol Rev.* 2021;86(3):532–65.
- [15] Huber M. Causal pitfalls in the decomposition of wage gaps. *J Business Econ Stat.* 2015;33(2):179–91.
- [16] Fortin NM, Lemieux T, Firpo S. Chapter 1 - Decomposition methods in economics. In: Ashenfelter O, Card D, editors. *Handbook of labor economics.* Vol. 4. Amsterdam: Elsevier; 2011. p. 1–102.
- [17] VanderWeele TJ, Vansteelandt S, Robins JM. Effect decomposition in the presence of an exposure-induced mediator–outcome confounder. *Epidemiology.* 2014;25(2):300–6.
- [18] Jackson JW, VanderWeele TJ. Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology.* 2018;29(6):825–35.
- [19] Vansteelandt S, Daniel RM. Interventional effects for mediation analysis with multiple mediators. *Epidemiology.* 2017;28(2):258–65.
- [20] Moreno-Betancur M, Moran P, Becker D, Patton GC, Carlin JB. Mediation effects that emulate a target randomised trial: Simulation-based evaluation of ill-defined interventions on multiple mediators. *Stat Methods Med Res.* 2021;30(6):1395–412.
- [21] Nguyen TQ, Schmid I, Ogburn EL, Stuart EA. Clarifying causal mediation analysis: Effect identification via three assumptions and five potential outcomes. *J Causal Infer.* 2022;10(1):246–79.
- [22] Miles CH. On the causal interpretation of randomised interventional indirect effects. *J R Stat Soc Ser B Stat Methodol.* 2023 June;85:1154–72.
- [23] Nguyen TQ, Schmid I, Stuart EA. Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychol Methods.* 2021;26(2):255–71.
- [24] VanderWeele TJ, Robinson WR. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology.* 2014;25(4):473–84.
- [25] Judd CM, Kenny DA. Process analysis: Estimating mediation in treatment evaluations. *Evaluat Rev.* 1981;5(5):602–19.
- [26] Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Personality Soc Psychol.* 1986;51(6):1173–82.
- [27] Valeri L, VanderWeele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol Methods.* 2013;18(2):137–50.
- [28] Shpitser I. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Sci.* 2013 July;37:37.

- [29] VanderWeele TJ. Commentary: On causes, causal inference, and potential outcomes. *Int J Epidemiol*. 2016;45(6):1809–16.
- [30] Didelez V, Dawid AP, Geneletti S. Direct and indirect effects of sequential treatments. In: *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. UAI'06. Arlington, Virginia, United States: AUAI Press; 2006. p. 138–46.
- [31] VanderWeele TJ, Tchetgen Tchetgen EJ. Mediation analysis with time varying exposures and mediators. *J R Stat Soc Ser B (Stat Methodol)*. 2017;79(3):917–38.
- [32] Blaikie K, Eisenberg-Guyot J, Andrea SB, Owens S, Minh A, Keil AP, et al. Differential employment quality and educational inequities in mental health: A causal mediation analysis. *Epidemiology*. 2023;34(5):747–58.
- [33] Rudolph KE, Gimbrone C, Díaz I. Helped into harm: Mediation of a housing voucher intervention on mental health and substance use in boys. *Epidemiology (Cambridge, Mass)*. 2021;32(3):336.
- [34] Didden C, Egger M, Folb N, Maartens G, Rohner E, Kassanjee R, et al. The contribution of non-communicable and infectious diseases to the effect of depression on mortality: a longitudinal causal mediation analysis. *Epidemiology*. 2024;36(1):88–98.
- [35] SOEP. Socio-Economic Panel (SOEP), version 35, data for years 1984–2018 (SOEP-Core v35); 2019.
- [36] Goebel J, Grabka MM, Liebig S, Kroh M, Richter D, Schröder C, et al. The German Socio-Economic Panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik*. 2019;239(2):345–60.
- [37] Becker GS. Investment in human capital: a theoretical analysis. *J Politic Econ*. 1962;70(5):9–49.
- [38] Pearl J. *Causality*. Cambridge: Cambridge University Press; 2009.
- [39] Loh WW, Moerkerke B, Loeyts T, Vansteelandt S. Heterogeneous indirect effects for multiple mediators using interventional effect models. *Epidemiol Methods*. 2021;9(1):20200023.
- [40] VanderWeele TJ. *Explanation in causal inference: methods for mediation and interaction*. New York: Oxford University Press; 2015.
- [41] Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3(2):143–55.
- [42] Pearl J. Direct and indirect effects. In: *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*. UAI '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2001. p. 411–20.
- [43] Moreno-Betancur M, Carlin JB. Understanding interventional effects: A more natural approach to mediation analysis? *Epidemiology*. 2018;29(5):614–7.
- [44] Avin C, Shpitser I, Pearl J. Identifiability of path-specific effects. In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. IJCAI'05. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2005. p. 357–63.
- [45] Zheng W, van der Laan MJ. Longitudinal mediation analysis with time-varying mediators and exposures, with application to survival outcomes. *J Causal Inference*. 2017;5:20160006.
- [46] Díaz I, Williams N, Rudolph KE. Efficient and flexible mediation analysis with time-varying mediators, treatments, and confounders. *J Causal Inference*. 2023;11(1):20220077.
- [47] Lin SH, Young JG, Logan R, VanderWeele TJ. Mediation analysis for a survival outcome with time-varying exposures, mediators, and confounders. *Stat Med*. 2017;36(26):4153–66.
- [48] VanderWeele TJ, Vansteelandt S. Mediation analysis with multiple mediators. *Epidemiol Methods*. 2014 Jan;2:95–115.
- [49] Tchetgen Tchetgen EJ, VanderWeele TJ. On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology*. 2014;25(2):282–91.
- [50] Vanderweele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Stat Interface*. 2009;2(4):457–68.
- [51] VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. *Amer J Epidemiol*. 2010;172(12):1339–48.
- [52] Tchetgen Tchetgen EJ. A note on formulae for causal mediation analysis in an odds ratio context. *Epidemiol Methods*. 2014;2(1):21–31.
- [53] Tchetgen Tchetgen EJ, Shpitser I. Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Ann Stat*. 2012;40(3):1816–45.
- [54] Rubinstein M, Branson Z, Kennedy EH. Heterogeneous interventional effects with multiple mediators: Semiparametric and non-parametric approaches. *J Causal Inference*. 2023;11(1):20220070.
- [55] Holland PW. Statistics and causal inference. *J Amer Stat Assoc*. 1986;81(396):945–60.
- [56] VanderWeele TJ, Hernán MA. Causal effects and natural laws: Towards a conceptualization of causal counterfactuals for non-manipulable exposures, with application to the effects of race and sex. In: *Hoboken, NJ: Causality*. John Wiley & Sons, Ltd; 2012. p. 101–13.
- [57] Glymour C, Glymour MR. Commentary: race and sex are causes. *Epidemiology*. 2014;25(4):488–90.
- [58] Glymour MM, Spiegelman D. Evaluating public health interventions: 5. Causal inference in public health research—do sex, race, and biological factors cause health outcomes? *Amer J Public Health*. 2017;107(1):81–5.
- [59] Pearl J. Does obesity shorten life? Or is it the Soda? On non-manipulable causes. *J Causal Inference*. 2018;6(2):20182001.
- [60] Fienberg SE, Haviland AM. Discussion of statistics and causal inference: A review. *Test*. 2003;12:319–27.
- [61] Kaufman JS. Epidemiologic analysis of racial/ethnic disparities: Some fundamental issues and a cautionary example. *Soc Sci Med*. 2008;66(8):1659–69.
- [62] Greiner DJ, Rubin DB. Causal effects of perceived immutable characteristics. *Rev Econ Stat*. 2011;93(3):775–85.
- [63] Goldin C, Rouse C. Orchestrating impartiality: The impact of “blind” auditions on female musicians. *Amer Econ Rev*. 2000;90(4):715–41.

- [64] Magnusson C. Why is there a gender wage gap according to occupational prestige?: An analysis of the gender wage gap by occupational prestige and family obligations in Sweden. *Acta Sociol.* 2010;53(2):99–117.
- [65] Becker GS. Age, earnings, wealth, and human capital. In: *Human capital: a theoretical and empirical analysis, with special reference to education*, Second edition. NBER. New York: The University of Chicago Press; 1975. p. 214–30.
- [66] Goldin C. The quiet revolution that transformed women's employment, education, and family. *Amer Econ Rev.* 2006 May;96(2):1–21.
- [67] Algan Y, Dustmann C, Glitz A, Manning A. The economic situation of first and second-generation immigrants in France, Germany and the United Kingdom. *Econ J.* 2010;120(542):F4–30.
- [68] Kleven H, Landais C, Posch J, Steinhauer A, Zweimüller J. Child penalties across countries: Evidence and explanations. *AEA Papers Proc.* 2019;109:122–26.
- [69] Correll SJ, Benard S, Paik I. Getting a job: is there a motherhood penalty? *Amer J Sociol.* 2007;112(5):1297–338.
- [70] Hipp L. Do hiring practices penalize women and benefit men for having children? Experimental evidence from Germany. *Europ Sociol Rev.* 2019 11;36(2):250–264.
- [71] SOEP Group. SOEP-Core v34 – PGEN: Person-Related Status and Generated Variables. SOEP Survey Papers 758: Series D – Variable Descriptions and Coding. Berlin: DIW Berlin/SOEP; 2019.
- [72] Brenke K, Müller KU. Gesetzlicher Mindestlohn: kein verteilungspolitisches Allheilmittel. *DIW-Wochenbericht.* 2013;80(39):3–17.
- [73] Grabka MM. SOEP-Core v38 - Codebook for the \$PEQUIV File 1984-2020: CNEF Variables with Extended Income Information for the SOEP. Berlin: DIW Berlin/SOEP; 2024. Technical Report 1333.
- [74] (Destatis) SB. Gender Pay Gap; 2024. Accessed: 2024-01-15. https://www.destatis.de/EN/Themes/Labour/Labour-Market/Quality-Employment/Dimension1/1_5_GenderPayGap.html.
- [75] Schomaker M, Leroy V, Wolfs T, Technau KG, Renner L, Judd A, et al. Optimal timing of antiretroviral treatment initiation in HIV-positive children and adolescents: a multiregional analysis from Southern Africa, West Africa and Europe. *Int J Epidemiol.* 2017;46(2):453–65.
- [76] Wooldridge JM. *Econometric analysis of cross section and panel data*. Vol. 1 of MIT Press Books. Cambridge: The MIT Press; 2010.
- [77] Díaz I. Non-agency interventions for causal mediation in the presence of intermediate confounding. *J R Stat Soc Ser B Stat Methodol.* 2024;86(2):435–60.
- [78] Ehrlich U, Möhring K, Drobnič S. What comes after caring? The impact of family care on women's employment. *J Family Issues.* 2020;41(9):1387–419.
- [79] Hirsch BT. Why do part-time workers earn less? the role of worker and job skills. *Industr Labor Relations Rev.* 2005;58(4):525–51.
- [80] Russo G, Hassink W. The part-time wage gap: a career perspective. *De Economist.* 2008;156:145–74.
- [81] O'Dorchai S, Plasman R, Rycx F. The part-time wage penalty in european countries: how large is it for men? *Int J Manpower.* 2007 Feb;28:571–603.
- [82] Wolf E. The German part-time wage gap: Bad news for men? SOEP paper No 663. 2014.
- [83] Nightingale M. Looking beyond average earnings: why are male and female part-time employees in the UK more likely to be low paid than their full-time counterparts? *Work, Employment and Society.* 2019;33(1):131–48.
- [84] Antonczyk D, Fitzenberger B, Sommerfeld K. Rising wage inequality, the decline of collective bargaining, and the gender wage gap. *Labour Econ.* 2010;17(5):835–47.
- [85] Babcock L, Laschever S. *Women don't ask: negotiation and the gender divide*. Princeton: Princeton University Press; 2003.
- [86] Dittrich M, Knabe A, Leipold K. Gender differences in experimental wage negotiations. *Econ Inquiry.* 2014;52(2):862–73.
- [87] Exley CL, Niederle M, Vesterlund L. Knowing when to ask: The cost of leaning in. *J Politic Econ.* 2020;128(3):816–54.
- [88] Lin SH, VanderWeele T. Interventional approach for path-specific effects. *J Causal Inference.* 2017;5(1):20150027.

Appendix

A Structural assumptions including unmeasured common causes of the exposure, intermediates and the outcome

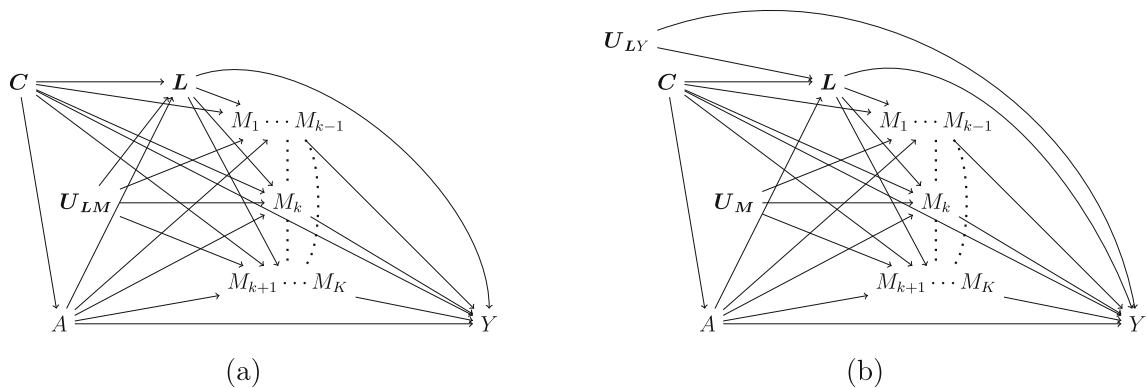


Figure A1: Structural assumptions encompassing unmeasured common causes of the intermediate factors in L and M (U_{LM}) (a), or unmeasured common causes of L and Y (U_{LY}), and of the mediators in M (U_M) (b), under which assumptions A1–A3 (main text) hold.

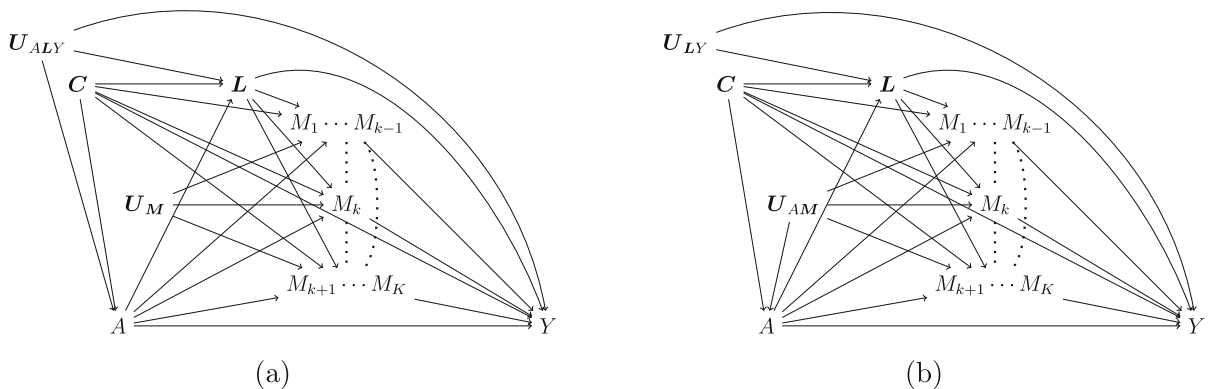


Figure A2: Structural assumptions encompassing unmeasured common causes of A , L and Y (U_{ALY}), and of the mediators in M (U_M) (a), or of A and M (U_{AM}), and L and Y (U_{LY}) (b), under which $IE_{I(C(ops))}$ (equation (5)) is identified.

B Alternative indirect effect measures in the presence of exposure-induced confounders

When operating under the DAG in Figure 1 (a) of the main text, one potential approach is to consider all intermediate factors of the causal model, i.e., L and M , in a single mediator set to evaluate the natural indirect effect (NIE) through L and M , as proposed by VanderWeele and Vansteelandt [48] and VanderWeele et al. [17]. This effect would be equivalent to the interventional indirect effect through L and M , since all exposure-induced confounders are included in the mediator set [22]. However, considering all intermediates jointly is not a suitable approach for investigating indirect pathways through specific mediators. Consequently, this approach does not address Question 1 in the main text. An alternative option could be to consider the path-specific indirect effect $A \rightarrow M \rightarrow Y$, which does not involve the paths from A to M via exposure-induced confounders L [44]. It can be defined as an interventional path-specific indirect effect where the intervention distribution of M is conditional on C and $L_0 = l$ [21,22,45]. This effect offers a mediational interpretation, but is not an interventional analogue to the

NIE through all mediators in $\mathbf{M}$, which includes $A \rightarrow \mathbf{M} \rightarrow Y$ and $A \rightarrow \mathbf{L} \rightarrow \mathbf{M} \rightarrow Y$ [22]. Path-specific indirect effects through a proper mediator subset $\mathbf{I}$, which do not involve the pathways from A to $\mathbf{I}$ via intermediates prior to $\mathbf{I}$, necessitate specifying the causal order among $\mathbf{I}$ and the other mediators in $\mathbf{R}$ [19,88]. Thus, they are not alternatives to the estimand targeted by Question 1 when the causal order among the mediators is unspecified.

C Identification results

Nguyen et al. [21] provide identification proofs for counterfactual outcomes when the mediator intervention distribution is a counterfactual distribution, including distributions that are either marginal or conditional on covariates $\mathbf{C}$, or on $\mathbf{C}$ and $\mathbf{L}_a$. The results can be applied to the identification of $\text{IE}_{I|C} = E[Y_{1\bar{I}|C} - Y_{1\bar{I}|C}\bar{R}_{1|C}]$ (equation (1)), $\text{IE}_{I|C} = E[Y_{1M_1} - Y_{1\bar{I}|C}\bar{R}_{1|C,L_1}]$ (equation (3)) of the main text. Thus, only the derivations of the intervention distributions $P_{I|C}(\mathbf{i}|\mathbf{C})$ and $P_{R|C,L_1}(\mathbf{r}|\mathbf{C}, \mathbf{L}_1)$ are provided here. These derivations closely follow those in Nguyen et al. [21]. As a notation reminder, $\mathbf{I}$ and $\mathbf{R}$ are subsets of $\mathbf{M}$, i.e., $\mathbf{M} = \mathbf{I} \cup \mathbf{R} = (M_1, \dots, M_K)$.

For any possible mediator subset $\mathbf{I} \subseteq \mathbf{M}$, we have for $P_{I|C}(\mathbf{i}|\mathbf{C}) = P(I_a = \mathbf{i}|\mathbf{C})$, $a \in \{0, 1\}$:

$$P(I_a = \mathbf{i}|\mathbf{C}) = P(I_a = \mathbf{i}|\mathbf{C}, A = a) \quad \text{conditional independence } (M_{1a}, \dots, M_{Ka}) \perp\!\!\!\perp A|\mathbf{C}, \quad (\text{A1})$$

$$= P(\mathbf{I} = \mathbf{i}|\mathbf{C}, A = a) \quad \text{consistency and positivity.} \quad (\text{A2})$$

By the weak union rule of conditional independence, the assumption $(\mathbf{I}_a, M_{1a}, \dots, M_{Ka}) \perp\!\!\!\perp A|\mathbf{C}$ implies that $(M_{1a}, \dots, M_{Ka}) \perp\!\!\!\perp A|\mathbf{C}, \mathbf{L}_a$ [21]. Thus, for $P_{R|C,L_1}(\mathbf{r}|\mathbf{C}, \mathbf{L}_1) = P(\mathbf{R}_1 = \mathbf{r}|\mathbf{C}, \mathbf{L}_1)$, we have:

$$P(\mathbf{R}_1 = \mathbf{r}|\mathbf{C}, \mathbf{L}_1) = P(\mathbf{R}_1 = \mathbf{r}|\mathbf{C}, \mathbf{L}_1, A = 1) \quad \text{conditional independence } (M_{1a}, \dots, M_{Ka}) \perp\!\!\!\perp A|\mathbf{C}, \mathbf{L}_a \quad (\text{A3})$$

$$= P(\mathbf{R} = \mathbf{r}|\mathbf{C}, \mathbf{L} = \mathbf{I}, A = 1) \quad \text{consistency and positivity.} \quad (\text{A4})$$

Identification results for $E[Y_{1\bar{I}|A=0, C\bar{R}}|A=1, C, L|A=1]$ in $\text{IE}_{I|C(\text{obs})}$ (equation (5)) build on the proofs in Jackson and VanderWeele [18] and Nguyen et al. [21]. Here, the counterfactual outcome is defined to be conditional on $A = 1$ and the mediator intervention distributions are functions of the observed data distribution.

$$E[Y_{1\bar{I}|A=0, C\bar{R}}|A=1, C, L|A=1] = E(E\{E[Y_{1\bar{I}|A=0, C\bar{R}}|A=1, C, L|A=1, \mathbf{C}, \mathbf{L}]|A=1, \mathbf{C}|A=1\}) \quad (\text{A5})$$

iterated expectation

$$= \sum_c \sum_l \sum_m E[Y_{ir}|A=1, \mathbf{c}, \mathbf{l}] P(\mathbf{i}|\mathbf{A}=0, \mathbf{c}) P(\mathbf{r}|\mathbf{A}=1, \mathbf{c}, \mathbf{l}) P(\mathbf{l}|\mathbf{A}=1, \mathbf{c}) P(\mathbf{c}|\mathbf{A}=1) \quad (\text{A6})$$

transition to summation¹¹

$$= \sum_c \sum_l \sum_m E[Y_{ir}|A=1, \mathbf{c}, \mathbf{l}, \mathbf{m}] P(\mathbf{i}|\mathbf{A}=0, \mathbf{c}) P(\mathbf{r}|\mathbf{A}=1, \mathbf{c}, \mathbf{l}) P(\mathbf{l}|\mathbf{A}=1, \mathbf{c}) P(\mathbf{c}|\mathbf{A}=1) \quad (\text{A7})$$

conditional independence $Y_{m_1, \dots, m_K} \perp\!\!\!\perp (M_1, \dots, M_K) | \{A=1, \mathbf{C}, \mathbf{L}\}$

$$= \sum_c \sum_l \sum_m E[Y|A=1, \mathbf{c}, \mathbf{l}, \mathbf{m}] P(\mathbf{i}|\mathbf{A}=0, \mathbf{c}) P(\mathbf{r}|\mathbf{A}=1, \mathbf{c}, \mathbf{l}) P(\mathbf{l}|\mathbf{A}=1, \mathbf{c}) P(\mathbf{c}|\mathbf{A}=1) \quad (\text{A8})$$

consistency and positivity.

For $E[M_0]$ in $\text{IE}_M = E[Y_1 - Y_{1E[M_0]}]$ (equation (9)), we have: $E[M_0] = E[M_0|A=0]$ by the independence assumption $M_a \perp\!\!\!\perp A$, and $E[M_0|A=0] = E[M|A=0]$ by consistency and positivity.

¹¹ Using short notation; for instance, $E[Y_{ir}|A=1, \mathbf{C} = \mathbf{c}, \mathbf{L} = \mathbf{l}] = E[Y_{ir}|A=1, \mathbf{c}, \mathbf{l}]$. For simplicity, the discrete case is considered; if continuous variables are involved, the transition would be to integration instead of summation.

D MC g-computation for the estimation of $E[Y_{a\tilde{a}^*|c}\tilde{R}_{a|c}]$

Table A1: Proposed algorithm to estimate $E[Y_{a\tilde{a}^*|c}\tilde{R}_{a|c}]$, based on the estimator described by Vansteelandt and Daniel [19]

Algorithm to estimate $E[Y_{a\tilde{a}^* c}\tilde{R}_{a c}]$	
Step 1	Specify and fit models for: 1) the distribution of $\mathbf{L}$ conditional on A and C (Model 1), 2) the distribution of $\mathbf{R}$ conditional on A and C (Model 2), 3) the distribution of $\mathbf{I}$ conditional on $A = a^*$ and C (Model 3), 4) the outcome Y conditional on $A, \mathbf{M}, \mathbf{L}$ and C (Model 4).
Step 2	Set the distribution of C to the respective empirical distribution.
Step 3	Set $A = a$.
Step 4	Set the variables in $\mathbf{L}$ to a random draw from Model 1 fitted in Step 1.
Step 5	Set the variables in $\mathbf{R}$ to a random draw from Model 2 fitted in Step 1.
Step 6	Set the variables in $\mathbf{I}$ to a random draw from Model 3 fitted in Step 1.
Step 7	Draw the outcome from Model 4 fitted in Step 1 based on the updated data generated through Steps 2–6.
Step 8	Estimate $E[Y_{a\tilde{a}^* c}\tilde{R}_{a c}]$ by calculating the mean of Y .
Step 9	Repeat steps 4–8 Z times with different seeds and calculate the mean of the Z estimates (to reduce Monte Carlo error).
Step 10	To obtain the nonparametric bootstrap standard error and percentile confidence intervals, repeat Steps 1–9 B times on the bootstrapped data.

E Application to the gender pay gap

E.1 Structural assumptions

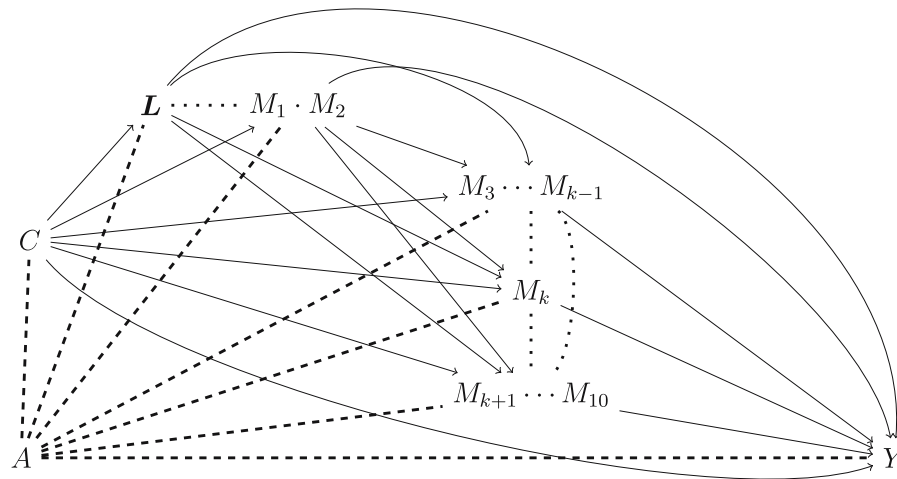


Figure A3: Gender pay gap: Assumptions regarding the causal influences among variables for a specific time period (here, the year 2017). Loosely dotted lines among the potential mediators signify that the structural dependence between these mediators remains unspecified. Dashed lines from A (female/male category) to other factors emphasize that causal effects of A are not addressed in the application of Approach 3.

Table A2: Targets of interventions, causal estimands, required assumptions, and interpretation when situating the results obtained for Aims 1a and 1b within Approaches (A) 1–3

Aims 1a-b				
A	Intervention on	Causal estimand	Assumptions	Interpretation
1	Gender, mediators	$\frac{E[Y_1 - Y_{1\bar{R} C,L_1}]}{TE} \times 100$	Consistency, Positivity,	% of TE mediated by all pathways from being assigned to the female category to wage through I , except those that pass through M_1, M_2 (which are prior to I), and those that pass through descendants of I in R .
		$\tilde{I}$ from	$Y_{am} \perp\!\!\!\perp A, Y_{am} \perp\!\!\!\perp \mathbf{M} \{A = a, C, L\}, M_a \perp\!\!\!\perp A$ [19], No unmeasured confounders of I and M_1, M_2 given A and C ,	
		$P(I_0 = i A = 0, C, M_1, M_2)$	Indirect effect measure criteria [22], Correct model specification	
		$\tilde{R}$ from		
		$P(R_1 = r A = 1, C, L)$		
2	Gender, mediators	$\frac{E[Y_1 - Y_{1\bar{R} C,L_1}]}{TE} \times 100$	Consistency, Positivity,	% of TE reduced by shifting the counterfactual distribution of I under the female category to match that under the male category, given C and M_1, M_2 , while holding the counterfactual distribution of R under the female category constant, given C and L_1 .
		$\tilde{I}$ from	$Y_{am} \perp\!\!\!\perp A, Y_{am} \perp\!\!\!\perp \mathbf{M} \{A = a, C, L\}, M_a \perp\!\!\!\perp A$ [19],	
		$P(I_0 = i A = 0, C, M_1, M_2)$	Correct model specification	
		$\tilde{R}$ from		
		$P(R_1 = r A = 1, C, L)$		
3	Mediators	$\frac{E[Y - Y_{\bar{R} A=1,C,L} A = 1]}{GPG} \times 100$	Consistency, Positivity,	% of GPG reduced by setting the distribution of I among women equal to that observed among men, given C and M_1, M_2 , while holding the distribution of R among women constant, given C and L .
		$\tilde{I}$ from	$Y_{am} \perp\!\!\!\perp \mathbf{M} \{A = a, C, L\}$ [18],	
		$P(I = i A = 0, C, M_1, M_2)$	Correct model specification	
		$\tilde{R}$ from		
		$P(R = r A = 1, C, L)$		

GPG denotes the observed marginal gender pay gap, defined as $E[Y|A = 1] - E[Y|A = 0]$, TE denotes the total effect of being assigned to the female category on wage, defined as $E[Y_1 - Y_0]$.

Table A3: Targets of interventions, causal estimands, required assumptions, and interpretation when situating the results obtained for Aims 2a and 2b within Approaches (A) 2–3

Aims 2a-b				
A	Intervention on	Causal estimand	Assumptions	Interpretation
2	Gender, mediators	$1 - \frac{E[Y_{TR 1,C,L}] - E[Y_{TR 0,C,L}]}{TE} (\times 100)$ $\tilde{I}$ from $P_I(\mathbf{i})$ (observed distribution of I in full sample)	Consistency, Positivity, $Y_{am} \perp\!\!\!\perp A$, $Y_{am} \perp\!\!\!\perp \mathbf{M} \{A = a, C, \mathbf{L}\}$ [41,42], $M_a \perp\!\!\!\perp A$, Correct model specification	% of TE reduced (“portion eliminated” [41]) by aligning the counterfactual distribution of I under the female and male categories with the observed sample distribution, while holding the counterfactual distribution of R constant, given C and L .
3	Mediators	$1 - \frac{E[Y_{TR 1,C,L} A = 1] - E[Y_{TR 0,C,L} A = 0]}{GPG} (\times 100)$ $\tilde{I}$ from $P_I(\mathbf{i})$ (observed distribution of I in full sample)	Consistency, Positivity, $Y_{am} \perp\!\!\!\perp \mathbf{M} \{A = a, C, \mathbf{L}\}$ [18], Correct model specification	% of GPG reduced by setting the distribution of I among women and among men equal to the observed sample distribution, while keeping the distribution of R among women and men as observed, given C and L .

GPG denotes the observed marginal gender pay gap, defined as $E[Y|A = 1] - E[Y|A = 0]$, TE denotes the total effect of being assigned to the female category on wage, defined as $E[Y_1 - Y_0]$.

E.2 Situating results within Approaches 1–3: Conditions and suggested interpretation

Tables A2 and Table A3 summarize the essential conditions under which the effects obtained for Aims 1a and 1b, as well as for Aims 2a and 2b of the applied analysis described in the main text (Sections 5.4.1 and 5.4.2), may represent specific target quantities within Approaches 1–3 (Sections 3.1–3.3). At this point, it is up to the reader to decide whether or not causal effects related to gender are meaningful in this context. As a reminder, the mediators M_1 and M_2 represent education levels (M_1 : ≥ 12 years of education, M_2 : college degree or higher). They are assumed to precede the other mediators, which include job characteristics and work experience (Figure A3). In Aim 1a, $\mathbf{I} = (M_3, \dots, M_{10})$, with $\mathbf{R} = (M_1, M_2)$. In Aim 1b, a separate analysis is conducted for each mediator M_j , where $\mathbf{I} = M_j$ for $j = 3, \dots, 10$. In Aim 2a, the analysis considers all mediators together, with $\mathbf{I} = (M_1, \dots, M_{10})$ and $\mathbf{R} = \emptyset$. In Aim 2b, a separate analysis is conducted for each mediator M_k , where $\mathbf{I} = M_k$ for $k = 1, \dots, 10$. C denotes age in years, L_1 indicates the presence of at least one child in the household, and L_2 indicates having a direct migration background. A is a binary indicator labeled “Gender”, where $A = 1$ denotes being female and $A = 0$ denotes being male. With respect to Approaches 1 and 2, it is assumed that A is independent of C . In Aims 1a and 1b, the intervention distribution of $\mathbf{I}$, from which the random draws $\tilde{\mathbf{I}}$ are obtained, is specified as

$$P_{\mathbf{I}|A=0, C, M_1, M_2}(\tilde{\mathbf{i}}|A = 0, C, M_1, M_2).$$

In both Aims 2a and 2b, the intervention distribution of $\mathbf{I}$ is specified as that observed in the total sample population:

$$P_{\mathbf{I}}(\mathbf{i}).$$

For all Aims, the joint distribution of the other mediators, summarized in $\mathbf{R}$, is held constant, given C and $\mathbf{L} = (L_1, L_2)$. For notational simplicity, let GPG represent the observed marginal gender pay gap, defined as $E[Y|A = 1] - E[Y|A = 0]$. Let TE denote the total effect on wage of being assigned to the female category, defined as $E[Y_1 - Y_0]$.

E.3 Detailed results for Aims 1b and 2b

Aim 1b: Which intervention would yield the largest reduction when setting the distribution of a single mediator M_j ($j = 3, \dots, 10$) in women equal to that in men, given age and educational background, while keeping the joint distribution of the other mediators in women constant as observed, given C and $\mathbf{L}$? (Table A4)

Table A4: Aim 1b. Percentage reductions in gender pay gap in log gross hourly wages, achieved by single-mediator interventions in women, with 95% bootstrap CIs

Intervention target	% Reduction in disparity in Y with 95% CI
M_3 : $\geq$ College degree job	3.6 [1.3; 6.3]
M_4 : ≥ 6.7 years in company	5.9 [4.6; 7.5]
M_5 : Female-dominated industry	7.4 [3.3; 11.3]
M_6 : Leading position	7.4 [6.0; 9.6]
M_7 : Full-time employment	6.3 [1.2; 11.4]
M_8 : Flexible working hours	−2.3 [−3.9; −0.6]
M_9 : Work experience	26.5 [20.9; 32.6]
M_{10} : Job prestige (SIOPS)	4.6 [2.3; 6.8]

A positive value indicates a reduction, while a negative value indicates an increase in the gender pay gap.

Aim 2b: Which intervention would yield the largest reduction when setting the distribution of a single mediator M_k ($k = 1, \dots, 10$) among women and among men equal to that in the total sample population, while keeping the joint distribution of the other mediators in women and men constant as observed, given C and L ? (Table A5)

Table A5: Aim 2b. Percentage reductions in gender pay gap in log gross hourly wages, achieved by single-mediator interventions in men and in women, with 95% bootstrap CIs

Intervention target	% Reduction in disparity in Y	% Reduction in disparity due to change in Y in men	% Reduction in disparity due to change in Y in women
M_1 : ≥ 12 years of education	-7.6 [-14.9; 0.7]	-6.8 [-13.8; 0.8]	-0.7 [-3.1; 2.2]
M_2 : $\geq$ College degree	-5.6 [-14.1; 3.5]	-4.3 [-12.3; 3.2]	-1.3 [-5.34; 2.6]
M_3 : $\geq$ College degree job	8.6 [0.3; 17.7]	7.4 [2.4; 12.8]	1.2 [-5.0; 7.6]
M_4 : ≥ 6.7 years in company	4.2 [-0.1; 8.2]	3.8 [-0.4; 6.7]	0.3 [-2.1; 2.8]
M_5 : Female-dominated industry	10.8 [7.7; 14.1]	7.8 [5.8; 9.7]	3.0 [0.8; 5.6]
M_6 : Leading position	10.3 [6.0; 14.2]	4.5 [1.3; 7.8]	5.8 [3.4; 8.2]
M_7 : Full-time employment	29.8 [20.2; 39.9]	26.7 [17.8; 36.4]	3.1 [0.5; 5.5]
M_8 : Flexible working hours	0.3 [-2.7; 3.5]	1.3 [-1.2; 4.0]	-1.0 [-2.7; 0.6]
M_9 : Work experience	21.5 [9.2; 35.3]	4.7 [-7.1; 16.0]	16.8 [9.6; 24.0]
M_{10} : Job prestige (SIOPS)	0.6 [-4.8; 6.0]	-2.3 [-5.1; 0.6]	2.9 [-1.8; 7.2]

A positive value indicates a reduction, while a negative value indicates an increase in the gender pay gap.