Soc Psychiatry Psychiatr Epidemiol (2017) 52:785–793
DOI 10.1007/s00127-017-1390-0

COMMENTARY

Explaining intersectionality through description, counterfactual thinking, and mediation analysis

John W. Jackson[1]

Received: 27 February 2017 / Accepted: 1 May 2017 / Published online: 24 May 2017

Introduction

I would like to thank Schwartz [1] for insightful questions about our paper [2] on whether and how quantitative approaches could be useful for understanding intersectionality, and I would also like to thank the Editor for providing an opportunity to respond. On many points, we agree. The additive joint disparity we considered is a policy-relevant measure that can be used to track the health of multiply marginalized populations, without any reference to interactions, and certainly this focus is consistent with intersectionality. Schwartz's emphasis that, when available, social theory should guide an analysis is an excellent one and is why we provided an extensive, though not exhaustive, review of additive interaction measures and modeling strategies. It would indeed be interesting to evaluate whether tests for synergism in the sufficient-cause framework [3, 4] could be mapped to concepts of intersectionality.

Many of Schwartz's questions revolve around the causal status of race, a topic which has been considerably debated in the statistics, social science, and more recently the epidemiology literature [5–11]. Contrary to Schwarz's assumption, our study did not consider race as a cause, nor did we consider potential outcomes for race. Early on, we made clear: "In attributing disparities to race, we mean the historical legacy of racism in the United States (US) taking shape through various means, including slavery, Jim Crow, and segregation [12]. This legacy involves discrimination which may be intentional or the result of policies, laws or practices that systematically disadvantage blacks and shape economic opportunities [13] (and also) indirect processes whereby blacks are more vulnerable to economic or political shifts because of this legacy" [14]. The reality of class in the US can reinforce this legacy by mediating access to quality neighborhoods, housing, and education. It is also important to remember that the joint decomposition originated long before potential outcomes was used to infer causal inference in epidemiology [15, 16], before any entrenched perspectives about restricting studies to well-defined interventions arose, so its use does not imply that paradigm. Our use of the joint decomposition focuses on describing and understanding how outcomes are patterned for a multiply marginalized group as compared to a non-marginalized group. It is not equipped to speak of the action of racism or socioeconomic disadvantage or their intersection at the person level; its focus is on understanding patterns among the population.

In what follows, I will explain how, in our descriptive implementation, the joint decomposition relates to certain features of intersectionality. The central point is that even though potential outcomes of race/SES were not considered, the notion of a disparity itself is inherently counterfactual, and this property helped us draw insights about disparities across multiple axes. This property is also what would encourage one to identify targets to reduce those disparities. Accordingly, I will demonstrate how a mediation analysis for the joint disparity and its decomposition can support this aim. I also outline how to repurpose existing software for mediation analysis to accomplish this aim, and also present non-parametric formulae that can be

This comment refers to the article available at doi:10.1007/s00127-016-1334-0.

✉ John W. Jackson
john.jackson@jhu.edu

[1] Departments of Epidemiology and Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

used to develop estimators for more general settings (e.g., when mediator-outcome confounders are affected by a social status).

Interaction and intersectionality

Intersectionality is an expansive theoretical framework [17]. Its wide scope concerns intersecting social positions of marginalization (e.g., blacks with low SES) as well as positions of power (e.g., blacks with high SES), intersecting social processes (e.g., how sexism operates in the context of social disadvantage), and social action (e.g., how interventions and policies to reduce racial disparities in unemployment can acknowledge and adapt to settings of socioeconomic advantage and disadvantage). A key tenet in all of these endeavors is to normalize a perspective on groups that may be typically overlooked when analyses or political action subsumes populations into broader categories such as race or class and, as a result, offer solutions that may not respond to the needs of certain groups—typically those at the nexus of marginalized statuses [18, 19].

One of our contributions was to emphasize that an excess intersectional disparity—a parameter that many quantitative investigations in intersectionality have estimated [20]—has the same form as a statistical interaction, and is thus a quantity of contrasts. The excess intersectional disparity is defined as the joint disparity minus the sum of the referent disparities. It would seem inconsistent to ask about the presence or magnitude of an excess intersectional disparity but not allow for the joint decomposition. But how are we to interpret the joint decomposition and, relatedly, measures of statistical interaction? Are they antithetical to intersectionality as Schwartz contends?

Counterfactual thinking and intersectionality

It may help to first review how we often conceptualize disparities. Even though disparities do not measure the effects of well-specified interventions, they are not without counterfactual meaning. If we consider excess rates of disease among disadvantaged groups as avoidable, then we can in some sense imagine an alternate world, however vague, where that disparity does not exist. That we can imagine a future without interpersonal discrimination, and further imagine one that remediates class-based structural legacies of racism that perpetuate disparities perhaps independently of interpersonal discrimination (e.g., the impact that real-estate tax-based funding has on the overrepresentation of low-income racial/ethnic minority children in lower quality schools), itself speaks to some sort of potential outcome. This is true regardless of whether we can enumerate what national and local policies and interventions are needed to advance social progress, or even predict potential barriers to them. It is true that disparities also reflect the effects of historical attempts to address them (in that they may be smaller than they once were), but again, we are still motivated to reduce even residual disparities that persist. Counterfactuals, even if ill-specified, seem central to the notion of disparities as unnecessary, avoidable, and unjust [21]. This perspective underpins the way the Institute of Medicine defines disparities in the receipt of healthcare [22]. But simply observing that a disparity exists does not tell us how to reduce it or what certain interventions might achieve. To make progress, we move from counterfactual thinking to counterfactual methods that conceptualize a finer model where a disparity in some outcome (e.g., wages) is perpetuated by disparities in, or the differential effects of, manipulable determinants of that outcome (e.g., educational attainment) [11, 23–25]. With counterfactual methods, we examine the extent to which an intervention on that determinant might reduce disparities (as in observational data) [26] or actually does so (as in a trial) [27, 28]. Although it may be possible to develop a formal counterfactual model to support these statements, this intuitive justification may suffice.

If we accept that disparities reflect the effects of a constellation of historical and current actions, policies, and other systematic forms of oppression [12], we can extend counterfactual thinking (and later, counterfactual methods) to examine the intersection of social statuses. The referent race disparity reflects effects of racism among those raised in families with greater socioeconomic resources (again, broadly defining racism to capture interpersonal and structural forms that disadvantage racial/ethnic minorities). One might point out that higher SES does not necessarily translate into similar gains for blacks and whites because of differences in social networks, or discrimination in housing and other spheres of life [29], but this could reasonably be construed as reflecting the legacy of racism. The referent SES disparity reflects the effects of socioeconomic disadvantage among those whose lineage did not endure the oppression of slavery, Jim Crow, and other structural effects of racism—but who, as Schwartz points out, also benefit from some form of white privilege [12]. If in fact the effects of racism on the outcome did not vary by childhood SES, and the effects of socioeconomic disadvantage did not vary by race, we would expect the joint disparity to equal the sum of the referent race disparity and referent SES disparity, and this would necessarily imply that the intersectional disparity equals zero. When the effects of racism are stronger in contexts of socioeconomic disadvantage than advantage, and the effects of socioeconomic disadvantage are stronger in the presence of racial oppression and/or its legacy than the presence of racial privilege, the excess intersectional disparity captures the differential.

Contrary to Schwartz's conjecture, this formulation is quite compatible with mutual constitutivism. Homogeneous effects of racism and socioeconomic disadvantage do not preclude heterogeneous manifestations of racism and socioeconomic disadvantage in the qualitative sense. However, the joint disparity can only exceed the sum of the referent when the effect of racism, whatever its form, leads to a higher rate of undesirable outcomes for those with low childhood SES and when the effect of socioeconomic disadvantage leads to a higher rate of undesirable outcomes among blacks. A positive excess intersectional disparity is a clear indication that intersectional processes are contributing to the joint disparity, even if we cannot explicitly describe what those processes are and how they operate. A positive excess intersectional disparity implies a quantitative manifestation of intersectionality. However, even when we cannot detect such manifestations, e.g., the excess intersectional disparity is null, the joint disparity can still be substantially larger than either referent disparity, and calling attention to this is still very much in line with the spirit of intersectionality.

Counterfactual methods and intersectionality

It is clear, then, that the joint disparity and its decomposition focus on understanding how outcomes are patterned, not how they are manifested or experienced. The decomposition unpacks dimensions of extent, but leaves unanswered questions of quality and kind, which may be unsatisfying. This is the purview of qualitative and mixed-methods research that can richly describe how intersectionality manifests in daily life, how it can guide and frame interventions and political action, and how it explains unique opportunities and barriers to health and well-being. Such data on target populations will be essential for developing interventions that are well suited for multiply marginalized groups of interest.

As we noted in our discussion, the potential outcomes framework could be leveraged to identify potential intervention targets to improve the health of multiply marginalized groups—a point Schwartz agrees with. Specifically, we noted that causal mediation analysis [11] could be used to examine how much the joint disparity in say, log-wages, comparing blacks with low SES to whites with high SES, would decrease after removing disparities in some target, say some test of pre-market skills obtained through education. Similar analyses examining racial disparities have garnered considerable attention in the field of labor economics. Fryer [30], extending analyses of Johnson and Neal [31], found in the 1997 National Longitudinal Survey of Youth [32] that controlling for a measure of pre-market skills (score percentiles from the Armed Forces Qualifying Test, AFQT scores) reduced male black–white disparities in 2006/2007 log-wages by seven units from −0.19 (0.02) to −0.12 (0.03). We can carry out a similar analysis for the joint disparity in log-wages by first fitting a model for log-wages (Y) conditional on race (A; black = 1, white = 0), dichotomized childhood SES (B; low = 1, high = 0), their product ($A*B$), and the covariate age (C) among non-Hispanic males:

$$E[Y|a,b,c] = \alpha_0 + \alpha_1 a + \alpha_2 b + \alpha_3 a*b + \alpha_5 c. \quad (1)$$

The quantity $\alpha_1 + \alpha_2 + \alpha_3$ can be interpreted as the joint disparity in log-wages comparing blacks with low SES to whites with high SES. We then can go on to re-fit this model with an additional term for AFQT scores M:

$$E[Y|a,b,m,c] = \beta_0 + \beta_1 a + \beta_2 b + \beta_3 a*b + \beta_4 m + \beta_5 c.$$

In the "Appendix 1", we show that under an assumption that the effect of AFQT scores on log-wages is unconfounded given race, childhood SES, ethnicity, gender, and age, the quantity $\beta_1 + \beta_2 + \beta_3$ can be interpreted as the residual joint disparity in log-wages that remains after an intervention to equalize the distribution of AFQT scores across race and childhood SES, i.e., to set the distribution of AFQT scores in the entire population to follow the distribution found among whites with high SES. Carrying this out (see "Appendix 2" for details on AFQT scores), we find that the joint disparity, comparing black males with low SES to white males with high SES, would reduce by eight units from −0.29 (0.10) to −0.21 (0.03). Though this query is inherently vague [33] (there are likely several ways to affect pre-market skills let alone to eradicate disparities in them), it does suggest that a large portion of the joint disparity in wages can be attributed to the joint disparity in pre-market skills. Future research could thus reasonably prioritize developing and evaluating explicit interventions to improve equity in pre-market skills.

Although our paper was focused on understanding the joint disparity, we can also examine the residual referent and excess intersectional disparities under an intervention to equalize test scores across race and childhood SES. We find that, upon controlling for AFQT scores, the referent race disparity comparing black males with high SES to white males with high SES reduces from $\alpha_1 = -0.16$ (0.04) to $\beta_1 = -0.11$ (0.04), a reduction of five units. Likewise, the referent SES disparity reduces by four units from $\alpha_2 = -0.16$ (0.03) to $\beta_2 = -0.12$ (0.03). We can also define the residual intersectional disparity as the difference between the residual joint disparity and sum of the residual referent disparities. The excess intersectional disparity remains null [$\alpha_3 = \beta_3 = 0.02$ (0.05)]. Applying the joint decomposition, the residual referent race and SES disparities are, respectively, 53 and 59% of the residual joint disparity's magnitude.

Overall, an intervention to equalize AFQT scores could potentially reduce the black–white wage gap among males by at least four units for all marginalized populations considered. Interestingly, even though the greatest potential gains (eight units) would occur for black males with low SES, their wages would not surpass the wages that black males with high SES or white males with low SES have before the intervention! A substantial joint disparity would remain. Moreover, it remains the case that both racism and socioeconomic disadvantage appear to contribute to the joint disparity; neither appears sufficient to fully account for it.

This example shows the clarity that the joint disparity and its decomposition can deliver in a mediation analysis, even when there is no evidence of an excess intersectional disparity. Readers may recall that for unemployment, there was a sizeable excess intersectional disparity that represented almost 40% of the joint disparity's magnitude. It would certainly be of scientific interest to examine how interventions to equalize test scores across race and class would change the excess intersectional disparity for unemployment, as this could be useful for advancing intersectionality theory. However, perhaps, what matters for policy making, from an intersectional perspective, is to examine how much intervening on potential targets might reduce the joint disparity and, moreover, to examine the absolute outcomes for multiply marginalized groups under such interventions.

Moving beyond this conceptual example, how is one to implement mediation analysis in practice? The formulae presented in the text only apply for linear outcomes when there are no further statistical interactions, and moreover, they require that the models for the outcome are correctly specified. In the "Appendix 1", I present non-parametric formulae that could be used to develop estimators for more complex settings. Fortunately, these formulae imply that we can appropriate existing software for mediation analysis on the additive scale that allows for further interactions and is appropriate even for non-linear outcomes [34]. It could be adapted to study the additive joint disparity by subsetting the data to the multiply marginalized group and non-marginalized group and using an appropriate indicator variable as the exposure (e.g., 1 = blacks with low SES, 0 = white with high SES). A similar strategy can be taken to study each referent disparity. The excess intersectional disparity could then be computed from these results as the difference between the residual joint disparity and the sum of the residual referent disparities, and its standard error could be obtained through bootstrapping (note that in the causal mediation analysis literature, the residual disparity is akin to the natural direct effect, and the disparity reduction is akin to the natural indirect effect). An alternative approach for non-linear outcomes would be to use software for mediation analysis on the multiplicative scale [35] to estimate the proportion by which the additive joint, referent, and intersectional disparities would change upon equalizing a target variable, and also estimate the proportion of the residual joint disparity for each residual referent disparity and the residual excess intersectional disparity. Unlike the parametric approach in the motivating example, these approaches rely on correctly specified models for the mediator and the outcome, and can handle statistical interactions between the joint social status (defined by race and childhood SES) and the mediator [34, 35]. Such interactions are plausible—and interesting in their own right—as they reflect the possibility that the effect of improving test scores on log-wages may differ across groups defined by race and childhood SES.

Though these approaches will be detailed in future work, it would be good to point out some key conceptual limitations and potential solutions. First, a mediation analysis as described here would only estimate how a disparity would change among those who share the same levels of covariates, and this may not always be of substantive interest. Emerging work in causal decomposition analysis [36] that overcomes this constraint for a single social status (e.g., race) could be extended to the case of multiple intersecting social statuses. Second, existing software for mediation analysis would assume that covariates are not affected by either social status, and could suffer from selection bias [37] if this assumption was violated. In the "Appendix 1", I present non-parametric formulae for mediation analysis of joint, referent, and excess intersectional disparities that can appropriately adjust for such covariates. Future work could use these formulae to develop flexible estimators in this setting. As a final reminder, any causal interpretation of a mediation analysis requires that all mediator-outcome confounders are measured and adjusted for. This is a very high standard and should guide the design of observational studies, particularly those that are expressly leveraged or conducted to advance our understanding of how to address health disparities.

Conclusion

There is no quantitative approach that can harness every dimension of intersectionality. What we have done is to provide a framework that in some instances speaks to quantitative manifestations of intersectionality, and hopefully, this will be used in concert with other qualitative approaches that describe the unique experiences of multiply marginalized groups in greater depth and nuance. An extension that we did not explore here would be to consider outcome differences defined not by the intersection of personal characteristics, but rather of those defined by the intersection higher level characteristics such as markers of institutional racism and/or neighborhood disadvantage [38]. This could capture the quantitative effects of intersecting power structures, and use counterfactual methods to understand how to reduce the risk among those who experience their intersection. By incorporating counterfactual thinking and allowing for further exploration through counterfactual methods, the joint decomposition provides an extensive framework for tracking the health of multiply marginalized groups and identifying potential opportunities to improve it.

Acknowledgements The author would like to thank Tyler J. VanderWeele for helpful comments on a draft of this manuscript. All the remaining errors belong to the author.

Compliance with ethical standards

Conflict of interest The author has no conflicts of interest to report.

Appendix 1: Mediation analysis formulae for two social statuses

Non-parametric results for causal mediation analysis

(In the absence of covariate feedback, i.e., mediator-outcome confounders not affected by race or childhood SES)

Consider some outcome Y among persons defined by variables race/ethnicity A (0 = non-Hispanic white, 1 = non-Hispanic black) and socioeconomic status B (0 = high, 1 = low), among some level of covariates gender and age $C = c$. Let M represent a measure of potentially manipulable characteristics in later life, e.g., measures of pre-market skills as reflected in test scores that may affect the outcome Y. Suppose that the effect of M on Y is unconfounded given A, B, and C, such that $E[Y(m)|a, b, m, c] = E[Y(m)|a, b, c]$ and also that consistency holds, such that $E[Y(m)|a,b,m,c] = E[Y|a,b,m,c]$.

The joint disparity that would remain if the distribution of test scores M for black persons with low SES ($A = 1$, $B = 1$) with covariates $C = c$ was set equal to its distribution for white persons with high SES ($A = 0$, $B = 0$) with $C = c$ would be

$$\mu_{m11} - E[Y|A = 0, B = 0, c],$$

and the joint disparity reduction would be

$$E[Y|\,A = 1, B = 1, c] - \mu_{m11},$$

where $\mu_{m11} = \Sigma_m E[Y|A=1,B=1,m,c]P(m|A=0,B=0,c)$.

The referent race disparity that would remain if the distribution of test scores M for black persons with high SES ($A = 1$, $B = 0$) with covariates $C = c$ was set equal to its distribution for white persons with high SES ($A = 0$, $B = 0$) with $C = c$ would be

$$\mu_{m10} - E[Y|\,A = 0, B = 0, c],$$

and the referent race disparity reduction would be

$$E[Y|\,A = 1, B = 0, c] - \mu_{m10},$$

where $\mu_{m10} = \Sigma_m E[Y|A=1,B=0,m,c]P(m|A=0,B=0,c)$.

The referent SES disparity that would remain if the distribution of test scores M for whites persons with low SES ($A = 0$, $B = 1$) with covariates $C = c$ was set equal to its distribution for white persons with high SES ($A = 0$, $B = 0$) with $C = c$ would be

$$\mu_{m01} - E[Y|A = 0, B = 0, c],$$

and the referent SES disparity reduction would be

$$E[Y|\,A = 1, B = 0, c] - \mu_{m01},$$

where $\mu_{m01} = \Sigma_m E[Y|A = 0, B = 1, m, c]P(m|A = 0, B = 0, c)$.

The excess intersectional disparity that would remain if, together, the distributions of test scores M for black persons with low SES ($A = 1$, $B = 1$) with covariates $C = c$, black persons with high SES ($A = 1$, $B = 0$) with covariates $C = c$, white persons with low SES ($A = 0$, $B = 1$) with covariates $C =$ c were each set equal to its distribution for white persons with high SES ($A = 0$, $B = 0$) with covariates $C = c$ would be

$$(\mu_{m11} - E[Y|A=0,B=0,c]) - \{(\mu_{m10} - E[Y|A=0,B=0,c]) + (\mu_{m01} - E[Y|A=0,B=0,c])\},$$

and excess intersectional disparity reduction would be

$$(E[Y|A=1,B=1,c] - \mu_{m11}) - \{(E[Y|A=1,B=0,c] - \mu_{m10}) + (E[Y|A=0,B=1,c] - \mu_{m01})\},$$

where

$$\mu_{m11} = \Sigma_m\, E[Y|A = 1, B = 1, m, c]P(m|A = 0, B = 0, c),$$

$$\mu_{m10} = \Sigma_m\, E[Y|A = 1, B = 0, m, c]P(m|A = 0, B = 0, c),$$

$$\mu_{m01} = \Sigma_m\, E[Y|A = 0, B = 1, m, c]P(m|A = 0, B = 0, c).$$

(In the presence of covariate feedback, i.e., a mediator-outcome confounder affected by race or childhood SES)

Consider again some outcome Y among persons defined by variables race/ethnicity A (0 = non-Hispanic white, 1 = non-

Hispanic black) and socioeconomic status B ($0 =$ high, $1 =$ low), among some level of covariates gender and age $C = c$, and M a measure of potentially manipulable characteristics in later life, e.g., measures of pre-market skills as reflected in test scores that may affect the outcome Y. Suppose now that there is a confounder L of test scores M that may be affected by race/ethnicity A ($0 =$ non-Hispanic white, $1 =$ non-Hispanic black) or socioeconomic status B ($0 =$ high, $1 =$ low) and that the effect of M on Y is unconfounded given A, B, L, and C, such that $E[Y(m)|a,b,m,l,c] = E[Y(m)|a,b,l,c]$ and also that consistency holds, such that $E[Y(m)|a,b,m,l,c] = E[Y|a,b,m,l,c]$.

The joint disparity that would remain if the distribution of test scores M for black persons with low SES ($A = 1$, $B = 1$) with covariates $C = c$ was set equal to its distribution for white persons with high SES ($A = 0$, $B = 0$) with $C = c$ would be

$$\mu_{m11} - E[Y|A = 0, B = 0, c],$$

and the joint disparity reduction would be

$$E[Y|A = 1, B = 1, c] - \mu_{m11},$$

where $\mu_{m11} = \Sigma_{m,l}E[Y|A = 1, B = 1, m, l, c]P(m|A = 0, B = 0, c)P(l|A = 1, B = 1, c)$.

The referent race disparity that would remain if the distribution of test scores M for black persons with high SES ($A = 1$, $B = 0$) with covariates $C = c$ was set equal to its distribution for white persons with high SES ($A = 0$, $B = 0$) with $C = c$ would be

$$\mu_{m10} - E[Y|A = 0, B = 0, c],$$

and the referent race disparity reduction would be

$$E[Y|A = 1, B = 0, c] - \mu_{m10}$$

where $\mu_{m10} = \Sigma_{m,l}E[Y|A = 1, B = 0, m, l, c]P(m|A = 0, B = 0, c)P(l|A = 1, B = 0, c)$.

The referent SES disparity that would remain if the distribution of test scores M for whites persons with low SES ($A = 0$, $B = 1$) with covariates $C = c$ was set equal to its distribution for white persons with high SES ($A = 0$, $B = 0$) with $C = c$ would be

$$\mu_{m01} - E[Y|A = 0, B = 0, c],$$

and the referent SES disparity reduction would be

$$E[Y|A = 1, B = 0, c] - \mu_{m01},$$

where $\mu_{m01} = \Sigma_{m,l}E[Y|A = 0, B = 1, m, l, c]P(m|A = 0, B = 0, c)P(l|A = 0, B = 1, c)$.

The excess intersectional disparity that would remain if, together, the distributions of test scores M for black persons with low SES ($A = 1, B = 1$) with covariates $C = c$, black persons with high SES ($A = 1$, $B = 0$) with covariates $C = c$, white persons with low SES ($A = 0$, $B = 1$) with covariates $C = c$ were each set equal to its distribution for white persons with high SES ($A = 0$, $B = 0$) with covariates $C = c$ would be

$$(\mu_{m11} - E[Y|A=0,B=0,c]) - \{(\mu_{m10} - E[Y|A=0,B=0,c]) + (\mu_{m01} - E[Y|A=0,B=0,c])\},$$

and excess intersectional disparity reduction would be

$$(E[Y|A=1,B=1,c] - \mu_{m11}) - \{(E[Y|A=1,B=0,c] - \mu_{m10}) + (E[Y|A=0,B=1,c] - \mu_{m01})\}$$

where

$$\mu_{m11} = \Sigma_{m,l}E[Y|A = 1, B = 1, m, l, c] \times P(m|A = 0, B = 0, c)P(l|A = 1, B = 1, c),$$

$$\mu_{m10} = \Sigma_{m,l}E[Y|A = 1, B = 0, m, l, c] \times P(m|A = 0, B = 0, c)P(l|A = 1, B = 0, c),$$

$$\mu_{m01} = \Sigma_{m,l}E[Y|A = 0, B = 1, m, l, c] \times P(m|A = 0, B = 0, c)P(l|A = 0, B = 1, c).$$

Results under linear models for causal mediation analysis in the absence of covariate feedback

Consider the following linear models:

$$E[Y|a,b,c] = \alpha_0 + \alpha_1 a + \alpha_2 b + \alpha_3 a * b + \alpha_5 c, \qquad (1)$$

$$E[Y|a,b,m,c] = \beta_0 + \beta_1 a + \beta_2 b + \beta_3 a * b + \beta_4 m + \beta_5 c. \qquad (2)$$

Consider an intervention to set each of the distributions of test scores M for black persons with low SES ($A = 1$, $B = 1$) with covariates $C = c$, black persons with high SES ($A = 1$, $B = 0$) with covariates $C = c$, and white persons with low SES ($A = 0$, $B = 1$) with covariates $C = c$ is equal to the distribution for white persons with high SES ($A = 0$, $B = 0$) with covariates $C = c$.

The residual joint disparity would be: $\beta_1 + \beta_2 + \beta_3$.

In addition, the amount the joint disparity is reduced would be: $(\alpha_1 + \alpha_2 + \alpha_3) - (\beta_1 + \beta_2 + \beta_3)$.

The residual referent race disparity would be: β_1.

In addition, the amount the referent race disparity is reduced would be: $\alpha_1 - \beta_1$.

The residual referent SES disparity would be: β_2.

In addition, the amount the referent SES disparity is reduced would be: $\alpha_2 - \beta_2$.

The residual excess intersectional disparity would be: β_3.

In addition, the amount the excess intersectional disparity is reduced would be: $\alpha_3 - \beta_3$.

Proof The non-parametric results under no covariate feedback follow directly from those of VanderWeele and Robinson [11] and also VanderWeele and Tchetgen

Tchetgen [16]. It follows from the results of VanderWeele and Robinson [11] that these formulae hold under an assumption that the effects of test scores M are unconfounded given race, SES, and covariates ethnicity, gender, and age (along with consistency and positivity assumptions for test scores). The results under covariate feedback follow directly from those of Jackson and VanderWeele [36] for Proposition five under an assumption that the effects of test scores M are unconfounded given race, SES, and covariates ethnicity, gender, and age, and a covariate L possibly affected by race and/or childhood SES (again, along with consistency and positivity assumptions for test scores).

The results under linear models follow, since we have

$$\begin{aligned}\mu_{m11} &= \Sigma_m E[Y|A=1,B=1,m,c]P(m|A=0,B=0,c)\\ &= \Sigma_m(\beta_0+\beta_1+\beta_2+\beta_3+\beta_4 m+\beta_5 c)\\ &\quad\times P(m|A=0,B=0,c)\\ &= \beta_0+\beta_1+\beta_2+\beta_3+\beta_4 E[M|A=0,B=0,c]\\ &\quad+\beta_5 c.\end{aligned}$$

Similarly

$$\begin{aligned}E[Y|A=0,B=0,c] &= \Sigma_m E[Y|A=0,B=0,m,c]\\ &\quad P(m|A=0,B=0,c)\\ &= \Sigma_m(\beta_0+\beta_4 m+\beta_5 c)\\ &\quad P(m|A=0,B=0,c)\\ &= \beta_0+\beta_4 E[M|A=0,B=0,c]\\ &\quad+\beta_5 c.\end{aligned}$$

Thus

$$\mu_{m11} - E[Y|A=0,B=0,c] = \beta_1+\beta_2+\beta_3.$$

Moreover

$$\begin{aligned}&E[Y|A=1,B=1,c]-E[Y|A=0,B=0,c]\\ &\quad=\alpha_1+\alpha_2+\alpha_3.\end{aligned}$$

And so

$$\begin{aligned}&E[Y|A=1,B=1,c]-\mu_{m11}\\ &\quad=\{E[Y|A=1,B=1,c]-E[Y|A=0,B=0,c]\}\\ &\qquad-\{\mu_{m11}-E[Y|A=0,B=0,c]\}\\ &\quad=(\alpha_1+\alpha_2+\alpha_3)-(\beta_1+\beta_2+\beta_3).\end{aligned}$$

We also have

$$\begin{aligned}\mu_{m10} &= \Sigma_m E[Y|A=1,B=0,m,c]P(m|A=0,B=0,c)\\ &= \Sigma_m(\beta_0+\beta_1+\beta_4 m+\beta_5 c)P(m|A=0,B=0,c)\\ &= \beta_0+\beta_1+\beta_4 E[M|A=0,B=0,c]+\beta_5 c.\end{aligned}$$

Similarly

$$\begin{aligned}E[Y|A=0,B=0,c] &= \Sigma_m E[Y|A=0,B=0,m,c]\\ &\quad\times P(m|A=0,B=0,c)\\ &= \Sigma_m(\beta_0+\beta_4 m+\beta_5 c)\\ &\quad\times P(m|A=0,B=0,c)\\ &= \beta_0+\beta_4 E[M|A=0,B=0,c]\\ &\quad+\beta_5 c.\end{aligned}$$

Thus

$$\mu_{m10} - E[Y|A=0,B=0,c] = \beta_1.$$

Moreover

$$\mathrm{E[Y|\ A=1,B=0,c]-E[Y|A=0,B=0,c]}=\alpha_1$$

And so

$$\begin{aligned}&E[Y|A=1,B=0,c]-\mu_{m10}\\ &\quad=\{E[Y|A=1,B=0,c]-E[Y|A=0,B=0,c]\}\\ &\qquad-\{\mu_{m10}-E[Y|A=0,B=0,c]\}\\ &\quad=\alpha_1-\beta_1.\end{aligned}$$

We also have

$$\begin{aligned}\mu_{m01} &= \Sigma_m E[Y|A=0,B=1,m,c]P(m|A=0,B=0,c)\\ &= \Sigma_m(\beta_0+\beta_2+\beta_4 m+\beta_5 c)P(m|A=0,B=0,c)\\ &= \beta_0+\beta_2+\beta_4 E[M|A=0,B=0,c]+\beta_5 c.\end{aligned}$$

Similarly

$$\begin{aligned}E[Y|A=0,B=0,c] &= \Sigma_m E[Y|A=0,B=0,m,c]\\ &\quad\times P(m|A=0,B=0,c)\\ &= \Sigma_m(\beta_0+\beta_4 m+\beta_5 c)\\ &\quad\times P(m|A=0,B=0,c)\\ &= \beta_0+\beta_4 E[M|A=0,B=0,c]\\ &\quad+\beta_5 c.\end{aligned}$$

Thus

$$\mu_{m01} - E[Y|A=0,B=0,c] = \beta_2.$$

Moreover

$$E[Y|A=0,B=1,c]-E[Y|A=0,B=0,c]=\alpha_2.$$

And so

$$\begin{aligned}&E[Y|A=0,B=1,c]-\mu_{m01}\\ &\quad=\{E[Y|A=0,B=1,c]-E[Y|A=0,B=0,c]\}\\ &\qquad-\{\mu_{m01}-E[Y|A=0,B=0,c]\}\\ &\quad=\alpha_2-\beta_2.\end{aligned}$$

From above, we have

$$\begin{aligned}&\{\mu_{m11} - E[Y|A = 0, B = 0, c]\}\\ &- \{\mu_{m10} - E[Y|A = 0, B = 0, c]\}\\ &- \{\mu_{m01} - E[Y|A = 0, B = 0, c]\}\\ &\qquad\qquad\qquad = \beta_3.\end{aligned}$$

In addition

$$\begin{aligned}&\{E[Y|A = 1, B = 1, c] - E[Y|A = 0, B = 0, c]\}\\ &-\{\mu_{m11} - E[Y|A = 0, B = 0, c]\}\\ &- (\{E[Y|A = 1, B = 0, c] - E[Y|A = 0, B = 0, c]\}\\ &-\{\mu_{m10} - E[Y|A = 0, B = 0, c]\})\\ &- (\{E[Y|A = 0, B = 1, c] - E[Y|A = 0, B = 0, c]\}\\ &-\{\mu_{m01} - E[Y|A = 0, B = 0, c]\})\\ &= \alpha_3 - \beta_3.\end{aligned}$$

This completes the proof.

Appendix 2: Analytic details for the variable AFQT scores used in the mediation analysis

Test score percentiles were obtained for the NLSY97 cohort from the Armed Forces Qualification Test which was administered as part of the Armed Services Vocational Aptitude Battery as reported in the 1999 survey year. These scores were similarly constructed to those used in the earlier 1979 NLSY79 cohort but are not considered official. Scores were standardized by 3-month age as described elsewhere [30]. All other variables followed the descriptions given in the main text of Jackson et al. [2].

References

1. Schwartz S (2017) Commentary: on the application of potential outcomes-based methods to questions in social psychiatry and psychiatric epidemiology. Soc Psychiatry Psychiatr Epidemiol 52(2):139–142
2. Jackson JW, Williams DR, VanderWeele TJ (2016) Disparities at the intersection of marginalized groups. Soc Psychiatry Psychiatr Epidemiol 51(10):1349–1359
3. Rothman KJ (1976) Causes. Am J Epidemiol 141(2):90–95
4. VanderWeele TJ, Robins JM (2007) The identification of synergism in the sufficient-component-cause framework. Epidemiology 18(3):329–339
5. Holland PW (1986) Statistics and causal inference. J Am Stat Assoc 81(396):945–960
6. Greiner DJ, Rubin DB (2011) Causal effects of perceived immutable characteristics. Rev Econ Stat 93(3):775–785
7. Glymour C (1986) Statistics and metaphysics. J Am Stat Assoc 81(396):964–966
8. Kaufman JS, Cooper RS (1999) Seeking causal explanations in social epidemiology. Am J Epidemiol 150(2):113–120
9. Krieger N, Smith GD (2000) Re: “Seeking causal explanations in social epidemiology”. Am J Epidemiol 151(8):831–833
10. VanderWeele TJ, Hernán MA (2012) Causal effects and natural laws: towards a conceptualization of causal counterfactuals for nonmanipulable exposures, with application to the effects of race and sex. In: Berzuini C, David P, Bernardinelli L (eds) Causality: statistical perspectives and applications. Wiley, New Jersey, pp 101–113
11. VanderWeele TJ, Robinson WR (2014) On the causal interpretation of race in regressions adjusting for confounding and mediating variables. Epidemiology 25(4):473–484
12. Reskin B (2012) The race discrimination system. Annu Rev Sociol 38(1):17–35
13. Williams DR, Mohammed SA (2013) Racism and health I. Am Behav Sci 57(8):1152–1173
14. Wilson WJ (2010) Structural and cultural forces that contribute to racial inequality. In: More than just race: being black and poor in the inner city. Norton & Company, New York
15. Rothman KJ (1986) Modern epidemiology, 1st edn. Little, Brown, Boston
16. VanderWeele TJ, Tchetgen Tchetgen EJ (2014) Attributing effects to interactions. Epidemiology 25(5):711–722
17. Hill Collins P (2015) Intersectionality’s definitional dilemmas. Annu Rev Sociol 41(1):1–20
18. Hancock A-M (2007) Intersectionality as a normative and empirical paradigm. Politics Gend 3(2):248–254
19. Crenshaw K (1991) Mapping the margins: intersectionality, identity politics, and violence against women of color. Stanf Law Rev 43(6):1241
20. Hancock A-M (2013) Empirical intersectionality: a tale of two approaches. UC Irvine Law Rev 3(2):259–296
21. Braveman P (2006) Health disparities and health equity: concepts and measurement. Annu Rev Public Health 27:167–194
22. Cook BL, McGuire TG, Zaslavsky AM (2012) Measuring racial/ethnic disparities in health care: methods and practical issues. Health Serv Res 47(3pt2):1232–1254
23. Williams DR, Mohammed SA (2013) Racism and health II. Am Behav Sci 57(8):1200–1226
24. Oaxaca R (1973) Male–female wage differentials in urban labor markets. Int Econ Rev (Philadelphia) 14(3):693
25. Blinder AS (1973) Wage discrimination: reduced form and structural estimates. J Hum Resour 8(4):436
26. Greenland S (2005) Epidemiologic measures and policy formulation: lessons from potential outcomes. Emerg Themes Epidemiol 2:5
27. Cooper LA, Hill MN, Powe NR (2002) Designing and evaluating interventions to eliminate racial and ethnic disparities in health care. J Gen Intern Med 17(6):477–486
28. Mackenbach JP, Gunning-Schepers LJ (1997) How should interventions to reduce inequalities in health be evaluated? J Epidemiol Community Health 51(4):359–364.
29. Brayboy Jackson P, Williams DR (2006) The intersection of race, gender and SES: health paradoxes. In: Gender, race, class & health: intersectional approaches. Jossey-Bass, San Francisco, pp 131–162
30. Fryer RG (2011) Racial inequality in the 21st century: the declining significance of discrimination. In: Handbook of labor economics, vol 4, pp 855–971
31. Neal DA, Johnson WR (1996) The Role of premarket factors in black–white wage differences. J Political Econ 104(5):869–895
32. Bureau of Labor Statistics, U.S. Department of Labor, and National Institute for Child Health and Human Development. Children of the NLSY79, 1979–2014. Produced and distributed by the Center for Human Resource Research, The Ohio State University, Columbus

33. Hernán MA (2016) Does water kill? A call for less casual causal inferences. Ann Epidemiol 26(10):674–680
34. Imai K, Keele L, Tingley D (2010) A general approach to causal mediation analysis. Psychol Methods 15(4):309–334
35. Valeri L, VanderWeele TJ (2013) Mediation analysis allowing for exposure–mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. Psychol Methods 18(2):137–150
36. Jackson JW, VanderWeele TJ (2017) Decomposition analysis to identify intervention targets for reducing disparities. http://arxiv.org/abs/1703.05899. Accessed 13 Apr 2017
37. Hernán MA, Hernández-Díaz S, Robins JM (2004) A structural approach to selection bias. Epidemiology 15(5):615–625
38. Wallace ME, Mendola P, Liu D, Grantz KL (2015) Joint effects of structural racism and income inequality on small-for-gestational-age birth. Am J Public Health 105(8):1681–1688