
Article:
Chamberlain, J., Holroyd, J., Jenkins, B. et al. (1 more author) (2023) Implicit bias, intersectionality, compositionality. Philosophical Psychology. ISSN 0951-5089

https://doi.org/10.1080/09515089.2023.2213245

Implicit Bias, Intersectionality, Compositionality

1. Introduction

Research in social psychology on implicit biases, over the past two decades, has used a range of experimental tools to try to access aspects of individual cognition that people may be unable, or unwilling, to report on. These measures aim to reveal whether people who profess to hold anti-racist, anti-sexist, and otherwise egalitarian beliefs and attitudes, might nonetheless harbour biases "under the radar".[1] Such research programmes have been enormously influential - participated in by millions of people,[2] leading to high profile narratives about the ways prejudice has “gone underground”, and informing large investments in workplace training that raises awareness about (and which sometimes dubiously promises to uproot, or mitigate the effects of) implicit biases.[3]

A decade ago, concerns were raised about the ability of measures used in these research programmes to engage with the complexities of biases. In particular, Goff and Kahn (2013) challenged the idea that these research programmes revealed anything about biases that target those facing intersectional oppressions. One of the concerns is this: a prominent measure in such research programmes is the Implicit Association Test. By design, this measure evaluates biases concerning one dimension of social identity - comparing, for example, the extent to which a particular attribute is associated with black rather than white people; or with men rather than women. One way of understanding these studies is as trying to access features associated with the concept WOMAN or BLACK PERSON. But what concepts do participants actually have in mind when engaging in these studies? Goff and Kahn present empirical evidence suggesting that participants are likely to have in mind *white* women, and black *men* in particular (2013, 375-376). Rather than reveal features of the participants’ concept WOMAN, the studies reveal features of the participants’ concept WHITE WOMAN, for example. That would mean that these studies cannot be assumed to tell us anything about the kinds of associations participants have in relation to black women (or the concept BLACK WOMEN), for example. More generally, biases that target people at the intersection of different forms of oppression will most likely not be detected by these kinds of experimental measures.

Since this critique, a range of high profile papers have reported on studies that aim to look at how biases might target those who face multiple intersecting oppressions, and hence to examine how biases might intersect.

Some of the most interesting findings in this literature concern the ways that perceptions of one social identity can shape attributions of other features. For example, a series of studies find that the social class status attributed to an individual affects the way they are racialised. People categorised as lower class or poor are more likely to be categorised as black (Freeman et al. 2011; Penner and Saperstein 2013; Lei and Bodenhausen 2017). In a similar interaction, attributions of race appear to affect attributions of maturity. Black girls in particular are likely to be "adultified": perceived as more mature, more knowledgeable about topics such as sexual relations, and less in need of nurture than

[1] The claim that bias has “gone underground” and is now operating “under the radar” has now been prevalent for decades (cf. Dovidio & Gaertner 1986; Monteith et al. 2001; Bergh & Hoobler, 2018).
[2] Project implicit are continually presenting data from online participation in implicit measures, totalling over two million performances on the test (Xu et al. 2014)
[3] See Atewologun et al. (2018) on the efficacy of implicit bias training, and Madva (2017) for the merits of such institutionally endorsed trainings.

white girls or black boys (Epstein et al. 2017, building on research on black boys by Goff et al. 2014). This indicates that perception of social categories is highly interdependent (though see Petsko et al 2022). This would suggest that biases targeting those facing multiple oppressions would be similarly complex. And indeed, reports of real world experiences at the intersection of multiple axes of oppression make visible this complexity. For example, while researching the real world effects of the adultification of black girls, Blake and Epstein heard about the following, humiliating case from a research participant:

> In ... sixth grade, ... the school nurse, like, ask[ed] my aunt if I was sexually active .... And I was, like, at the time, like, what? Like, what? Nobody has sex. Like, I didn't know anyone that had sex. And it was so crazy to me. And then just thinking, like, she would never think to ask my [white] friend that. (2019, 6)[4]

The participant describes a racial bias in the nurse's attitudes; the bias is also gendered and related to age. It appears that stereotypes specific to black girls are operative here: black women and girls have historically been subjected to the stereotype of the hypersexualised "jezebel".[5] The specific intersections of race, gender and age culminate in forms of discriminatory treatment that cannot be separated into the effects of gender, race, or age, in isolation. Thus, understanding how biases and stereotypes target those at the intersection of multiple oppressions is a valuable task.

This task is undoubtedly a difficult one. For one thing, it is not obvious how we should understand the relationship between *experiences of* intersecting oppressions due to the social categories of which one is a member, and *cognition about* intersecting social categories. These might come apart (as folk psychology about psychological kinds might not track what psychological kinds there are). In what follows, our focus will be on cognition about intersecting social categories. We will make no claims about how we should model intersectionality or intersectional oppression as a metaphysical phenomenon.[6]

Rather, the aim of this paper is to evaluate the success and limitations of some recent attempts within social psychology to grapple with the ways biases operate towards those facing intersecting forms of oppressions. Our contention is that problematic conceptualisations of intersectionality inform, in some instances, the predictions generated and the experimental designs intended to test them. This, alongside unarticulated assumptions about how concepts combine, has thwarted the success of these efforts to address intersectional biases.

In section 2, we show how inadequate conceptualisations of intersectionality have undermined research on intersecting biases and negatively impacted upon experimental design. In section 3, we identify assumptions made about how concepts and their associated stereotypes combine, and show how these assumptions have thwarted identification of biases faced by those experiencing multiple forms of oppression. In section 4, we collate some recommendations, made throughout the paper, for how these problems can be avoided in future research on implicit biases.

[4] For those not familiar with US schooling grades: sixth graders are typically 11-12 years old.
[5] See Collins (1990) for a discussion of this jezebel stereotype and controlling images more broadly.
[6] For some helpful papers on this topic, see Bright et al (2016), and Bernstein (2020).

2. Intersectionality

The idea of intersectional oppression has a long history, but was coined as such and incorporated (not without ongoing resistance) into academic thinking following Kimberlé Crenshaw's influential 1989 paper.[7] There Crenshaw aimed to critique both the frameworks used in anti-discrimination law, and aspects of feminist and civil rights thinking, that adopted a "single axis" approach: focusing on differential treatment based on membership of one social category (one's gender, or one's race), and assuming that all members of that social category had similar experiences of discriminatory treatment. Thus, if black women's experiences of discrimination did not align with those of white women, or those of black men, they were not recognised as experiencing discrimination on the basis of either gender or race.

Such assumptions overlook the fact that "Black women can experience discrimination in ways that are both similar to and different from those experienced by white women and Black men" (1989, 149). A framework is needed, Crenshaw argues, that acknowledges the experiences of intersectional oppressions - the qualitatively different experiences that one may have as a result of experiencing oppression based on one's gender *and* one's race *and* indeed other social group memberships (class, age, sexuality, disability and more). This would not only make available remedies, through anti-discrimination law, to those experiencing intersectional oppressions. It would also, more broadly, enrich feminist and civil rights theory and practice by showing the false assumptions of neutrality present in the supposition that our societies and institutions are just "but for" those instances of discrimination against individuals based on one aspect of their social identity. Rather than supposing that our society is just "but for" those cases when race or gender is unduly considered, it would reveal more fully the social transformations needed if social institutions are not to be premised on the experiences of privileged white men. It would enable social movements to more fully address those experiencing marginalisation and oppression and more radical transformations to be envisaged: those that focus on structural, as well as individual, forms of oppression (1989, 166-167).

The key point we want to emphasise in Crenshaw's call to heed intersectionality is this: that experiencing multiple forms of oppression can lead to *qualitatively different experiences of oppression* from other persons who share any one social group. We emphasise this because, as we will see in the following sub-section, it has not been adequately heeded by those taking up the idea of intersectionality in experimental social psychology.[8] We turn to the ways the concept has been used in recent social psychological studies now.

2a. Conceptualisations of intersectionality in the social psychology literature:

The aim of researchers we discuss below is to make progress in understanding cognition about intersecting forms of oppression - for example, about the kinds of biases that target people who experience multiple forms of oppression. Rather than asking what kinds of stereotypes or associations participants have regarding, e.g. black or white people, or men

[7] See Crenshaw's (1991) paper for further development of the observations and concepts in Crenshaw (1989). For earlier work, see Truth (2020/1851), Cooper (1988/1892), Beal (2009/1969), Collins' (1990). For more recent work, see Henning (2020), and Dotson (2016).

[8] See Henning (2020) for worries that, under certain conditions, such misuses may constitute forms of "methodological microaggression".

or women, these researchers are aiming to understand how biases might operate at the intersection of multiple social categories. For example, they might ask how biases target people who are encountered as, for example, black women who are working class; or white men who are middle class; or black girls; or white women who are adults.

A terminological note: Crenshaw and those concerned with intersectionality focus on the *experiences* of those living at the intersection of multiple forms of oppression. Researchers in social psychology, such as those we discuss below, are concerned with the *biases* that those who experience intersecting oppressions face. For clarity (and brevity), when discussing the biases that those facing intersecting oppressions experience, we will refer to this phenomenon as that of *intersecting biases*. The social psychologist's question is how these biases behave.

Crucially, understandings of intersecting biases should be able to inform strategies for addressing those biases and other problematic or discriminatory cognitions - and indeed, do a better job of doing so than those based on "single axis" measures. However, we argue that both understanding about biases, and recommendations for how to combat them, are hindered due to inadequate conceptualisations of intersectionality. We present two indicative misconceptualisations: first, intersectionality is elided with non-binary social identities. Second, intersectionality is conceived of in uni-dimensional terms.

i. Intersectional and non-binary social categories

One problematic conceptualisation features in a summary piece by Kang and Bodenhausen (2015) outlining the challenges and opportunities in social psychological research that are presented by the reality of the intersections of social category attributions. Whilst offering a helpful survey of a range of experimental and theoretical work to date, Kang and Bodenhausen slide between talk of "intersections" in social identities and the stereotypes that target them, and of persons whose "identities lie within the intersections of […] conventionally binary distinctions" (2015, 548). By the first locution, Kang and Bodenhausen clearly mean the sorts of intersections that Crenshaw had in mind - the distinct stereotypes that might target someone on the basis of race *and* gender *and* class, for example. Such intersections, they write, "make it quite tenuous to think in unqualified, general terms about the psychological impact of any particular category (e.g., race or gender)" (2015, 548). Meanwhile, the second kind of "intersection" they have in mind concerns non-binary social identities - being attributed membership of racial groups other than black or white - such as multiracial - or being categorised as having a non-binary gender (rather than as a binary gender: man or a woman).[9] Whilst Kang and Bodenhausen aim to survey a broad range of literature, we believe that to avoid confusion and problematic prescriptions, it is of the utmost importance to clearly distinguish cases in which intersectional oppression is at stake, and cases in which membership of non-binary social categories is at issue (and of course, members of non-binary social categories might themselves face intersecting forms of oppression).

To see this, consider their discussion, later in the paper, of strategies for combating biases. Their discussion builds on studies that focus on the intersection of race, gender and sexuality. In particular, they focus on the finding that black gay men and black straight men

[9] Kang and Bodenhausen use "transgender" as an example of this kind of identity - but note that whilst some identities falling under the umbrella term "trans" - some gender queer or some non-binary persons - will fit this description, some trans men or women fall within binary gender categories (not the category assigned at birth).

are stereotyped differently, and that stereotypes about gay men may modify the stereotypes associated with black (straight) men, concerning hostility, danger and threat (Remedios et al. 2011, discussed in Kang and Bodenhausen 2015 at 555-556). Kang and Bodenhausen tease out the following prescription: "Thus, perceivers can avoid or reduce unwanted forms of bias in their social perceptions by attending to multiple identity dimensions that bring clashing stereotypes into focus" and by "selectively highlighting particular social categories" (2015, 556). Having discussed similar studies (concerned with focusing on flexible assignment of ingroup status) Kang and Bodenhausen make the following more general prescription: that attending to multiple identities presents the possibility

> "that a negative stereotype associated with one broad social identity will be undermined by contradictory stereotypes about another identity. Thus, by recognizing the multifaceted social identities of others, perceivers can take an important step toward less biased decisions and more positive social interactions." (2015, 556)

The suggestion, in short, is to pitch stereotypes against one another, and aim to focus on the one that might cancel out other negative stereotypes. First, one concern that we might have with this strategy is that rather than undermining stereotypes, it simply trades on them in the hope that some positive stereotype will trump any negative stereotypes - scant consolation to those whose multiple social group memberships are all negatively stereotyped in some way or another. Second, there is also something insulting about suggesting that a perceiver should strategically "cancel out" or "overcome" an aspect of social identity ("focus on their sexuality in order to overcome biases associated with their race") - as if both might not be important parts of identity! This normalises the erasure of certain social identities. A third concern is that such prescriptions are obviously completely inapt in relation to non-binary social categories, and if applied there could perpetuate harms. (To clarify: Kang and Bodenhausen do not explicitly suggest that such strategies should be applied in cases of non-binary social identities. But nor do they take care to specify that such strategies are inapt in those cases.) Consider what such a bias prevention strategy would prescribe: "when encountering multiracial individuals, try to focus on positive stereotypes associated with whiteness in order to overcome any negative stereotypes associated with blackness". Or: "when encountering non-binary individuals, try to overcome any negative stereotypes associated with non-binary individuals by focusing on the positive stereotypes associated with men (or indeed women, depending on the context)". Such strategies are offensive, and harmful due to miscategorising those they target: ignoring an individual's multiracial identity; or misgendering a non-binary person.[10] This conceptualisation therefore fails to adequately model intersectional social categories.[11]

ii. Models of intersectionality

Now consider the second kind of misconceptualisation. In a large multi-author experimental paper reporting on a number of studies investigating how implicit biases might interact, Connor et al. (2022) survey different possible ways of conceptualising how biases about individuals who experience multiple forms of oppression, due to membership in multiple stigmatised social categories, might interact. Their studies aim to examine the interaction of

[10] See Kapusta (2016) on harms of misgendering.

[11] Thanks to an anonymous reviewer for helping us frame the concerns in this section.

biases associated with gender, race, social class, and age. In what follows, we will focus heavily on their work since it is a prominent, much downloaded, large (and presumably costly!) study into intersecting biases. The authors include leading figures in research on implicit biases.

Connor et al.'s experimental work is framed in terms of three models for how biases targeting those experiencing multiple forms of oppression (that is, those perceived to belong to multiple marginalised or stigmatised social groups) interact.[12] One model predicts that the effects of these biases will compound. In doing so, they might interact additively. A second compounding model considered here, and drawn from Crenshaw's work, is that the effects of these biases might produce "multiple jeopardy", characterised as "a negative bias that exceeds the sum of the negative biases associated with each category" (2022, 2). The third model considered is the "category dominance model", which predicts that one perceived social categorisation will be the most salient (which is not determined, and may be contextually dependent), and that this social categorisation will drive any biased effects.

The first thing to note is that none of these three models (compounding: additive or multiplicative, and category dominance) for how those experiencing multiple oppressions might be targeted by biases (including the second model, which putatively builds on Crenshaw's work) respect Crenshaw's key insight about intersectional oppressions: namely, that it can produce *qualitatively distinct experiences of oppression* (see also McCall 2005). The experiences may not be the mere sum, nor the multiplicative output ("a negative bias that exceeds the sum of negative biases"), of the experiences of other group members. Rather, they might be distinctive, as a result of how gendered oppression and racial oppression (and other experiences of oppression based on social category membership) interact. Put differently, some of the effects of intersecting biases will elude the uni-dimensional perspective present in each of the models considered by Connor et al.

To see what is missing from these three models, recall our earlier example of the intersecting biases exhibited by a school nurse. The experience of being subjected to such biases is unlikely to be the result of an additive or multiplicative interaction between, for example, the biases which target either white girls in sixth grade or black boys in sixth grade. Nor is one perceived social category (race, gender or age) driving their experiences.

The second thing to note is that these models of intersecting bias appear to detrimentally inform the experimental designs the authors use, thus in our view rendering the data gathered at best unhelpful, and at worst highly misleading. In the following sub-section, we detail what we take to be the key missteps that follow from these misconceptualisations of intersectionality. These pertain in particular to the choice of stimuli used, and experimental design, in attempting to evaluate the ways in which attributes associated with different social categories might interact.

2.b. Impact of these conceptualisations in empirical studies

Connor et al. (2002) present 5 large studies (totalling 5,204 participants), which aim to shed light on which of their hypotheses, about how biases targeting multiple intersectional

[12] Note that experiencing multiple forms of oppression, and being perceived as a member of multiple stigmatised or marginalised groups, come apart, e.g. if one experiences oppression based on non-perceptible features, as may be the case with so-called "non-visible" disabilities (for discussion see Cureton 2018, McGuire & Carel, 2019, Stramondo forthcoming). Clearly, the studies focus primarily on those social categories that are perceptibly discernable. Somewhat surprisingly, studies suggest sexuality to be amongst these categories (see e.g. Remedios et al. 2011).

oppressions interact, is best supported. In so doing, they also seek to provide new experimental paradigms for how to investigate intersecting implicit biases. Recall, the competing hypotheses in play are the compounding hypothesis, in either additive or multiplicative version; and the category dominance hypothesis. Their focus was on biases that might target individuals at the intersection of different gender, racial, class and age categories.[13] Consider the predictions generated by each of these three models:

> P1: the compounding models would predict that those at the intersection of multiple axes of oppression would be subject to negative effects to a greater extent than those who are targeted by just one form of bias.

That is, the negative output of intersecting biases are greater than the output of biases targeting one social category, on the basis of which one is oppressed. Precisely how these biases compound differs according to the details of the model, additive or multiplicative:

> P1-a: if intersecting biases are additive, the negative effects of biases will be the sum of the negative biases associated with each category.
>
> P1-m: if intersecting biases are multiplicative, the negative effects of biases will be greater than the sum of the negative biases associated with each category.

For example, given biases that target women, black people, lower class people, and the elderly, compounding models would predict that lower class elderly Black women be subject to negative biases that either sum the biases associated with lower class status, the elderly, women and black people (additive model), or exceed the sum of the negative biases associated with each category. Thus, on either model, lower class elderly black women would be associated with negative biases to a greater extent than those who are targeted by some but not all of these biases (e.g. young white middle class women, elderly Black working class men). We might already have some reservations here: what does it even mean to "sum" or "multiply" qualitatively different stereotypes? We return to these concerns below.

In contrast, the category dominance model would generate a different prediction:

> P2: where multiple biases may be in play, one social category will be dominant and will drive the effect. Thus where (e.g.) gender biases are dominant, we would expect to see women targeted by negative biases to the same degree (irrespective of race, class, age); where racial biases are dominant, we would expect to see all black people targeted by biases to the same degree (irrespective of gender, class, age).

This prediction is motivated by observations about the limited cognitive resources of individuals, and the subsequent need for a parsimonious handling of these resources. The model makes no prediction about *which* social category will be the most salient, and therefore drive the bias effects.

The category dominance and compounding models are in disagreement as to which group will be subject to the negative effects of bias to the greatest extent. The additive and

[13] The perceived income of the pictured people was treated as a proxy for the perceived social status of those persons.

multiplicative models differ on the extent to which we should expect biases to manifest. Connor et al. (2022) intend their studies to advance our understanding of how biases intersect by settling these disagreements.[14] In the following, we show how these understandings of intersectionality, and the predictions they inform, shape the experimental tools used, and data gathered.

i. Choices of experimental tools

The first three of Connor et al.'s studies depend upon deployment of (or innovations in) the use of Implicit Association Tests (IATs). Such measures are categorisation tasks. The speed with which participants are able to complete the categorisations (in different blocks of the experimental task) are taken to indicate the strength of associations between the target and the stimuli; or the readiness with which individuals can access that stimuli, given exposure to the target.

One such measure they deploy is a Single-target IAT.[15] We focus on their deployment of this tool to illustrate where the inadequate conceptualisations of intersectionality have an impact. In study 1a (307 participants), Connor et al. introduced stimuli in the form of words (positively and negatively valenced) and target persons that varied according to race and class (the gender and age of the targets were held fixed). Participants are instructed to categorise words (positively and negatively valenced) and persons (varying according to race and class) as good (congruent trials) or bad (incongruent trials). The positively valenced words include: *Beautiful, Glorious, Joyful, Lovely, Marvellous, Pleasure, Superb, Wonderful* (2022, 5). And the following words were used as negative cues: *Agony, Awful, Horrible, Humiliate, Nasty, Painful, Terrible, Tragic* (ibid). We emphasise these particular positive and negative terms, as we return critically to these stimuli in ii below.[16] The measure is of whether participants will be able to associate the targets more quickly as good, rather than bad, depending on features of the person (their race or class). This is with a view to initially measuring how race and class biases might interact. Study 1a was a "within subjects" study (meaning participants were each exposed to targets from the varying social groups (the intersections of black/white; high/middle/lower class) and potentially differential responses recorded). Study 1b (533 participants) was instead a "between subjects" study (meaning participants were randomly selected to respond to targets from one of four groups (the intersections of black/white; high/low class), and potentially differing responses to different conditions across participants measured). Study 1b also introduced another ST-IAT with different stimuli - instead of positive/negative terms, intended to identify evaluative biases associated with target groups, the stimuli included

[14] Ultimately, Connor et al. (p.23) argue that the results of their studies support a hybrid account of bias interactions, reporting that both category type salience, and compounding effects, can be detected in the results. As we will see, our view is that the data is insufficiently robust to support any such conclusions.

[15] For a useful summary and evaluation of ST-IATs see Bluemke & Friese (2008). Note that the ST-IAT requires participants to respond to stimuli (representations of individuals belonging to particular social groups), rather than specified social categories. As such other methods that rely on similar techniques (such as evaluative priming tasks) may also face these concerns.

[16] Note these are well established and much used stimuli. See e.g. race evaluative IATs used at Project Implicit https://implicit.harvard.edu/implicit/iatdetails.html.

wealth and poverty associated terms, to identify whether racial groups with different class statuses are more strongly associated with notions to do with wealth or poverty.[17]

Consider what the competing models would predict (we restrict our attention to the multiplicative and category dominance models for brevity's sake):

> P1m (race/class): if intersecting biases are multiplicative, the negative effects of biases will be greater for lower class black men than the sum of the negative biases associated with each category (and thus greater than the biases associated with black middle class, black high class; or white lower class targets).
>
> P2 (race/class): where multiple biases may be in play (here: race and class) one social category will be dominant and will drive the effect. If class bias drives the effect, we would expect to see the same magnitude of bias expressed towards black and white lower class targets. If race drives the effect, we would expect to see the same magnitude of bias against black targets, irrespective of class status.

Connor et al. report findings which, they claim, support the category dominance model: in both study 1a and study 1b the social class of the targets drove the effect (2022, 6). Namely, participants more strongly associated lower class targets - whatever their race - with negative constructs. And (perhaps unsurprisingly!) participants more strongly associated lower class individuals with poverty-related constructs. In contrast to the effects of social class, no significant main effects of race were found in either 1a or 1b: race was not more or less strongly associated with positive or negative terms; nor more or less strongly associated with wealth terms; and race did not appear to moderate the effects of class biases.

Our view is that we cannot suppose these findings tell us much at all about the relationship between race and class, and the biases associated with them, precisely because the experimental design fails to adequately grapple with the concept of intersectionality. Firstly, note that the studies don't countenance at least two possible ways in which the social categories might interact. Various studies indicate that a person's perceived class status affected the way in which they were racialised: being viewed as lower class, or poor, correlated with being perceived as black (Freeman et al 2011; Penner and Saperstein 2013; Lei and Bodenhausen 2017). Moreover, there are strong racial associations with class: black people are more strongly associated with "poverty" (a key component of class status) (Cox and Devine 2015; Brown-Iannuzzi et al 2019). Either kind of influence would undermine the category dominance hypothesis. This is highly relevant to how we interpret the results, but the experimental tools used don't appear to be able to speak to it.[18]

---

[17] The constructs reported on for this ST-IAT include: wealth terms: *rich, wealthy, affluent, prosperous, well off, loaded, fortune, lucrative*; poverty terms: *poor, poverty, destitute, needy, impoverished, broke, bankrupt, penniless* (2022, 5). A concern one might have here is the relationship of identity, rather than association, of some of the stimuli with the category labels (poverty, wealth). This is troubling, but distinct from our main thread of concern so we set it aside. See Haider et al. 2011 for IAT stimuli including class related constructs.

[18] One might think that pre-experimental stimuli selection could control for the possible interactions between social categorisations; in personal correspondence with Connor, he indicates that attempts to match black and white targets on the basis of perceived class was done without controls for other variables in perceived traits (warmth, competence, attractiveness), in a way that may artificially suppress racial biases. Indeed, Connor suggests: "no matter what set of perceived traits you choose to match [black and white targets] on, creating mismatches on additional non-matched perceived traits

Second, consider that the predictions, and hence the experimental tools designed to evaluate them, presuppose that any interactions between biases (here, race and class biases) will show up in the *magnitude* of biases exhibited on the ST-IATs. As we have previously put the problem: these models all indicate an assumption that the interactions between biases will be uni-dimensional. This precludes the possibility that a *qualitatively different kind* of bias might be expressed. Yet this is what we might expect were we to engage with the literature on the different kinds of stereotypes associated with race and class. True to Crenshaw's insight about the distinct experiences at the intersections of oppressions, existing literature indicates that the sorts of stereotypes and biases that target black lower class men and white lower class men differ in quality. Consider stereotypes to do with gang affiliations and criminality that face black lower class men (Steffensmeier 1998); and the "white trash" stereotype associated with white lower class men (Hartigan 2013). This is obscured by implicit measures that focus solely on the magnitude of positive or negative biases. Indeed, it is not at all clear that the different ways in which these biases are negative would be expected to show up, in linear fashion, on a positive/negative evaluative IAT. Given that the structure of IATs delivers these uni-dimensional findings about the *magnitude*, but not *kind* of bias, we might think that such experimental tools are simply not adequate for investigating such intersecting biases. Note that this is not simply a function of the IAT: it is applicable to any measure that is designed to exclusively test the magnitude of positivity and negativity in participants' responses.

Of course, there are sometimes good methodological reasons to focus on such measures. For example, consider the studies of Perszyck et al. (2018). Their studies use the Affect Misattribution Procedure (also used by Connor et al. in their study 4) to investigate negative biases that might be expressed towards children who differ with respect to race (black/white) and gender (girl/boy). Such tools expose participants to an unfamiliar visual symbol (a Chinese character), having been primed by a racialised, gendered face, and measure the extent to which the symbol is evaluated positively or negatively ("nice looking" or "not nice looking"). In this case, Perzyck et al.'s participants were children, and so they had good reason to choose methods which require no reflective thought, and which don't rely on fast reaction times. Nevertheless, note that such methods deliver uni-dimensional negative or positive evaluations, and are therefore insensitive to the kind of rich stereotypical contents that studies examining qualitative stereotypes associated with black and white girls and boys can reveal.

The risk is that what we do find out will be driven by the experimental tools available, and the hypotheses these inform (to do with magnitude or dominance of biases); and not by the actual qualitative experiences of intersecting oppressions. Moreover, simply recognising that several social categories, or several intersecting oppressions, co-exist is insufficient. When these are simply added into the experimental task, *without* attending to the possibility of qualitatively different experiences of oppression, further worries emerge.

ii. Biased stimuli

In study 2 (371 participants), Connor et al. (2022) aimed to investigate a greater range of intersections: the biases that might be in play given targets that varied along the following dimensions: "three different races (e.g., Asian, Black, and White), two genders (female vs. male), two levels of social class (high vs. low), and two levels of age (old vs. young)" (2022,

is likely and perhaps inevitable as long as race is perceived as being causally linked to the traits (as it is in the case of income)." (personal correspondence, 20/02/2023)

6). A ST-IAT was again used, with words (positively and negatively valenced) and target persons that varied according to the four social category dimensions mentioned above: crucially *now including gender*.[19] As before, participants were instructed to categorise words (positively and negatively valenced) and persons (varying according to race, class, age *and gender*) as good (congruent trials) or bad (incongruent trials).

Here's the important thing to note: the stimuli used for the ST-IAT were those used in what is standardly referred to as an "evaluative IAT" - an IAT which assesses comparative positive and negative strengths of association. These tests and their stimuli have been much used in evaluating implicit racial biases, where the aim is to consider the extent to which racial categories - typically cued either by words (black/white), or by male faces (pictures of faces that are black or white) - are differentially associated with positive or negative valence.[20] Connor et al. deploy a ST-IAT, built around evaluative IAT stimuli, to evaluate the valences attached to individuals of multiple social categories (race, gender, class, age). To see what is problematic here, consider one of the recent recommendations for best practice in using IATs:

> "A4. For IATs designed to measure stereotypes, avoid confounding the stereotype's contrasted attributes with valence" (Greenwald et al. 2022, 1166).

The thought here is that some stereotyped content - such as that used in gender-potency IATs - includes stereotypical content that is also valenced. For example, on a gender/potency stereotype measure, "strong" is positively valenced; "weak" is negatively valenced. If that feature pervades one's stimuli, then it will be difficult to ascertain whether any effect is indeed driven by the stereotyped content of the stimuli, or rather by the valences of the stimuli, and differential associations between the target and those positive or negative valences. The recommendation, then, is to try to "valence match" the stimuli, to avoid this potential confound (see Rudman et al. 2001, Greenwald et al. 2022 for discussion). Note that this recommendation should go both ways, although Greenwald et al. do not make this explicit. That is:

> A4*. For IATs designed to measure evaluative associations, avoid confounding the valences of attributes with stereotyped content.

Consider again the evaluative notions incorporated into the evaluative ST-IAT used by Connor et al:

> Positive: *Beautiful, Glorious, Joyful, Lovely, Marvellous, Pleasure, Superb, Wonderful*;

---

[19] The method was somewhat different from study 1, to accommodate the fact that given the larger number of social categories in play, potentially 24 conditions would be needed (Asian woman, of low class, elderly; Black woman, of low class, elderly; White woman, of low class, elderly; Asian man, of low class, elderly; Black man, of low class, elderly; White man, of low class, elderly…. Etc to a total of 24 combinations of the social categories). Rather than assign participants to one of 24 conditions, they looked at the relationship between a participant's difference rating of the stimuli (along the dimensions of race, age, class and gender, in relation to the ST-IAT score) (see pp.9-10).

[20] Crucially, it is not the standard measure used in various studies of gender bias. Instead, IAT stimuli have focused on specific gender stereotypes, including: Gender/career stereotypes (project implicit); gender/potency stereotypes (Rudman et al. 2001); gender/STEM vs. arts stereotypes (Charlesworth et al. 2022); Gender and leadership stereotypes (Dasgupta and Asgari 2004).

Negative: *Agony, Awful, Horrible, Humiliate, Nasty, Painful, Terrible, Tragic* (2022, 5).

Note that some of the positive evaluative terms contain some (racialised) gendered content: *beautiful* and *lovely* are more strongly gendered: stereotypically associated with (white) women rather than men, according to dominant social norms. Of course, this doesn't matter when the target social categories all share the same gender, so the use of such stereotyped valences won't affect any differential effects in response times there. This is why the use of such constructs is unproblematic in Connor et al's study 1a (where gender is held fixed). But the assumption that such stimuli can be used in later studies - where multiple social categories, including gender, are part of the target identity associations with which are being evaluated - is problematic, since gender stereotypes associated with the stimuli confound the valences of the attributes.

Moreover, this choice of stimuli is underpinned by an inadequate conceptualisation of intersectionality. If, as per the compounding models, one assumes that intersecting biases interact only in ways that sum, or multiply, the independent negative biases - such that the overall outcome could only be *more of the same biases* - then it is easy to overlook the extent to which existing biases, or constructs that reveal them, might themselves encode stereotyped assumptions. If one assumes that *nothing changes* when social categories and their associated stereotypes intersect, other than the magnitude of those expressed associations, then one can easily overlook the extent to which existing biases might reflect stereotypes that would shift in relation to other (perhaps non-paradigm) members of the target group. More concretely: if one uses stimuli that are apt for single gender studies into race and class, and if one supposes that any racial or class biases demonstrated on evaluative IATs will only change in magnitude, and in no other dimension, when other social categories that may activate biases (gender, age) are considered, then it is easy to overlook the ways in which gendered stereotypes might inform the associated evaluative concepts. Or, if - as per the category dominance hypothesis - one supposes that the negative biases associated with different social categories operate largely independently, interacting only insofar as the most salient swamps the effects of other potential biases, then it is easy to overlook the ways that biases associated with new intersections of oppression might shape the responses to those same stimuli, which may encode stereotyped confounds of the valenced attributes.

The extent to which this causes problems can be seen when we consider the findings reported on by Connor et al. Reporting on the effects of study 2, which, as described above, aimed to measure the evaluative associations (positive/negative) activated based on race, class, gender and age, a strong gender effect was found, such that positive terms were most closely associated with high class women. When one thinks of the classist and gender-based stereotypes associated with the evaluative constructs used - stereotypes concerning who is *lovely, beautiful*, and perhaps also *glorious, marvellous* - it is hardly surprising that these concepts are most strongly associated with high class women! It is impossible to tell if this finding reveals a genuine evaluative bias on the part of the participants, or is the result of the confounding effects of the gender stereotyped content of the stimuli.

Is there evaluative content that would have been more neutral, or at least "matched" in terms of class, gender (and age) associated content? Perhaps. But we might also return to the concern, articulated above, about exactly what these measures are aiming to reveal. The stated aim, recall, is to evaluate "the simultaneous effects of multiple intersecting social categorizations on the expression of intergroup bias" (2022, 4). Consider the depressingly

rich and varied stereotypes and biases that are associated with people assigned to different gender, race, class and age categories: the stereotype of the gang affiliated black (and male and working class) youth; the stereotype of the welfare dependent black mother; the stereotype of the white girlboss middle class young woman; the stereotype of overly assertive black working woman.[21]

Once the (depressing) heterogeneity and multidimensionality of these qualitatively different intersecting biases is visible, why think that a linear scale, shaped around positive or negative attitudes, will be well placed to capture much at all about how these biases interact? It is unclear that a linear scale of negative to positive attitude will be particularly informative, given these varied stereotypes.

These simple measures appear unlikely to inform our understanding of the rich heterogeneity of stereotypes at different intersections of oppression. Similarly, the compounding and category dominance models of intersectionality appear unlikely to capture *how* such stereotypes intersect. We have argued that these models have a detrimental effect in driving research questions, shaping experimental design, and generating problematic data. We have argued that the notion of intersectionality is being misconceptualised. Next, we suggest that these problems are also premised on assumptions about how concepts combine.

3. Compositionality

In this section we tease out the relationship between misconceptualisations about intersectionality and assumptions concerning the compositionality of concepts and associated stereotypes. We contend that several studies are premised on two mistaken assumptions:

1) Studies that focus on cognition about single social categories are investigating "simple" social concepts; studies that focus on cognition about multiple social categories are investigating "complex" social concepts (cf Goff and Khan 2013)
2) When putatively "simple" social concepts combine, the complex concepts inherit the associated stereotypes of their "simpler" constituent concepts.

We start with the assumptions involved in 2.

i. Concepts and combination

Classical analyses of concepts have it that a concept is structured definitionally, such that it provides the necessary and sufficient conditions for any item to come under that concept (Laurence & Margolis 1999, 9; Murphy 2002, 15). A challenge for such a view is the difficulty, in many instances, of articulating any such definition, which is supposed to be fundamental to a cogniser's possession of a concept. In contrast, prototype theories have been appealing: such views maintain that concepts are constituted not definitionally, but by their prototypes.[22] According to this way of thinking about concepts, they are (typically or always) "structured mental representations that encode the properties that objects in their

[21] See e.g. Dotson 2016, Williams 2014, Collins 1990.

[22] For summaries of the various arguments against the classical view, see Laurence & Margolis 1999, 13–27, and Murphy 2002, 16–24. "Prototype theories" refers to a broad class of similar views (Laurence & Margolis 1999, 28; Murphy 2002, 41)

extension *tend* to possess" (Laurence & Margolis 1999, 28, our emphasis). Prototype theorists typically understand the features of a concept to be weighted, such that certain features of a concept are given greater weight over other features (Laurence & Margolis 1999, 28; Murphy 2002, 42). Indeed, a "prototype" can be understood *as* a structured set of weighted features (Del Pinal and Spaulding 2018, 97; Gleitman et al. 2012, 422).

Let us assume (for now!) that, sometimes, simple concepts combine to make more complex ones. Given this, a key challenge for theories of concepts of any stripe - but particularly prototype theorists - is that they should have something to say about - and generate verifiable predictions for - how concepts combine (note: there may not be any one way in which they do this - perhaps multiple models are needed). The prototype theorist holds that a "new concept of some kind is constructed from the summary representations of the component concepts" (Murphy, 2002, 470). Difficulties arise, however, where the combined concept and its associated features are not easily derived from the component concepts. The difficulties for prototype theorists are as Gleitman et al. (2012) describe:

> Prototype theory says that the concept PET is itself represented as the set of stereotypic properties of pets and FISH is represented as the stereotypic properties of fish. Compositionality under prototype theory thus entails that to understand the linguistic expression "pet fish" we must compute the prototype as a function of the prototypes for "pet" (i.e. something like a golden retriever) and "fish" (i.e. something like a trout). Given these prototypes, the derivation of the prototype for "pet fish", which is neither dog-like nor trout-like, appears on its face to be intractable (2012, 429-430, see also Laurence and Margolis 1999, 38-44 and Murphy 2002, 443-475, for discussion of challenges about compositionality)

There are two distinct problems to emphasise. One problem concerns *emergent* features. A theory of concepts and their associated stereotypes should be able to accommodate features that (seemingly) complex combinations possess, but the (seemingly) simpler constituents do not. For example, although people rarely (if ever) think of "talks" when they think of either PETS or BIRDS, they often *do* think of "talks" when they think of PET BIRDS (Murphy 2002, 467). That is, TALKS is often associated with PET BIRD, despite not being associated with either PET or BIRD. A second problem concerns features that are *cancelled* under combination. MIGRATES is a property stereotypically associated with BIRD, but not a feature of PET BIRDS (Murphy 2002, 467). Any adequate theory of concepts must have some story to tell that can accord with a range of data about how, sometimes, features *emerge* or are *cancelled* in combinatorial contexts.

Why is this relevant to the discussion of intersecting biases? It isn't entirely clear what model of concepts underpins contemporary work on implicit biases in social psychology. Nor is it clear (to us, at least!) what model of concepts *should* be endorsed. However, what seems clearer is that the models of intersectionality by which their investigations are constrained leave little room for accommodating the two desiderata we identified for any adequate account of concepts: namely, accommodating the way that features can emerge, or be cancelled, in conceptual combinations.

In supposing (as the compounding models do) that negative biases will be the sum or multiplicative output of negative biases associated with each category, the authors seemingly assume that there will be no relevant, bias-related features associated with the complex categories that are not also a feature of each simple category. Or, in supposing (as the category dominance model does) that any negative effects will be driven by the dominant

category, the authors assume that the stereotypes associated with the complex category will be the same as those associated with each "simple" category (the question is simply which variety of social category (gender, or race, or class, etc) will be most salient in any particular context). These models obscure the idea that as social categories combine, qualitatively different content is associated with those combinations.

Consider the implications of this in terms of stereotypes that those experiencing multiple oppressions face. Black women report that they are often quickly labelled as "angry" if they argue for their beliefs or stand up for themselves. Particularly in the USA, but elsewhere too, the "angry black woman" is a well-known trope (Childs 2005). However, neither white women nor black men are typically stereotyped as angry. Whilst black men are often stereotyped as aggressive, the quality of anger in the "angry black woman" stereotype is distinctive, as an anonymous black lawyer, interviewed by Williams (2014), articulates:

> I am allowed to be passionate, even to demonstrate some level of anger, but it better not be personal. It better not be about me. If I become angry about anything personal, then that is perceived as being an angry black woman. (Williams 2014, 202)

Neither the concept WOMAN nor BLACK appear to contain the associated stereotype ANGRY, at least, not in the *personal* sense involved in the "angry black woman" stereotype. There is, therefore, no easy access to or explanation of that emergent racialised, gendered stereotype. It is occluded on a view where the features associated with a complex concept are those inherited from, or given by, the putatively simpler categories.

Nor do the compounding or category dominance models of intersecting biases appear to countenance the idea of bias related features of the compounding concepts being *cancelled* - or perhaps modified - in combinatorial contexts. The assumption is that a negative bias attached to simple concept C will be inherited by a complex concept incorporating concept C. Yet, other data suggests this assumption is at least sometimes mistaken. Consider the study by Remedios et al. (2011) into how expressions of likeability were shaped by race (black/white) and sexuality (gay/straight). In the abstract, one might suppose that a negative bias is associated with both black people, and gay people; these group memberships are both bases of marginalisation and oppression. One might therefore suppose that black gay men would be judged less likeable than, for example, black straight men. However, Remedios et al. found that the stereotypes associated with each social category concept interacted in non-linear ways, such that whilst white straight men were judged more likeable than white gay men, this pattern was not found for black men: black gay men were judged as more, not less, likeable than black straight men. Remedios et al. hypothesise that different aspects of sexuality-related stereotypes are activated when combined with different racial categories. Crucially - and recall this is on the assumption that there are ever "simple" concepts in play - the relevant negative stereotypes regarding the concepts GAY and BLACK do not simply inherit the features of the "simple" concepts.[23]

According to the compounding view of intersectionality, the negative biases attending GAY and BLACK would compound, leading to the prediction that (either in sum, or multiplicatively, so in excess of the sum) people perceived as both black and gay would experience the greatest degree of negative bias. Likewise, the category dominance model seems to assume that bias relevant features of the "simple" category will be inherited in the

[23] Of course, this is *not* to say that black gay men face no negative stereotypes!

conceptual combinations. These assumptions about conceptual combination appear both philosophically contentious and empirically unsupported.[24]

ii. “Simple” concepts; failures of intersectional thinking

In the above discussion, we have been conceding the idea that seems to be underpinning the investigation into intersecting biases: that complex social concepts (black working class elderly woman) are composed from the “simpler” concepts (black; working class; elderly; woman). And the thought is that the “simpler” concepts are ones that we already have some data about: from studies on gender bias, on age bias, on racial bias. These studies putatively investigate the biases associated with one social category. Yet recall the concern with which the paper started: that such empirical work likely overlooked the extent to which the studies targeted not *women* simpliciter (the concept WOMAN), or *black people* in general (the concept BLACK PERSON), but rather specific paradigmatic group members: *white* women, or black *men* (or rather, the concepts WHITE WOMEN and BLACK MEN) (Goff and Khan 2013, 375-376).

If Goff and Khan are correct in this claim, then there is little reason to suppose that the “single axis” studies *are* investigating a “simple” social category at all: assumptions about the race, class and age (and more) of the target social group members are already imported. Given this, researchers must be attentive to the possibility that such assumptions, or the psychological processes by which such assumptions are imported or resisted, might have an unforeseen influence on their studies.

For example, consider Thiem et al's (2019) research into whether associations between black men and danger-related concepts generalise to black people of other genders or ages. They used a sequential-priming measure to test the response times and error rates for respondents who were categorising either objects (as guns versus tools) or words (as “threatening” versus “safe”), following a series of primes. To ensure that the primes varied in race, age, and gender, Thiem et al. used a series of facial photos: “six each of Black and White girls and boys… and of Black and White women and men” (2019, 1429). They understood these to be “easily identifiable with respect to membership in the social categories under investigation” (2019, 1429), although they conceded that the adults were all "relatively young" (2019, 1436). Across three experiments, they found that "seeing Black face primes facilitated the rapid and accurate categorization of danger-related objects and words relative to seeing White face primes" (2019, 1435), thus suggesting a racial bias. Additionally, their findings suggest that children appear less strongly associated with danger than adults, and that females appear less strongly associated with danger than males.

These findings appear plausible. However, Thiem et al. also argue for more fine-grained conclusions, by manipulating their results in various ways. For example, they infer from the differing response times after child primes to those after adult primes that "racial bias was weaker after child primes than after adult primes", but that, since racial bias appeared to be significant for both adult and child primes, "racial bias emerged across prime age" (2019, 1429-30). This inference requires the assumption that participants identified black children *as* children as easily - and quickly - as they identified white children as

[24] See also Del Pinal and Spaulding (2018) for an argument that supports our contention here: namely, that salient statistical associations are unlikely to survive conceptual combination (whilst stereotypes that are central to a concept will do so). Implicit measures target statistical associations, so access stereotype or evaluative content unlikely to survive conceptual combination. For further discussion of conceptual centrality and implicit bias, see also Del Pinal, Madva and Reuter (2017).

children. If this were *not* the case, then at least some variations in response time might be attributable to difficulties in identifying the ages of black children. Since, as we have seen, black girls are very likely to be perceived as older than they are, more must be done if this possibility is to be ruled out. Such problems are unlikely to extend only to the adultification of black girls. We might consider, for example, that study participants will not only typically think of *white* women when they think of women, but that they might also therefore think of a woman as white more quickly than they can think of a woman as black.

In short, the problem is worse than that of failing to recognise the ways in which associated features may be emergent, or cancelled or modified, under conceptual combination. The problem is that of assuming that the features associated with one complex social category (white middle class woman) can be transposed onto all other members of those groups. This is precisely the problem that Crenshaw aimed to draw attention to: the problem of assuming that all members of a social category have similar experiences of discriminatory treatment (1989, 149). If this assumption is underpinning the research into intersectional bias, it is a grave failure to adequately grapple with the concept of intersectionality. There is a real need to better understand intersectional bias, but any research which ignores Crenshaw's insight is unlikely to offer meaningful findings. Such research is also very unlikely to do justice to the experiences of those, like the participants in Blake and Epstein's (2019) research, who experience intersectional oppression.

4. Conclusions

How should research proceed if it is to do a better job of investigating intersectional biases? Here we tease out eight recommendations. First, where implicit measures (such as the IAT) are used:

1. Clearly distinguish research questions concerning intersectional biases from those concerning biases of other kinds, including biases towards members of non-binary social categories (see section 2.a.i).

2. Where using IATs, avoid confounding valences with stereotyped attributes (section 2.b.ii).

3. Where possible, move beyond measures structured around “positive” and “negative” evaluative attitudes (see section 2.i, 3.i).

4. Scrutinise whether “simple” social categories are ever being measured, or whether participants import assumptions about multiple group memberships such that paradigm group members are the targets of studies (section 3.ii)

5. Attend to qualitative empirical literature to identify specific stereotype content that may target social groups facing different intersecting oppressions (section 2).

Paying close attention to the empirical literature on the experiences of people who are subject to intersectional oppression can facilitate the identification of stereotype content for investigation. For example, such literature articulates the “adultification” stereotype reported by the young black women in Blake and Epstein’s (2019) interviews, or the “personalised

anger" stereotype described by Williams' (2014) participants. Identifying whether there are implicit biases of these kinds requires using implicit measures guided by and structured in light of the qualitative literatures on experiences of discrimination.

Note that we do not make the strong claim that no quantitative measurements are useful for understanding intersecting biases. Nor do we at this stage recommend against the use of quantitative implicit measures such as the IAT; though that more radical conclusion may be supported if future attempts to use it to measure intersecting biases fare no better than the studies we have critiqued here. Given our critiques, its use certainly needs strong justification.

For now, we stress that any such experimental tools will have to be attentive to the concerns we have raised in this paper, in particular regarding how those tools might be distorted when used for measuring intersectional biases. Moreover, it is worth being explicit about the limitations of any quantitative measures, and contextualising their use alongside wider data about experiences of bias, stereotypes and oppression.Engagement with those literatures might also provide reasons for moving beyond implicit measures of attitudes and stereotypes. In doing so, we recommend:

6. Jettison the compounding and category dominance models of intersectionality.

7. Consider explicitly how non-linear expressions of bias might be measured, with both quantitative and qualitative measures.

8. Design measures that are specifically attuned to the qualitatively different experiences of oppression that those facing multiple forms of oppression might face.

Empirical work that genuinely grapples with the concept of intersectionality, as it was intended to be used, should consider these methodological recommendations. More generally, researchers should consider how they can better do justice to the insight from Crenshaw: that the experiences of all social group members may differ, and may qualitatively change depending on the intersections of oppression that they face. To fully embrace Crenshaw's claims about the transformational import of engaging in thinking about intersectional oppressions, we might consider how social psychological research would look if our starting questions were of a different kind, namely: what data would we need in order to dismantle the structures of oppression that so many find themselves at the intersections of? What would studies that gather this data look like?[25]

References

Atewologun, D., Cornish, T., & Tresh, F. (2018) Unconscious Bias Training: an assessment of the evidence for effectiveness. EHRC Research Report.

[25] Acknowledgements: thanks to audiences at the Nature of Bias conference, Claremont College, California; the Centre for Philosophical Psychology, Antwerp; and Hiljke Hänel's online series on biases in academia. Particular thanks also to Luca Barlassina, Jerry Viera, Jenny Saul, and Stephen Laurence, for valuable discussion of early ideas. Thanks also to the anonymous reviewers for the journal for helpful suggestions.

Beal, F. M. (2008). Double jeopardy: To be Black and female. *Meridians*, 8(2), 166-176.

Bergh, C., & Hoobler, J. M. (2018). Implicit Racial Bias in South Africa: How Far Have Manager-Employee Relations Come in 'The Rainbow Nation'?. *Africa Journal of Management*, 4(4), 447-468. https://doi.org/10.1080/23322373.2018.1522173

Bernstein S. (2020)The metaphysics of intersectionality. *Philosophical Studies* 177 (2):321-335 DOI https://doi.org/10.1007/s11098-019-01394-x

Bright, L. K., Malinsky, D., & Thompson, M. (2016). Causally interpreting intersectionality theory. *Philosophy of Science*, 83(1), 60-81 DOI: 10.1086/684173

Blake, J. J., & Epstein, R. (2019). Listening to black women and girls: Lived experience of adultification bias. Washington, DC: Georgetown Law Center on Poverty and Inequality. https://www.law.georgetown.edu/poverty-inequality-center/wp-content/uploads/sites/14/2019/05/Listening-to-Black-Women-and-Girls.pdf

Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): assessing automatic affect towards multiple attitude objects. *European journal of social psychology*, *38*(6), 977-997. https://doi.org/10.1002/ejsp.487

Brown-Iannuzzi, J. L., Cooley, E., McKee, S. E., & Hyden, C. (2019). Wealthy Whites and poor Blacks: Implicit associations between racial groups and wealth predict explicit opposition toward helping the poor. *Journal of Experimental Social Psychology*, *82*, 26-34. https://doi.org/10.1016/j.jesp.2018.11.006

Charlesworth, T. E., & Banaji, M. R. (2022). Patterns of implicit and explicit stereotypes III: Long-term change in gender stereotypes. *Social Psychological and Personality Science*, *13*(1), 14-26. https://doi.org/10.1177/1948550620988425

Childs, Erica Chito (2005). Looking Behind the Stereotypes of the "Angry Black Woman": An Exploration of Black Women's Responses to Interracial Relationships. *Gender & Society* 19(4): 544–61 https://www.jstor.org/stable/30044616

Collins, P. H. (1990). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. Routledge.

Connor, P., Weeks, M., Glaser, J., Chen, S., & Keltner, D. (2022). Intersectional implicit bias: Evidence for asymmetrically compounding bias and the predominance of target gender. *Journal of Personality and Social Psychology.* Advance online publication. https://doi.org/10.1037/pspa0000314.

Cooper, A. J. (1988). A Voice from the South. Oxford University Press.

Cox, W. T., & Devine, P. G. (2015). Stereotypes possess heterogeneous directionality: A theoretical and empirical exploration of stereotype structure and content. *PloS one*, *10*(3), DOI: 10.1371/journal.pone.0122292

Crenshaw, K. (1989) Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*: 139-169.

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. Stanford Law Review, 43(6), 1241–1299. https://doi.org/10.2307/1229039

Cureton, A. (2018), 'Hiding a Disability and Passing as Non-Disabled', in Adam Cureton, and Thomas E. Hill, Jr. (eds), *Disability in Practice: Attitudes, Policies, and Relationships*, Engaging Philosophy (Oxford University Press). https://doi.org/10.1093/oso/9780198812876.003.0002

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of experimental social psychology*, *40*(5), 642-658. https://doi.org/10.1016/j.jesp.2004.02.003

Del Pinal, G & Spaulding, S. (2018) Conceptual centrality and implicit bias *Mind & Language,* (2018). 33(1): 95-111. DOI: 10.1111/mila.12166

Del Pinal, G, Madva A & Reuter K. (2017) Stereotypes, Conceptual Centrality and Gender Bias: An Empirical Investigation *Ratio* 30(4):384-410. https://doi.org/10.1111/rati.12170

Dotson, K. (2016). Between rocks and hard places: Introducing Black feminist professional philosophy. *The Black Scholar*, *46*(2), 46-56. https://doi.org/10.1080/00064246.2016.1147992

Dovidio, J. F., & Gaertner, S. L. (Eds.). (1986). *Prejudice, discrimination, and racism*. London: Academic Press.

Epstein, R., Blake, J. and González, T., (2017) Girlhood Interrupted: The Erasure of Black Girls' Childhood (June 27, 2017). Available at SSRN: https://ssrn.com/abstract=3000695 or http://dx.doi.org/10.2139/ssrn.3000695

Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M., & Ambady, N. (2011). Looking the part: Social status cues shape race perception. *PloS one*, *6*(9), e25107. https://doi.org/10.1371/journal.pone.0025107

Gleitman, L., Connolly, A. & Armstrong, S. L. (2012). Can prototype representations support composition and decomposition. In Oxford handbook of compositionality (pp. 418–436). Oxford: Oxford University Press.

Goff, P. A., & Kahn, K. B. (2013) How Psychological Science Impedes Intersectional Thinking. *Du Bois Review: Social Science Research on Race*, *10*(2), 365-384. DOI:10.1017/S1742058X13000313

Goff, P. A., Jackson, M. C., Di Leone, B. A. L., Culotta, C. M., & DiTomasso, N. A. (2014). The essence of innocence: consequences of dehumanizing Black children. *Journal of personality and social psychology*, *106*(4), 526. DOI: 10.1037/a0035663

Greenwald AG, Brendl M, Cai H, Cvencek D, Dovidio JF, Friese M, Hahn A, Hehman E, Hofmann W, Hughes S, Hussey I, Jordan C, Kirby TA, Lai CK, Lang JWB, Lindgren KP, Maison D, Ostafin BD, Rae JR, Ratliff KA, Spruyt A, Wiers RW. (2022) Best research practices for using the Implicit Association Test. *Behavior research methods*, 54(3):1161-1180. https://doi.org/10.3758/s13428-021-01624-3

Haider, A. H., Sexton, J., Sriram, N., Cooper, L. A., Efron, D. T., Swoboda, S., ... & Cornwell, E. E. (2011). Association of unconscious race and social class bias with vignette-based clinical assessments by medical students. *JAMA*, *306*(9), 942-951. DOI: 10.1001/jama.2011.1248

Hartigan, J. (2013). Who are these white people?:“Rednecks,”“hillbillies,” and “white trash” as marked racial subjects. In White out (pp. 100-116). Routledge.

Henning, T. M. (2020). Racial methodological microaggressions: When good intersectionality goes bad. In *Microaggressions and Philosophy* (pp. 251-271). Routledge.

Hill Collins, P. (1990). Black Feminist Thought: knowledge, consciousness, and the politics of empowerment. New York, NY: Routledge.

Kang, S. K., & Bodenhausen, G. V. (2015). Multiple identities in social perception and interaction: Challenges and opportunities. *Annual review of psychology*, *66*(1), 547-574.

Kapusta, S. J. (2016). Misgendering and Its Moral Contestability. Hypatia, 31(3), 502–519. http://www.jstor.org/stable/44076489

Laurence, Stephen & Eric Margolis (1999). ‘Concepts and Cognitive Science’. In Eric Margolis & Stephen Laurence (eds.), *Concepts: Core Readings* (1999: 3–81). MIT Press.

Lei, R. F., & Bodenhausen, G. V. (2017). Racial assumptions color the mental representation of social class. *Frontiers in psychology*, *8*, 519. https://doi.org/10.3389/fpsyg.2017.00519

Madva, Alex (2017) Biased against Debiasing: On the Role of (Institutionally Sponsored) Self-Transformation in the Struggle against Prejudice, *Ergo* 4: 145-79. http://dx.doi.org/10.3998/ergo.12405314.0004.006

McGuire, Coreen, and Havi Carel, (2018) The Visible and the Invisible: Disability, Assistive Technology, and Stigma, in Adam Cureton, and David T. Wasserman (eds), *The Oxford Handbook of Philosophy and Disability*, Oxford Handbooks (Oxford University Press), https://doi.org/10.1093/oxfordhb/9780190622879.013.14

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition*, *19*(4), 395-417.

Murphy, Gregory (2002). *The Big Book of Concepts*. MIT Press.

Penner, A. M., & Saperstein, A. (2013). Engendering racial perceptions: An intersectional analysis of how social status shapes race. *Gender & Society*, *27*(3), 319-344.

Perszyk, D. R., Lei, R. F., Bodenhausen, G. V., Richeson, J. A., & Waxman, S. R. (2019). Bias at the intersection of race and gender: Evidence from preschool-aged children. *Developmental science*, *22*(3), e12788.

Petsko, C. D., Rosette, A. S., & Bodenhausen, G. V. (2022). Through the looking glass: A lens-based account of intersectional stereotyping. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspi0000382

Project Implicit (2011) https://implicit.harvard.edu/implicit/

Remedios, J. D., Chasteen, A. L., Rule, N. O., & Plaks, J. E. (2011). Impressions at the intersection of ambiguous and obvious social categories: Does gay+Black=likable? *Journal of Experimental Social Psychology*, 47, 1312-1315. https://doi.org/10.1016/j.jesp.2011.05.015

Rudman, L. A., Greenwald, A. G., & McGhee, D. E. (2001). Implicit Self-Concept and Evaluative Implicit Gender Stereotypes: Self and Ingroup Share Desirable Traits. *Personality and Social Psychology Bulletin*, *27*(9), 1164–1178. https://doi.org/10.1177/0146167201279009

Stramondo, J (forthcoming) The Ethics of Passing and Disability Disclosure in Academic Philosophy." In *The Bloomsbury Guide to Philosophy of Disability: Radical Resistances and Intersectional Imaginings*. Ed. Shelley Tremain. New York, NY: Bloomsbury Publishing

Steffensmeier, D., Ulmer, J., & Kramer, J. (1998). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. *Criminology*, 36(4), 763-798. https://doi.org/10.1111/j.1745-9125.1998.tb01265.x

Thiem, K. C., Neel, R., Simpson, A. J., & Todd, A. R. (2019). Are Black Women and Girls Associated With Danger? Implicit Racial Bias at the Intersection of Target Age and Gender. *Personality and Social Psychology Bulletin*, *45*(10), 1427–1439. https://doi.org/10.1177/0146167219829182

Truth, S. (2020). Ain't I A Woman? Penguin UK.

Williams, J. C. (2014). Double jeopardy? An empirical study with implications for the debates over implicit bias and intersectionality. *Harvard journal of law & gender*, 37(1), 184–242.

Xu, K., Nosek, B., & Greenwald, A. (2014). Psychology data from the race implicit association test on the project implicit demo website. *Journal of open psychology data*, 2(1), e3. DOI: http://dx.doi.org/10.5334/jopd.ac