Tackling Jencks' Red Hair Challenge: Methods and Pilot Evidence from the Black-White Infant Health Gap*

Henrietta Bada
University of Kentucky

Ariana Napier

James Wagner
Transylvania University

Geoffrey Fain Williams
Transylvania University

November 10, 2026

Abstract

Christopher Jencks noted that beyond its social cost, pervasive discrimination creates an empirical challenge - if society discriminates aggressively against a particular set of biological markers, for example red hair, then the genes for red hair and the damaging effects of discrimination are perfectly collinear. While in the most simplistic cases the challenge can be insoluble, there are several approaches used by researchers that can be used with great effect. Here we use causal inference to analyze the challenge and solutions, focusing on the disparity in Black-White infant health in the United States. One promising approach combines data on both socially perceived ancestry/identity and genetic evidence of ancestry to create conditional independence between identity and any biological drivers of low birthweight that are not socially mediated. We test the approach on a pilot data set and present power computations for a full-sized study. Overall, the methods and evidence underline the importance of tracking socially mediated causal pathways and improving methods in this area.

*Funding for online advertisements and random cash prizes came from the Transylvania University Office of the Dean and the Jones Faculty Development Fund. The work was done as Transylvania IRB Projects 2020-08 and 2021-05.

1 Introduction

Over the past 50 years, Harvard sociologist Christopher Jencks has been central in elucidating the complicated issues in linking genes and social outcomes. Among his points, one of the most famous is the "red hair" discussion of gene and discriminatory social environment interaction: "If, for example, a nation refuses to send children with red hair to school, the genes that cause red hair can be said to lower reading scores. This does not tell us that children with red hair cannot learn to read" (Jencks et al., 1972, pp.66-67). Clearly, in this example the genetic effect of red hair and the social effect of discrimination are perfectly collinear, making it impossible to disentangle the role of genes and environment.[1]

While the "red hair" story does a good job of setting up the basic challenges, in real world examples there are a number of factors that allow researchers to split out the effects. Our hope is that by approaching the issue using frameworks from the causal inference and potential outcomes literature we can help advance research in this area. Our key contributions are (a) analyzing the "red hair" problem using causal inference tools, (b) identifying and evaluating several existing empirical strategies where conditional independence may apply, and (c) presenting results from a pilot study that builds on a promising approach developed in the 1970s but fallen into disuse.[2] We focus on the Black-White infant health gap, specifically disparities in birthweight and rates of low birthweight (LBW) as well as maternal and child health (MCH).

We focus on LBW and MCH disparities for several reasons. First, it is an issue of great importance with substantial scholarly analysis over the last few decades. Second, for maternal and child health analysis can focus on a few measures of well-being; other areas where researchers look at discrimination outcomes can be complicated by issues of categorization, for example the issue of occupation and industry in labor markets. Third, it is an area where the range of peer-reviewed

[1] Another place where he made the same point but without using "red hair" as an example is Jencks (1992, pp.96-100). In this writing, about 20 years after his first discussion, he also referenced the possible role of social interactions, paralleling similar contemporaneous discussion in Manski (1993). The "red hair" issue comes up with some frequency in work looking at social and genetic interactions, for example in Meyer et al. (2023).

[2] From Loehlin et al. (1973) and Scarr et al. (1977), referenced in Jencks and Phillips (2011, page 90).

research in the last few decades includes papers that focus on (a) the direct effects of genetics (Anum et al., 2009; Frey et al., 2016),[3] (b) the direct effects of discrimination (Geronimus, 1992), and (c) research focusing on mediating factors such as access to resources (Collins and Thomasson, 2004). It represents a particularly good arena to think about the various issues of "red hair," and we believe that much of the theoretical analysis has the potential to be applied in other contexts.

The "red hair" issue is knotty and involves ideas and problems studied by a broad range of disciplines. Most obviously, it involves questions of discrimination and social identity, but other ideas from healthcare and genomics as well. Economists have played a significant role in developing approaches in all of these areas. For at least seventy years, for example, economists have been analyzing discrimination, deploying a range of tools (Becker, 1957; Bertrand and Duflo, 2017; Lang and Spitzer, 2020; Chelwa et al., 2022). And in the last few decades work with genomics data has been increasingly incorporated into economic analysis (Benjamin et al., 2012, 2024). Fields such as social economics and the study of social interactions are further evidence of this process (Manski, 1993; Benhabib et al., 2010).[4]

Of particular importance are methodological innovations over the past few decades that help to measure the component causal paths of discrimination and genetics. With regard to understanding genetics, improvements in technology and a collapse in costs make it feasible to record data on genotypes on a large scale. This has revolutionized a range of fields, including work in the social sciences. There is significant awareness of the potential role of socially mediated effects (sometimes termed "environmental pathways") with powerful methodological approaches continuing to develop (Benjamin et al., 2024).

The causal pathway of discrimination was relatively simple to directly measure in the early 20th century, as a substantial portion of the white population viewed racial prejudice as quite normal, discriminatory laws were common, and discriminatory policies by organizations were explicit. After the Civil Rights movement successes in the 1950s and 60s, self-reported prejudice and explicit discrimination have gone down considerably, with debate by social scientists as to how much is an actual reduction in discrimination and how much is simply less

[3]For other examples, see David and Collins Jr (2007).

[4]We do not directly look at social interactions, or endogenous effects of social identity (Manski, 1993), in this analysis. Some specific issues regarding this will be discussed at relevant points in the text.

transparency about discrimination and prejudice. A number of measures have been developed that appear to elicit reliable data on prejudices (Quillian, 2006; Bertrand and Duflo, 2017; Carlana, 2019). Measures of individual prejudice aggregated at the county or metropolitan area level vary considerably across the United States, and variation in these measures can predict outcomes across geographies (Charles and Guryan, 2008; Stephens-Davidowitz, 2014; Hehman et al., 2019).

We now move from the general questions involved in the "red hair" problem to the specifics of Black-White LBW and MCH disparities.

1.1 The Black-White Gap in Infant Mortality and Morbidity

The Black-White gap in infant mortality and morbidity can be tracked to some of the earliest reliable data in the United States (Shin, 1975), and despite shifting patterns over time continues to be significant. Recent research estimates that reducing black infant mortality to white levels would save over four thousand lives per year (Speights et al., 2017). While infant mortality is a relatively visible and measurable pattern of loss, research indicates that disparities in preterm birth, low birthweight and other measures of morbidity are likely signs of more difficult to measure costs of similar or greater magnitude. There is increasing evidence that in utero insults and early childhood morbidity are associated with a range of negative longterm outcomes in lifetime health, education and employment (Almond and Currie, 2011; Almond, 2006). While research continues to shed light on the costs of low birthweight and other measures, there is no doubt that they are substantial (Almond et al., 2005; Oreopoulos et al., 2008). Given its importance, research into causes of racial disparities in both infant mortality and morbidity is active and ongoing, with researchers looking at a range of genetic and social causes.

In the context of maternal and child health outcomes, there are a number of theories as to how discrimination creates disparities (none of which are exclusive of the others). Health researchers have found it helpful to divide causes into upstream, midstream and downstream as a way of structuring analysis of the different issues (Bharmal et al., 2015; Braveman, 2023). For example, looking at preterm birth disparities, Braveman et al. (2021) point to access to income, wealth and education as midstream causes which may then contribute to downstream causes such a prenatal care, preconception care and gestational diabetes. Dis-

crimination based on social identity represents the critical upstream cause driving multiple midstream causes (such as income) and through them additional downstream causes (such as prenatal care).

What is particularly interesting is that education, income and general access to healthcare seem to have steadily diminished in explanatory power and in studies using data from the last several decades the majority of the gap cannot be explained by background characteristics (Lhila and Long, 2012; Elder et al., 2016). Additionally, disparities between European-Americans and first generation African immigrants are minute, but second and third generation African immigrants show patterns similar to the general African-American population, suggesting the importance of cumulative factors (David and Collins, Jr., 2007). Health outcomes for Black women in the United States deteriorate with age much faster than health outcomes for whites and other ethnicities, providing further support (Geronimus, 1992).

There are a broad range of possible causal pathways by which discrimination could work. Geronimus sketches out the possibilities

> This can take concrete, observable forms, such as diminished access to health services and health education or residence in segregated neighborhoods with excessive exposure to environmental hazards. It may also be the product of evolving concepts of psychosocial risk. For example, there are empirical indications that prolonged, effortful, active coping with social injustice may, itself, exact a physical price. (page 210)

One pathway may simply be a ubiquitous lack of respect. In Collins Jr et al. (2004) the researchers interviewed 312 Black women who had recently delivered infants about experiences with racism. Sample questions included "Whites often assume that you work in a lower status job than you do and treat you as such" and "You are treated with less dignity and respect than you would be if you were White". Self-reported racism over the lifetime in multiple domains predicted an odds ratio of very low-birthweight of 2.6 (CI 1.2-5.3).

In theory a causal pathway for the disparities is social or cultural influences within a particular identity group.[5] However, we do not see evidence that it plays a vital role (Braveman et al., 2021). While some behaviors around maternal and child health that have a social compo-

[5]Such pathways might be modeled using ideas from the social interactions literature (Manski, 2000).

nent are linked to negative health outcomes, the effects are generally uncertain and small relative to the disparities. For example, while not relevant to birthweight, patterns of infant sleeping (supine vs. prone) do increase the rate of SIDS and vary between Black and white parents. With that said, the base rate of SIDS is small enough the overall effect is at most 5% of the disparity in mortality, and less than 1% of the overall disparities in morbidity.[6]

The rest of the paper proceeds as follows. In Section 2 we present an analytical framework, using ideas from the causal inference and potential outcomes literatures to look at the "red hair" challenge and several strategies that offer traction against it. In Section 3 we discuss some of the issues involved in empirical measurement in the context of Black-White LBW and MCH disparities, as well as the data of the pilot study. In Section 4 we present the results of a pilot study, based on roughly 100 respondents who were recruited via African-American genealogical societies and advertising on Facebook. Combining data on social identity, genetic ancestry and pregnancy outcomes, we attempt to split out the two pathways of the "red hair" effect.[7] In Section 5 we use sample statistics from the pilot study to compute overall power requirements for a larger study. We conclude by highlighting what seem to be important prospects for future research.

2 Analytical Framework

In what follows, we use the tools of causal inference to set out a framework for "red hair" situations. In particular, we pull from both the potential outcomes (PO) and directed acyclic graph (DAG) approaches (Imbens and Rubin, 2015; Pearl, 2009; Imbens, 2020).

2.1 Using a Directed Acyclic Graph

Because a key issue in the "red hair" scenario is the two possible causal pathways, we begin by illustrating with a simplified directed acyclic

[6]The infant mortality rate for Blacks was 1.08% and 0.49% for white in 2017 (Singh and Stella, 2019). In contrast in the same period the overall rate of SIDS was 0.046% for whites and 0.103 for Blacks while the estimated difference in the rate of prone sleeping was about 20-30% in Black households and about 10-15% in white households (Colson et al., 2009; Huang et al., 2024).

[7]Our approach in the pilot study builds on ideas from roughly 50 years ago, in Loehlin et al. (1973) and Scarr et al. (1977), referenced in (Jencks and Phillips, 2011, page 90).

graph (DAG) in Figure 1.

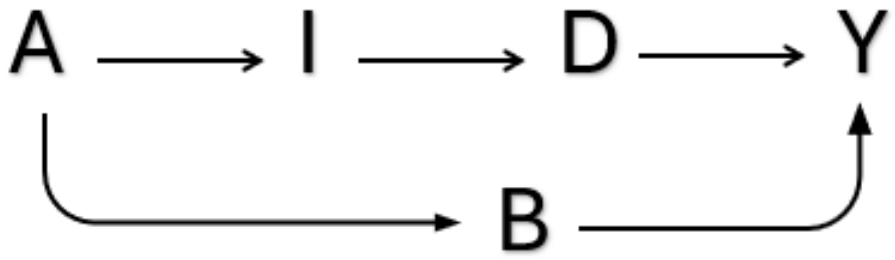


Figure 1: A schematic model of the causal links of "red hair" phenomenon - ancestry affecting genes for both social identity and direct biological impacts.

Walking through each of the five nodes of the DAG in turn, we map them to the Jencks' example of a society where children with red hair (A, I) are not allowed to go to school (D) and hence are illiterate (Y), creating the illusion that children with red hair are genetically incapable of reading (B). At this stage, we keep to the binary logic of his example and fix it that each node can only take the value of zero or one (later we will allow a range of real values for several of the nodes).

The left-most node (or *source* node in DAG terminology) is A_i, individual i's ancestry. In the context of a "red hair" situation, this has two parts: (a) being descended from members of a specific group of individuals $\mathbb{A}$ who (distinctly in the society) have red hair, and (b) possessing the group of genes that determine the visible red hair trait. Fixing $A_i \in \{0, 1\}$, $A_i = 1$ would mean being descended from red haired individuals.[8]

The node $I \in \{0, 1\}$ represents the social identity of the individual. Note that in this case, not only are values limited to two options, there is a perfect correlation between A and I: $I = 1$ if i has red hair, $I = 0$ otherwise.[9]

Social discrimination $D \in \{0, 1\}$ is the child node of I. Again, we limit the outcomes to zero or one, and assume perfect correlation: If

[8]For the remainder of the discussion of the pure binary case we suppress the index.

[9]When we loosen the restrictions on A and I and allow (to take values between zero and one, the relationship between them will become more complicated.

$I = 1$ then $D = 1$ and the individual has no access to schooling, or $D = 0$, with full access to education.

The outcome is symbolized with $Y \in \{0, 1\}$. Keeping the chain of binary variables to a value of 1 in the "red hair" path, illiteracy would be marked with $Y = 1$, literacy $Y = 0$.

The above discussion reviews the path $A \to I \to D \to Y$. In the most extreme and simplest example of Jencks' story, there would a chain of socially constructed causality such that $E(I|A = 1) = 1$, $E(D|I = 1) = 1$ and $E(Y|D = 1) = 1$.[10] That is to say, any individual with $A = 1$ would, with 100% probability, (a) have red hair, (b) lack access to schooling, and (c) be illiterate.

The DAG also posits a second path from A to Y of $A \to B \to Y$. The conjecture behind this path is that A might be correlated with biological traits (B) that could have an effect on the outcome Y that is not mediated by social discrimination. If we are focused on the role of discrimination on Y, then $A \to B \to Y$ is a *backdoor path*. The standard method for handling backdoor paths is to control for either A or B. More than 50 years ago, that would basically be impossible; we had no information on A separate from I. With improvements in genomics, it is possible but challenging. To understand why, we now walk through the "red-hair" problem using the language of potential outcomes.

2.2 Potential Outcomes

Having outlined the basic issues with the help of a DAG, we can start to use the potential outcomes (PO) framework to think through identification issues and strategies (Imbens, 2020).

If we keep to a model where all variables are binary, the potential outcomes framework for analyzing the effects of discrimination (D) and a biological channel without a social component (B) is relatively simple.[11] We can think of each one having a treatment effect, τ_D and τ_B, defined in turn as:

[10] And since each variable takes only values $\{0, 1\}$, conditional variance would be zero.

[11] Since we are comparing two "treatments" - discrimination D and a biological mechanism B - the model would be a multifactorial experiment. There is an extensive literature setting up such experiments with randomization, including Fisher (1949). Some more recent discussions include Barnard et al. (2003) and Lopez and Gutman (2017). This is the model for research design, but of course a true experiment is not possible in this context.

$$\tau_D = E(Y|D=1) - E(Y|D=0)$$

$$\tau_B = E(Y|B=1) - E(Y|B=0)$$

In order for these treatment effects to be reliably causal effects, in each case we need the relevant treatment to be conditionally independent from other attributes, which is to say

$$D \perp\!\!\!\perp ((Y|D=1),(Y|D=0))|A,B,I$$

$$B \perp\!\!\!\perp ((Y|B=1),(Y|B=0))|A,D,I$$

We can immediately see that in the case of binary attributes and perfect correlation these are impossible, almost nonsensical constraints; as set up, $A = I = D = B$ and in particular we get the table:

	$D=1$	$D=0$
$B=1$	individuals with red hair	no examples
$B=0$	no examples	all other individuals

In this worst case scenario, the lack of observations in the $D=1, B=0$ and $D=0, B=1$ cells, combined with the $Y=1$ outcomes for all cases with $D=1, B=1$, means we cannot disprove the possibility that $Y=1$ outcomes are driven by B, even if there is *no* direct, non-mediated, biological pathway.

2.3 Allowing Non-Binary Values

As reviewed in the previous subsection, the binary "red hair" model is structurally simple but analytically insoluble. Here we adapt it to make it more realistic, which adds some complexity but also furnishes strategic openings for researchers.[12]

Working, as in the binary version, from ancestry A to outcome Y:

For ethnic ancestry A and genetic inheritance, ancestry A_i can now take any value $[0,1]$, measuring the percent of individual i's ancestry from a historical population $\mathbb{A}$.

Ethnic identity I is separate from ancestry, and remains binary: $I_i \in \{0,1\}$.[13] For any ancestry A_i it is a realization of a Bernoulli

[12] We revert to indexing by individual for this section.

[13] Identity can also be expanded to include more than two categories. So long as the focus is on comparing two categories, this would not change the overall analysis.

process with $p_i = f(A_i)$ where, with probability 1, $I_i = 0$ if $A_i = 0$, $I_i = 1$ if $A_i = 1$, and $f'() > 0$.

The variable B, a *hypothesized* biological vulnerability correlated with ancestry A but *not* socially mediated, can now be more precisely defined. We normalize values of $B_i \in \{0, 1\}$. As with variable I_i, B_i is a realization of a Bernoulli process with $p_i = A_i$, so that $E(B_i) = A_i$ and, with probability 1, $B_i = 0$ if $A_i = 0$ and $B_i = 1$ if $A_i = 1$.[14] That is to say, there is a particular gene which is correlated with the amount of group $\mathbb{A}$ ancestry; as the proportion of this ancestry increases, the probability of the gene's presence increases.

It is worth taking a moment to take stock: For the three variables A, I, and B, the only change from the original binary model is that A can now take values across the range $[0, 1]$. Given the randomness of the overall process of meiosis, specifically Mendelian segregation, in genetic inheritance (Benjamin et al., 2024), it is now the case that I_i and B_i are no longer necessarily identical. Even more than that,

$$I_i \perp\!\!\!\perp B_i | A_i$$

It also becomes possible the conditional independence assumption holds with regard to B;

Proposition 1. *For any individual i with $0 < A_i < 1$,*

$$B_i \perp\!\!\!\perp ((Y_i|B = 1), (Y_i|B = 0))|A_i \tag{1}$$

Proof. From construction (based on the mechanics of Mendelian segregation) that B_i is Bernoulli with $p = A_i$.

□

Thus conditioning on genetic ancestry allows us to separate out the possibility of a non-socially mediated pathway.

Another approach to separate out socially and non-socially mediated pathways is family fixed effects in genome-wide association studies (GWAS) (in some variants called family-based genome-wide association studies or FGWAS (Veller and Coop, 2024)). There are a range of

[14] An alternative model might be $B_i \in [0, 1]$ where $B_i = A_i$. This is not identical to a Bernoulli model of B_i but quite similar. In theory it is possible that B_i is based on an interaction, where multiple SNPs can push the value to higher levels. This is not developed here but note that the interactive term is even easier to identify than a simple Bernoulli, the probability would be exponential with n determined by the number of SNPs necessary to create the effect, so p_i^n. As A_i increases, p_i^n increases exponentially.

approaches but the overall idea is straightforward - the use of family fixed effects to control for a range of environmental and social stratification confounders when looking at the genetic pathway. An example specification (Guan et al., 2025, based on model 1, page 1053) for a sibling i from family j is:

$$y_{ji} = \delta g_{ji} + \alpha g_{par(j)} + \varepsilon_{ji}$$

Here y_{ji} is the individual-level phenotype (i.e., measures of outcomes such as BMI), g_{ji} is the genotype data and $g_{par(j)} = g_{p(j)} + g_{m(j)}$ are genetic controls for the parents.[15] This makes δ the direct genetic effect; α the NTC or non-transmitted coefficient (this is a control for a number environmental and gene interaction pathways); and ε_{ji} the error term.[16]

Proposition 2. *For any individual i from family j*

$$A_i \perp\!\!\!\perp D_j, I_j | A_j \tag{2}$$

Proof. As with Proposition 1, the proof depends on the mechanics of meiosis: A_i is a completely independent draw from within the possibilities of A_j.

□

Controlling for family genes A_j inherently helps to control for family-level social stratification and environmental effects, allowing us to better understand the effect of individual genes A_i. With regard to "red hair" questions they are focused on the genetic pathway, not the socially-mediated pathway. By looking across thousands or tens of thousands of SNPs they have the potential in large samples to sensitively measure non-mediated genetic effects. However, this comes with a loss of power for understanding the socially mediated paths for any particular sample size.

Before we move on from identity I_i, a brief discussion of the role of social interactions (Manski, 2000) is worthwhile. This would be a pathway where an individual adjusts their behavior because of behavior by another person they have some social tie to (such as a similar

[15] To keep consistency within this text the meaning of the indexes i and j are swapped relative to the original.

[16] There are a range of methods developed recently that support imputation of family genotypes for observations where only one member of a family is observed (Young et al., 2022; Guan et al., 2025).

identity). Recent studies in this area have tended to focus on the possibility of conformity (where an individual wants to be more like other people) or spillovers (where an individual learns something from other people). In large part because of the complexity of the empirical issues in the overall field of social economics, researchers have largely avoided a "one size fits all" approach (Athey and Imbens, 2017). Direct investigation of a social interactions pathway is beyond the scope of this paper.[17]

Moving to discrimination, D_i, we can think in terms of measurable variation in this, conditional on $I_i = 1$. A powerful tool that has been used in a variety of ways is looking at variation in discrimination across geographies. Across a set of geographies $S_G = \{0, 1, 2, \dots G\}$ each geography has an expected discrimination, $E(D_i) = d_{gi}$ for any individual with $I_i = 1$ living in g. Variation in discrimination has limits, so $d_g \in [\underline{d}, \bar{d}], 0 <= \underline{d} < \bar{d} <= 1$. If we believe that individuals' geography of residence exogenous, or independent of A_i, B_i after controlling for observables, a more limited version of the conditional independence assumption can be used:[18]

Proposition 3. *Given* $g_i \perp\!\!\!\perp X_i$ *then*

$$D_{g_i} \perp\!\!\!\perp Y_i(D_i) | I_i, g_i, D_g$$

Proof. The conditional independence follows from the independence of g_i from X_i. □

Moving to the ultimate (or *sink*) node, the outcome Y_i, we make the simple point that measuring inputs is an important and possibly

[17] Manski (1993) presents a linear-in-means equation for social interactions,

$$E(y|x, z) = \alpha + \beta E(y|x) + E(z|x)'\gamma + x'\delta + z'\mu$$

(equation 1, page 533). In the context of low birthweight and many other issues, there is no clear story where the outcome y is affected by a behavior associated with identity x that could not be identified by researchers, measured, and incorporated in the vector of observable individual traits z. In contrast, without measurement, discrimination and any conjectured non-mediated biological cause would become correlational effects, represented in expression $x'\delta$ (where x are group level variables such as school or ethnicity). With measurement at more granular levels they can be controlled for fully, moving to the last expression $z'\mu$.

[18] Here we slightly shift notation in keeping with standards for variable treatment levels (Angrist and Imbens, 1995; Hirano and Imbens, 2004).

complementary empirical approach. Just as in Jencks' original example, if there are clear differences in access to key inputs (education), this can support some of the other strategies listed. A production function[19] can be helpful to frame the pathways by which discrimination D_i can affect the outcome:

$$Y_i = h()$$

The inputs to $h()$ will include resources such as education and income, access to healthcare, environmental stressors, etc. While the inputs themselves are fairly constant across studies and contexts, which of them are observed and which are not will vary considerably.

Discrimination can both reduce observable positive inputs - employment discrimination can reduce income, for example - but also can affect unobservables - by increasing exposure to hostile treatment.

2.4 Examples from published research

Here we briefly review uses of these methods in previous research, particularly focusing on research on Black-White MCH and LBW disparities when available.

The idea of Proposition 1, using a rigorous measure of A to control for any non-mediated pathways, comes from Loehlin et al. (1973) and Scarr et al. (1977), referenced in (Jencks and Phillips, 2011, page 90). Loehlin et al. look at a pool of White and Black adolescents where they have both IQ and blood type data. They compute the correlations between IQ and frequency of blood types associated with European ancestry in the two social groups (White and Black) separately, and find correlations to be generally low and frequently negative. Scarr et al. computed a relative probability measure (pages 71-72) of African ancestry based on blood type genetic data (Table 3, pages 78-79), and also found little to no relationship between the genetic measure and outcomes, after controlling for social identity. We are not aware of any other studies using these methods generally, or looking at LBW and MCH issues specifically.

An example of FGWAs (touched on above as Proposition 2) as an approach to handling population stratification and possible discrimination in another context is Wang et al. (2025). Researchers look

[19] In the context of maternal and child health, functions such as Rosenzweig and Schultz (1983); Corman et al. (1987); Reichman et al. (2009); Corman et al. (2018).

at the role of genetic ancestry among residents of Mexico city. They found that FGWAS drastically reduced the importance of genetic effects on educational attainment. We are not aware of any studies using FGWAS to look at LBW and MCH disparities.

Similarly, the idea of Proposition 3 has been used implicitly by a number of researchers, frequently with proxies of discrimination. For much of the 20th century, the role of discrimination in racial birthweight disparities could be inferred from variation in access to resources across geographies. In Collins and Thomasson (2004), visible measures of resource access had a strong relationship with low birthweight. They estimate the equation (equation 1, page 755)

$$lnIMR_{it} = (lnX_{it})\,\beta + \gamma_t + \epsilon_{it}$$

separately for both white and Black populations, where IMR is infant mortality rate specific to race, X is state-year characteristics (specific to race) and γ is year fixed effect. Even as the absolute gap in infant mortality fell from 30 per thousand births in 1940 to 13.8 in 1970, the relative rate increased. Measured access to resources explains a substantial percentage of the relative gap from 1920 to 1970, but goes from 82% of the gap in 1920 to 33% in 1970, with unidentified factors thus playing an increasingly important role.

Chae et al. (2018) use frequency of racist Google searches in a metropolitan area as a measure of racism. The overall effect, with controls for mother's age, geographic region and several area covariates, is estimated at 1.05 (1.02, 1.07) for every standard deviation increase in area prejudice (Chae et al., 2018, Model 3, Table 3, Page 53).

Orchard and Price (2017) use county-level tallies of residents' scores in bias tests – both implicit and explicit racial prejudice – from Project Implicit as a measure of county level of prejudice. Using weighted least squares they estimate the equation

$$y_i = \gamma z_c + \alpha\,(z_c \times \iota_i) + \delta w_c + \beta x_i + \varepsilon_{ic}$$

where y_i is birth outcome (either preterm birth or low birthweight) for mother i, z_c is county prejudice, ι_i is an indicator for mother's race, w_c is a vector of county-level controls and x_i is controls for the mother and the pregnancy. They find that shifting two standard deviations down in implicit racial prejudice is associated with a 14% fall in the LBW gap. A similar shift in measures of explicit prejudice from the

same source is associated with a 22% fall in the LBW gap. In both cases these include controls for maternal education and risk factors.

In both cases the researchers find ways to proxy the overall level of discrimination against people with Black social identities as it varies across geographies. It seems plausible that other relevant factors are either controlled for or are independent of the geography.[20]

In Collins, Jr. et al. (2002) researchers look at two paired groupings of US and foreign-born White and Black families. They do not directly measure or proxy discrimination, but the patterns suggest this plays a critical role. For each of the four groups they track how pregnancy outcomes develop across two generations - mothers and their daughters. For the daughters of both US and foreign-born White mothers, average birthweights against the previous generation show an increase that is statistically significant and of 45 grams or more. For the families of US-born Black mothers average birthweight, which was the lowest by a meaningful margin, births to their daughters show a statistically significant increase of 17 grams. For the daughters of foreign-born Black mothers, average birthweight drops by 56 grams relative to the previous generation, and the rate of LBW increases from 6.7% to 9.6%, moving from 53% of the US-born average rate across two generations of 12.6 to 76%. Thus, controlling for A and B (by tracking within family changes), a shift to the US environment is shown to have deleterious effect on birth outcomes, presumably via discrimination.

3 Data

In the next two sections we present data and results from a pilot study. This requires developing the ideas outlined above for the real-world issue of Black-White disparities in infant health and specifically issues in measuring each one.

3.1 Background

The process of constructing and implementing a rigorous measure of actual ancestry A only began in the past few decades. In particular,

[20] As noted above, an example from outside the context of maternal and child health, Chetty et al. (2020) use family fixed effects and other methods to further reduce endogeneity.

improvements in genomics and efforts to sample from across many human populations worldwide have been important, even though there remain limitations. In the context of racial discrimination in America, a plausible and relevant specification of the concept might be "what percentage of individual i's ancestors in 1400 were living in Sub-Saharan Africa?" By processing a sample of DNA from saliva, it is possible to make precise estimates how much of an individual's DNA matches DNA from individuals in samples from different populations worldwide (Noto et al., 2013; anc, 2022). With that said, while reasonable assumptions can be made, there is no perfect information about how much the population of a particular geography today matches the population of 1400. Additionally, different groups have been sampled at vastly different rates, meaning match accuracy can vary substantially by ancestry group. A more significant issue is that autosomal DNA is inherited in linked chromosomal blocks rather than as independently sampled base pairs. Each parent passes on an average 50% of their nuclear DNA, but the randomization is not at a granular level such as base pair or allele. During meiosis, a few dozen crossovers make each transmitted set of chromosomes a mosaic of the parent's two chromosome copies.[21] This means that while individuals are almost certain to have some DNA from each of their great-grandparents or even each great-great-great-grandparent, when you go more than 6 generations back, it is increasingly likely that some ancestors are "missing" in the DNA . If we are trying to construct a measure of ancestry based on location on the eve of early modern history, 500-600 years ago, we are looking at more than 10 generations, where the probability is high that 50% or more of the individual's ancestors have become invisible (Reich, 2018, Chapter 1).

Despite these limitations, genomics offers a remarkable tool for tackling the red hair issue. Instead of a binary ancestry attribute $A \in \{0, 1\}$ that is defined by identity, we now have an $A \in [0, 1]$ and methods that allow us to measure the latent details of A.

Economic research on identity (I) has become more common (Akerlof and Kranton, 2000; Dahis et al., 2019).[22] In this context I

[21] That is, the odds of a particular base pair being passed on by parent A is very heavily correlated with the base pairs right next to it.

[22] There is a small but growing body of research on the endogeneity of race (Dahis et al., 2019; Bodenhorn and Ruebeck, 2003). This is of great interest generally, with persuasive estimates that 16.6% of the Black population was "passing for white" before World War II. However, estimates for changes in racial identity from the past 50 years are less than 2%

would mean the presence or absence of an identity as Black or African-American. As we will see in more detail, this is highly but not perfectly correlated with an ancestry linked to Sub-Saharan Africa, with enough disconnection to allow analysis. Unlike A, I seems to be quite "lumpy," even if not perfectly binary. Self-reported ethnic identity is generally used. There are complexities, including the fact that there are individuals who declare a Black identity that is combined with other identities such as Hispanic ethnicity or Multiracial (with White, Asian, other ethnic identities included).[23]

Discrimination was not measured in the pilot study. American citizenship was used a screen in an effort to limit variation in environment. Otherwise the assumption was made that $D_i = I_i$. This is an obvious area for future work to develop.

Next would be B, biological causes that are not socially mediated. In other contexts, there are known examples of health outcomes Y where biological pathways correlated with A but not affected by D play a significant causal role. Perhaps the most well-developed example is Tay-Sachs disease.[24] For Tay-Sachs, the genes in question have been observed and the pathway is well understood. In the context of birthweight disparities, nothing like this has been found. Investigation has followed the standard genomic method of searching for SNPs that show strong statistical associations with poor outcomes, in particular those that can be matched with a plausible mechanism such as genes regulating inflammation. A number of candidate SNPs have been put forward (Anum et al., 2009; Frey et al., 2016) but there is no consensus on which if any play a role.

In the absence of any specified SNPs, as shown in Proposition 1 we can assume that any possible non-mediated genetic cause increases linearly with Sub-Saharan African ancestry. That is to say, for an individual mother i, $P(B_i = 1) = E(B_i) = A_i$. Importantly, once we control for A_i, $E(B)$ is independent of I, D, Y.

With regard to outcomes, respondents were asked a number of questions about birth outcomes including birthweight and if there were

of any group and hence unlikely to have an effect on the results discussed here (Saperstein and Penner, 2012).

[23]The current census uses seven large groupings, and roughly 3010 minor groupings `https://www.census.gov/newsroom/blogs/random-samplings/2023/09/what-you-should-know-detailed-dhc-a.html`.

[24]For a discussion of the mechanism of Tay-Sachs, see Gravel (2003). For a discussion of the discriminatory response to Tay-Sachs after it was first identified see Reuter.

complications.

3.2 Collection

In May-June 2020 and May-September 2021 we recruited respondents to fill out a confidential questionnaire online. As an incentive, randomly selected respondents were awarded $100. The initial part of the form screened respondents as to whether or not (a) they were American citizens, (b) they had been pregnant, (c) they had genetic testing done to determine their ancestry and (d) that ancestry was determined to be at least 2.5% African.

We used a number of outreach methods to recruit respondents, including advertising on a range of social media and reaching out to African-American genealogy groups and services. The outreach media gave a brief description of the project and a link to the Google Form questionnaire. By far the most effective method was paid, targeted advertising on Facebook. In August 2021, Facebook began to reject ads mentioning ethnicity and other social characteristics, and our respondents after that were almost completely women with $< 1\%$ African ancestry.

All data was self-reported via Google Forms, stored in a Google Sheet spreadsheet. After identifying information was removed, the responses were downloaded to a laptop computer and the analysis was done using Stata/MP 19.[25]

Out of 153 total responses, only 98 included birthweight as well as all 6 regressors (genetic ancestry, self-identification, perceived African ancestry, income, education and age at pregnancy).

[25]Regressions were performed using the command **regress**. Standard errors were computed using the **hc3** option when unweighted and the **rob** option when weighted.

3.2.1 Birthweight

Respondents were asked the birthweight of their most recent pregnancy in English units (pounds and ounces) and this was converted to grams.

3.2.2 Genetic Ancestry

Recruited respondents indicated that they had already gotten their genetic ancestry analyzed by a retail service. They were asked to report the percentage of ancestry that was African, European, Asian, Indigenous or Native American and Other. Responses that did not sum to approximately 100 were dropped.

3.2.3 Social Identity and Perception

Respondents were asked to note any and all of the standard U.S. Census racial categories ("White", "Black or African-American", "American Indian or Alaska Native", "Asian" or "Other") with which they identified. For the analysis in the main paper, the measure of "any self-identification" is used which corresponds to those who checked "Black or African-American" regardless of what other categories they checked.

Respondents were also asked how often strangers perceived them to have African, European, Native American/Pacific Islander, Asian and Other ancestries (they could respond "Always", "Sometimes", "Never" for each of these). For the analysis, the measure of "any perceived African ancestry" is used which correspond to those who checked either "Always" or "Sometimes" for African ancestry.

3.2.4 Controls

Respondents were asked about their income, education and age at pregnancy. For income, they could respond with any of four categories: "$<$ \$25,000", "\$25-50,000", "\$50-100,000", and "$>$ \$100,000." For education, they could respond "less than high school", "high school degree", "bachelor's degree", "graduate degree."

Summary data for our responses can be seen in Table 1. Our usable responses skew towards White identities (as discussed above, after August 2021 it became much more difficult to reach Black women using targeted Facebook advertising). We are able to use several regression methods to mitigate this issue. Our responses are largely middle income (most report being in the \$25,000-100,000 range) and are

quite well-educated (more than 50% have at least a bachelor's degree). Breakdown of income and education by social categories is provided in the appendices.

Summary Data	
Responses	98
Birthweight	3,344 (692)
African genetic ancestry	23.5 (32.9)
Black self-identification	
Yes	20.4
Yes (with other categories)	14.3
No	65.3
Perceived African ancestry	
Always	31.6
Sometimes	12.2
Never	56.1
Income	
< $25,000	14.3
$25-50,000	22.5
$50-100,000	42.9
> $100,000	20.4
Education	
less than high school	4.1
high school degree	39.8
bachelor's degree	23.5
graduate degree	32.7
Age at pregnancy	30.1 (5.4)

Table 1: Means and percentages for the variables used. Standard deviations for means in parentheses.

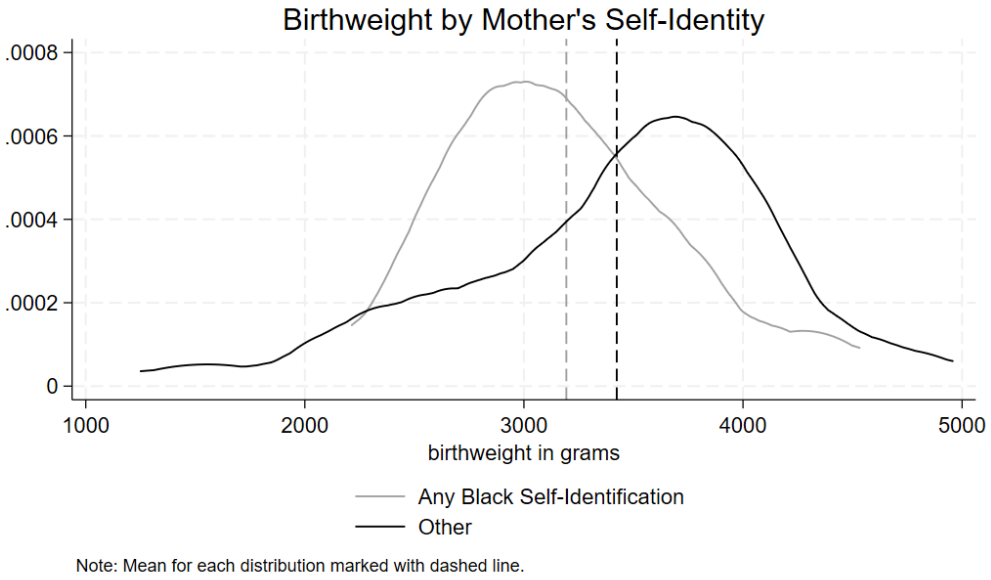


Figure 2: Kernel density graphs of birthweights reported among respondents, for those mothers who report any self-identification as Black, and for those reporting different self-identification.

Both the means and the distributions of birthweights are consistent with national data, although there is of course sampling error given the small sample (see Table 7 in Appendix A for a summary comparison).

In Figure 3 we show actual and predicted Black self-identification as a function of percent of genetic ancestry that is African. Two patterns can be seen that are very helpful for the purposes of testing the relative importance of socially-mediated and non-socially mediated genetic influences on LBW. First, between roughly 5% and 50% African ancestry we find examples of both Black self-identity and no Black self-identity. Second, above 50% African ancestry Black self-identification is universal, giving a range of $A \in [0.5, 1]$ with $I = 1$. Both of these help us to compare the effects of variation in I and B.

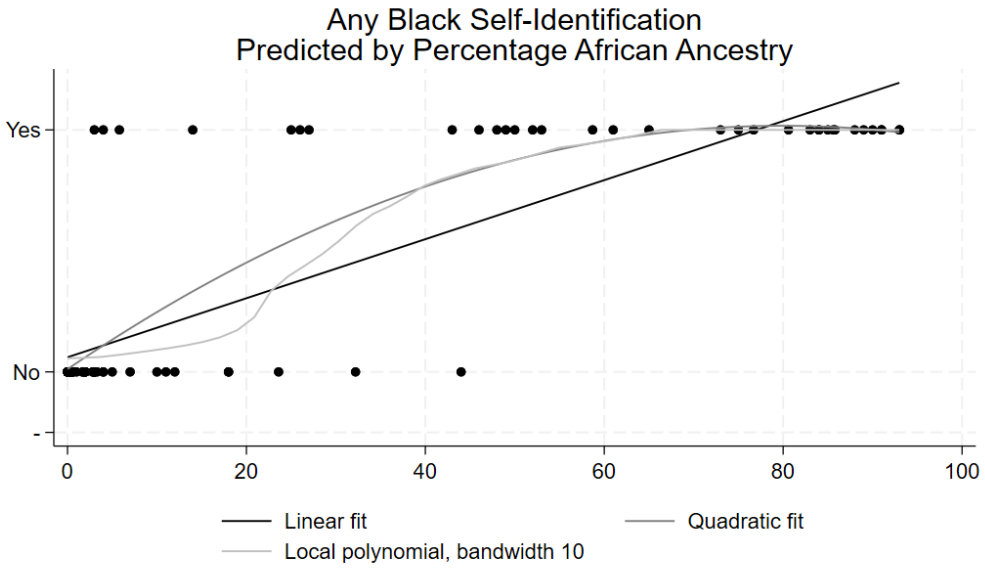


Figure 3: The relationship between Sub-Saharan African ancestry and any Black self-identity is stochastic as ancestry increases from a few percent to roughly 50%.

4 Results

Here we show results from the preferred specification - using all 98 responses with birthweight information, statistically weighting to give equal weight to each decile in the 0%-100% range of African ancestry.

We begin with African ancestry with controls but no information

Birthweight Regressed on Controls, Genetic African Ancestry (GA), Perceived African Ancestry (PA) and Black Self-Identification (SI)					
	GA only	PA only	SI only	GA & PA	GA & SI
Genetic Ancestry	-4.5†			-0.9	-0.5
	(2.7)			(4.3)	(4.5)
Perceived African Ancestry		-425.6*		-376.4	
		(180.5)		(303.8)	
Black Self-Identification			-382.8*		-357.1
			(182.9)		(308.0)
Controls	Education, Income, Age at Pregnancy *Included and not statistically significant*				
R^2	0.183	0.209	0.210	0.210	0.210
Adj. R^2	0.110	0.138	0.139	0.129	0.130
N	98	98	98	98	98

† < 0.1, * p < 0.05

Table 2: Summary results

on social identity

$$y_i = \delta w_i + \beta x_i + \varepsilon_i$$

it has a coefficient of -4.5; that is, going from 0% African ancestry to 100% is associated with a 450 gram reduction in birthweight (it is only barely significant at the 10% level).

Then we use social identity measures (separately) with controls

$$y_i = \alpha z_i + \beta x_i + \varepsilon_i$$

Both perceived African ancestry (respondent answers that others are likely to view them as having African ancestry) and Black self-identity show statistically significant drops in predicted birthweight of 425 grams and 383 gram respectively.

Thus, using the genetic ancestry (GA) and identity (PA, SI) measures separately clearly show the "red hair" effect of racial disparities. We now use them in combination to test the relative strengths of the two plausible pathways.

With genetic ancestry and a social identity measure (PA or SI) is combined, both lose significance.

$$y_i = \delta w_i + \alpha z_i + \beta x_i + \varepsilon_i$$

However, in the case of genetic ancestry it is largely because the coefficient falls close to zero, while for the social measures it is because the standard errors go up substantially even as the point estimate shifts only slightly. This seems to be driven by the collinearity of genetic and social measures - the correlations between them are in the range of 0.85. A larger dataset would obviously improve power; we estimate in Section 5 that 438 observations would probably be sufficient in samples following a similar pattern to our data here.

Tables 3, 4 and 5 present fuller details on regressions with just controls, the three measures (GA, PA, SI) separately, and then the genetic and social measures in co

Birthweight Regressed on Controls and Genetic African Ancestry (GA)						
	Controls Only			Controls and Ancestry (GA)		
		GA > 2.5%	GA reg wgt		GA > 2.5%	GA reg wgt
Percentage African ancestry				-2.9	-5.1	-4.5
				(2.1)	(2.8)	(2.7)
High school	-484.1	450.2	-436.6	-421.5	601.9	-250.3
	(596.4)	(1063.5)	(509.9)	(596.9)	(948.1)	(531.1)
Bachelor's degree	-757.2	216.7	-949.4	-670.3	433.9	-700.4
	(588.5)	(1090.1)	(519.9)	(595.0)	(973.2)	(555.4)
Graduate degree	-403.1	364.7	-585.2	-307.0	664.6	-333.6
	(578.4)	(1096.8)	(525.7)	(579.7)	(986.6)	(542.8)
$25-50,000	54.6	321.8	461.9	46.4	349.0	472.9
	(276.7)	(363.4)	(338.0)	(276.8)	(353.1)	(312.6)
$50-100,000	210.9	232.4	368.4	175.6	198.0	332.7
	(246.1)	(332.4)	(321.7)	(243.8)	(320.7)	(312.7)
More than $100,000	-40.1	-176.7	26.7	-84.9	-257.7	-11.0
	(326.6)	(488.7)	(415.8)	(323.0)	(465.2)	(386.2)
Age at pregnancy	-2.1	5.8	11.3	-1.7	7.7	11.6
	(14.6)	(19.3)	(16.9)	(14.2)	(18.9)	(16.5)
R-squared	0.068	0.099	0.138	0.087	0.157	0.183
Adj R-squared	-0.004	-0.041	0.071	0.005	0.004	0.110
F value	0.877	0.540	1.548	1.229	1.202	1.893
N	98	53	98	98	53	98
Education Wald F	1.44	0.31	3.01	1.33	0.41	2.06
Education Wald p	0.24	0.82	0.03	0.27	0.75	0.11
Income Wald F	0.80	0.81	1.20	0.77	1.16	1.58
Income Wald p	0.50	0.50	0.32	0.51	0.33	0.20
Ed & Inc F	1.02	0.61	1.69	0.92	0.80	1.53
Ed & Inc p	0.42	0.72	0.13	0.48	0.58	0.18

* $p < 0.05$

Table 3: Regression of birthweight of controls and controls with genetic ancestry data.

Birthweight Regressed on Controls, Perceived African Ancestry (PA), and Black Self-Identification (SI)						
	Perceived African Ancestry (PA)			Black Self-Identification (SI)		
		GA > 2.5%	GA reg wgt		GA > 2.5%	GA reg wgt
Others perceive	-226.2	-523.2*	-425.6*			
African ancestry	(161.1)	(231.4)	(180.7)			
Black self identity				-236.1	-436.1*	-382.8*
				(151.2)	(209.5)	(182.9)
High school	-424.9	277.7	-221.7	-469.7	207.0	-300.5
	(575.5)	(1732.7)	(480.2)	(570.2)	(464.4)	(495.1)
Bachelor's degree	-711.7	45.4	-703.6	-715.0	44.7	-738.0
	(571.9)	(1740.7)	(502.7)	(567.7)	(499.7)	(522.1)
Graduate degree	-326.1	317.6	-291.9	-364.8	221.7	-429.9
	(560.6)	(1749.4)	(493.8)	(554.3)	(533.6)	(504.1)
$25-50,000	62.2	421.2	412.4	71.6	462.3	527.4
	(277.7)	(354.5)	(320.9)	(278.2)	(350.9)	(302.3)
$50-100,000	162.9	217.6	278.5	196.6	305.9	392.8
	(241.8)	(317.2)	(313.9)	(240.4)	(313.1)	(302.4)
More than $100,000	-135.4	-376.5	-195.8	-88.2	-244.1	-7.5
	(325.8)	(463.8)	(384.4)	(320.7)	(471.4)	(386.0)
Age at pregnancy	-2.5	5.0	8.0	-3.5	2.9	8.1
	(13.5)	(16.2)	(15.5)	(13.9)	(18.4)	(15.6)
R-squared	0.092	0.207	0.209	0.094	0.182	0.210
Adj R-squared	0.010	0.062	0.138	0.013	0.034	0.139
F value	1.151	1.507	2.620	1.282	1.776	2.747
N	98	53	98	98	53	98
Education Wald F	1.60	0.50	2.65	1.31	0.23	1.92
Education Wald p	0.19	0.68	0.05	0.28	0.88	0.13
Income Wald F	0.88	1.91	1.93	0.93	1.97	2.32
Income Wald p	0.46	0.14	0.13	0.43	0.13	0.08
Ed & Inc F	1.10	1.33	2.17	1.01	1.33	1.90
Ed & Inc p	0.37	0.27	0.05	0.42	0.26	0.09

* $p < 0.05$

Table 4: Regression of birthweight on measures of identity.

Birthweight Regressed on Controls, Genetic African Ancestry (GA) Perceived African Ancestry (PA) and Black Self-Identification (SI)						
	Perceived African Ancestry (PA)			Black Self-Identification (SI)		
		GA > 2.5%	GA reg wgt		GA > 2.5%	GA reg wgt
Percentage African ancestry	-1.1	-0.4	-0.9	0.0	-1.2	-0.5
	(2.9)	(4.2)	(4.3)	(4.1)	(4.0)	(4.5)
Others perceive	-173.1	-503.4	-376.4			
African ancestry	(228.8)	(350.9)	(303.8)			
Black self identity				-236.7	-373.0	-357.1
				(297.2)	(304.3)	(308.0)
High school	-416.1	295.3	-209.3	-469.9	277.3	-288.6
	(589.7)	(774.4)	(498.9)	(584.3)	(532.4)	(506.0)
Bachelor's degree	-691.0	67.8	-682.1	-715.2	119.9	-724.0
	(589.3)	(814.6)	(530.5)	(586.9)	(582.4)	(538.7)
Graduate degree	-309.5	341.2	-275.4	-365.0	311.9	-411.9
	(574.8)	(829.2)	(512.9)	(570.6)	(613.3)	(518.1)
$25-50,000	57.5	419.4	420.3	71.6	448.3	524.2
	(284.2)	(373.3)	(306.9)	(282.6)	(352.8)	(299.7)
$50-100,000	161.4	215.7	281.7	196.7	287.3	387.2
	(244.4)	(333.3)	(307.9)	(244.3)	(311.0)	(296.9)
More than $100,000	-129.2	-374.9	-177.6	-88.1	-253.1	-9.5
	(325.4)	(467.9)	(370.5)	(325.1)	(468.7)	(384.2)
Age at pregnancy	-2.3	5.1	8.4	-3.5	3.7	8.3
	(13.7)	(16.4)	(15.5)	(14.2)	(18.4)	(15.8)
R-squared	0.093	0.207	0.210	0.094	0.184	0.210
Adj R-squared	0.000	0.041	0.129	0.001	0.013	0.130
F value	1.086	1.526	2.312	1.131	1.546	2.444
N	98	53	98	98	53	98
Education Wald F	1.45	0.51	2.41	1.28	0.27	1.81
Education Wald p	0.23	0.67	0.07	0.29	0.84	0.15
Income Wald F	0.83	1.63	1.75	0.92	1.80	2.30
Income Wald p	0.48	0.20	0.16	0.44	0.16	0.08
Ed & Inc F	0.98	1.21	1.88	0.98	1.29	1.78
Ed & Inc p	0.44	0.32	0.09	0.45	0.28	0.11
Ed & Inc p	0.37	0.27	0.05	0.42	0.26	0.09

* $p < 0.05$

Table 5: Regression of birthweight on measures of genetic ancestry and identity.

5 Power Analysis

While the results of the pilot are illuminating, the loss of significance when social and genetic measures are combined is frustrating. However, the data provides a joint distribution of social identity and African ancestry; we are not aware of other sources that document this. As can be seen in Figure 3, Black self-identity increases as African ancestry increases from roughly 5% to 50%.

The effect of Black social identity and perceived African ancestry are between 200 and 350 grams, generally at the higher end of the range. A standard error of 100 grams should thus give sufficient power. A major factor is the strong correlation (collinearity) between social and genetic measures; this increases the standard errors when both social and genetic measures are used.

The power computation[26] in this case is:

$$SE_{B_i} = \frac{sd_Y}{sd_X}\sqrt{\frac{1}{1-R_i^2}}\sqrt{\frac{1-R_Y^2}{n-k-1}}$$

Solving for N:

$$N = \left(\frac{sd_Y}{sd_X}\right)^2 \frac{1}{1-R_i^2}\frac{1-R_Y^2}{SE_{B_i}^2} + k + 1$$

The variables and their values are as follows:

- Standard deviation for US birthweight (grams): $sd_Y = 490$
- Standard deviation for Black identity/perceived African ancestry (a Bernoulli variable, with a mean in this population of 0.35-0.4): $sd_X = 0.45$
- The lowest R^2 regressions are for the unweighted: $R_Y^2 = 0.09$
- Basic regressors are the social and genetic measures, as well as income, education, and age at pregnancy: $k = 5$
- R^2 for regressing identity/perceived ancestry on genetic ancestry and controls varies from roughly 0.45 to 0.75. Here we use the minimum to prepare for a worst case: $R_i^2 = 0.75$

In order to achieve a standard error on identity equal to or less than 100 grams, at least 438 observations are needed. Because this is likely to be sensitive to exact distributions, this should be viewed as a lower bound.

[26] Formula 3.6.1 on page 86 of Cohen et al. (2003).

6 Conclusions

We have reviewed some of the empirical considerations involved in thinking about "red hair" challenges in discrimination, using causal inference and potential outcomes frameworks. We showed that several approaches that have been used by researchers have the potential to create conditional independence of the key independent variables under the right circumstances. We have looked at specific studies focused on infant mortality and morbidity disparities, and reviewed the analysis and results from a pilot study.

While racial discrimination has changed dramatically since Jencks first put forward the "red hair" question, substantial evidence shows that it continues to be important. Measures of explicit and implicit bias, while noisy, have been shown to match comparable measures and to be useful in research (Hehman et al., 2019; Krosnick et al., 2025). A major challenge is mapping from the bias of individuals in a region to the discrimination experienced by individuals - in the context of this paper, particularly Black women who are pregnant or approaching pregnancy, but more generally as well. Linking local measures of bias in a region to measures of stress among Black residents such as cortisol or hair cortisol levels might be one path (Staufenbiel et al., 2013).

We believe the approach of the pilot study has the potential to be developed further and included in the mix of methods used by economists and other researchers to think about social stratification. A rough comparison is with the standard use of polygenic indices, which summarize the genomic effect of a multitude of genes. As with PGIs, this method allows a relatively small sample to generate well-powered results on the role of discrimination and ancestry. Using ancestry data similarly is able to work with smaller samples while maintaining power. And by integrating ancestry data with social identity allows us to isolate the effect of identity in stratification. Methods from genomics such as FGWAS are obviously superior for research focused on the role of *genes*. But while they control for environment in a number ways the sheer number of SNPs they investigate mean that quite large samples are required for sufficient power. Using summary ancestry data as in the pilot study above thus represents a potential complementary tool.

Overall, the "red hair" problem is at the epicenter of some of the most important and analytically challenging issues in the social sciences. Further building out the array of tools available to researchers, and allowing different aspects of the question to be targeted is a vital

concern. We believe this study helps to further this project.

References

Ancestry composition: 23andme's state-of-the-art geographic ancestry analysis. /urlhttps://www.23andme.com/ancestry-composition-guide/, 2022. Accessed: 7/15/2024.

George A Akerlof and Rachel E Kranton. Economics and identity. *The quarterly journal of economics*, 115(3):715–753, 2000.

Douglas Almond. Is the 1918 influenza pandemic over? long-term effects of in utero influenza exposure in the post-1940 us population. *Journal of political Economy*, 114(4):672–712, 2006.

Douglas Almond and Janet Currie. Killing me softly: The fetal origins hypothesis. *Journal of economic perspectives*, 25(3):153–172, 2011.

Douglas Almond, Kenneth Y. Chay, and David S. Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3): 1031–1083, 2005.

Joshua D Angrist and Guido W Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, 90(430): 431–442, 1995.

Emmanuel A. Anum, Edward H. Springel, Mark D. Shriver, and Jerome F. Strauss. Genetic contributions to disparities in preterm birth. *Pediatric research*, 65(1):1, 2009.

Susan Athey and Guido W Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic perspectives*, 31(2):3–32, 2017.

John Barnard, Constantine E Frangakis, Jennifer L Hill, and Donald B Rubin. Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in new york city. *Journal of the American Statistical Association*, 98(462):299–323, 2003.

Gary S Becker. *The economics of discrimination*. University of Chicago press, 1957.

Jess Benhabib, Alberto Bisin, and Matthew O Jackson. *Handbook of social economics*. Elsevier, 2010.

Daniel J Benjamin, David Cesarini, Christopher F Chabris, Edward L Glaeser, David I Laibson, Gene/Environment Susceptibility-Reykjavik Study: Age, Vilmundur Guðnason, Tamara B Harris, Lenore J Launer, Shaun Purcell, et al. The promises and pitfalls of genoeconomics. *Annu. Rev. Econ.*, 4(1):627–662, 2012.

Daniel J Benjamin, David Cesarini, Patrick Turley, and Alexander Strudwick Young. Social-science genomics: Progress, challenges, and future directions. 2024.

Marianne Bertrand and Esther Duflo. Field experiments on discrimination. *Handbook of economic field experiments*, 1:309–393, 2017.

Nazleen Bharmal, Kathryn Pitkin Derose, Melissa Felician, Margaret M Weden, et al. Understanding the upstream social determinants of health. *California: RAND*, pages 1–18, 2015.

Howard Bodenhorn and Christopher Ruebeck. The economics of identity and the endogeneity of race, 2003.

Paula Braveman. *The social determinants of health and health disparities*. Oxford University Press, 2023.

Paula Braveman, Tyan Parker Dominguez, Wylie Burke, Siobhan M Dolan, David K Stevenson, Fleda Mask Jackson, James W Collins, Deborah A Driscoll, Terinney Haley, Julia Acker, et al. Explaining the black-white disparity in preterm birth: a consensus statement from a multi-disciplinary scientific work group convened by the march of dimes. *Frontiers in Reproductive Health*, 3:684207, 2021.

Michela Carlana. Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, 134(3):1163–1224, 2019.

David H. Chae, Sean Clouston, Connor D. Martz, Mark L. Hatzenbuehler, Hannah LF Cooper, Rodman Turpin, Seth Stephens-Davidowitz, and Michael R. Kramer. Area racism and birth outcomes among blacks in the united states. *Social science and medicine*, 199:49–55, 2018.

Kerwin Kofi Charles and Jonathan Guryan. Prejudice and wages: An empirical assessment of becker's the economics of discrimination. 2008.

Grieve Chelwa, Darrick Hamilton, and James Stewart. Stratification economics: Core constructs and policy implications. *Journal of Economic Literature*, 60(2):377–399, 2022.

Raj Chetty, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. Race and economic opportunity in the united states: An intergenerational perspective. *The Quarterly Journal of Economics*, 135(2): 711–783, 2020.

Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. *Applied multiple regression/correlation analysis for the behavioral sciences*. L. Erlbaum Associates, Mahwah, NJ, 3. edition, 2003.

William J Collins and Melissa A Thomasson. The declining contribution of socioeconomic disparities to the racial gap in infant mortality rates, 1920-1970. *Southern Economic Journal*, 70(4):746–776, 2004.

James W Collins, Jr., Shou-Yien Wu, and Richard J David. Differing intergenerational birth weights among the descendants of us-born and foreign-born whites and african americans in illinois. *American journal of epidemiology*, 155(3):210–216, 2002.

James W Collins Jr, Richard J David, Arden Handler, Stephen Wall, and Steven Andes. Very low birthweight in african american infants: the role of maternal exposure to interpersonal racial discrimination. *American journal of public health*, 94(12):2132–2138, 2004.

Eve R Colson, Denis Rybin, Lauren A Smith, Theodore Colton, George Lister, and Michael J Corwin. Trends and factors associated with infant sleeping position: the national infant sleep position study, 1993-2007. *Archives of pediatrics & adolescent medicine*, 163(12): 1122–1128, 2009.

Hope Corman, Theodore J Joyce, and Michael Grossman. Birth outcome production function in the united states. *Journal of Human Resources*, pages 339–360, 1987.

Hope Corman, Dhaval Dave, and Nancy E Reichman. Evolution of the infant health production function. *Southern Economic Journal*, 85(1):6–47, 2018.

Ricardo Dahis, Emily Nix, and Nancy Qian. Choosing racial identity in the united states, 1880-1940. Technical report, National Bureau of Economic Research, 2019.

Richard David and James Collins, Jr. Disparities in infant mortality: what's genetics got to do with it? *American Journal of Public Health*, 97(7):1191–1197, 2007.

Richard David and James Collins Jr. Disparities in infant mortality: what's genetics got to do with it? *American journal of public health*, 97(7):1191–1197, 2007.

Todd E Elder, John H Goddeeris, and Steven J Haider. Racial and ethnic infant mortality gaps and the role of socio-economic status. *Labour economics*, 43:42–54, 2016.

Ronald A Fisher. The design of experiments. 1949.

Heather A. Frey, Molly J. Stout, Laurel N. Pearson, Methodius G. Tuuli, Alison G. Cahill, Jerome F. Strauss III, Luis M. Gomez, Samuel Parry, Jenifer E. Allsworth, and George A. Macones. Genetic variation associated with preterm birth in african-american women. *American Journal of Obstetrics and Gynecology*, 215(2): 235. e1–235. e8, 2016.

Arline T Geronimus. The weathering hypothesis and the health of african-american women and infants: evidence and speculations. *Ethnicity & disease*, pages 207–221, 1992.

Roy Gravel. Tay-sachs disease. *Genetics*, 4:98–102, 2003.

Junming Guan, Tammy Tan, Seyed Moeen Nehzati, Michael Bennett, Patrick Turley, Daniel J Benjamin, and Alexander Strudwick Young. Family-based genome-wide association study designs for increased power and robustness. *Nature Genetics*, 57(4):1044–1052, 2025.

Eric Hehman, Jimmy Calanchini, Jessica K Flake, and Jordan B Leitner. Establishing construct validity evidence for regional measures of explicit and implicit racial bias. *Journal of experimental psychology: General*, 148(6):1022, 2019.

K Hirano and GW Imbens. The propensity score with continuous treatments. applied bayesian modeling and causal inference from incomplete-data perspectives, 2004.

Ryan Huang, Andrea R Spence, and Haim A Abenhaim. National sids trends in the united states from 2000 to 2019: A population-based study on 80 million live births. *Clinical Pediatrics*, 63(9):1216–1224, 2024.

Guido W Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–1179, 2020.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Christopher Jencks. *Rethinking Social Policy: Race, Poverty and the Underclass*. Harper Collins, 1992.

Christopher Jencks and Meredith Phillips. *The Black-White test score gap*. Brookings Institution Press, 2011.

Christopher Jencks, Marshall Smith, Henry Acland, Mary Jo Bane, David Cohen, Herbert Gintis, Barbara Heyns, and Stephan Michelson. *Inequality: A Reassessment of the Effect of Family and Schooling in America*. Basic Books, 1972.

Jon A Krosnick, Tobias H Stark, and Amanda L Scott. *The Cambridge handbook of implicit bias and racism*. Cambridge University Press, 2025.

Kevin Lang and Ariella Kahn-Lang Spitzer. Race discrimination: An economic perspective. *Journal of Economic Perspectives*, 34(2):68–89, 2020.

Aparna Lhila and Sharon Long. What is driving the black–white difference in low birthweight in the us? *Health Economics*, 21(3): 301–315, 2012.

John C. Loehlin, Steven G. Vandenberg, and R. Travis Osborne. Blood group genes and negro-white ability differences. *Behavior genetics*, 3(3):263–270, 1973.

Michael J Lopez and Roee Gutman. Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, pages 432–454, 2017.

Charles F Manski. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542, 1993.

Charles F Manski. Economic analysis of social interactions. *Journal of economic perspectives*, 14(3):115–136, 2000.

Michelle N Meyer, Paul S Appelbaum, Daniel J Benjamin, Shawneequa L Callier, Nathaniel Comfort, Dalton Conley, Jeremy Freese, Nanibaa'A Garrison, Evelynn M Hammonds, K Paige Harden, et al. Wrestling with social and behavioral genomics: risks, potential benefits, and ethical responsibility. *Hastings Center Report*, 53:S2–S49, 2023.

Keith Noto, Yong Wang, Shiya Song, David Turissini, Alisa Sedghifar, Barry Starr, Jake Byrnes, Eurie Hong, Catherine Ball, and Kenneth Chahine. Ethnicity estimate white paper. `https://www.ancestry.com/cs/dna-help/ethnicity/whitepaper`, 2013. Accessed: 7/15/2024.

Jacob Orchard and Joseph Price. County-level racial prejudice and the black-white gap in infant health outcomes. *Social science and medicine*, 181:191–198, 2017.

Philip Oreopoulos, Mark Stabile, Randy Walld, and Leslie L. Roos. Short-, medium-, and long-term consequences of poor infant health an analysis using siblings and twins. *Journal of human Resources*, 43(1):88–138, 2008.

MJK Osterman, BE Hamilton, JA Martin, AK Driscoll, and CP Valenzuela. Births: Final data for 2022, 2024.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Lincoln Quillian. New approaches to understanding racial prejudice and discrimination. *Annu. Rev. Sociol.*, 32(1):299–328, 2006.

David Reich. *Who we are and how we got here: Ancient DNA and the new science of the human past*. Pantheon Press, 2018.

Nancy E Reichman, Hope Corman, Kelly Noonan, and Dhaval Dave. Infant health production functions: What a difference the data make. *Health Economics*, 18(7):761–782, 2009.

Shelley Reuter. Tay-sachs disease. *Encyclopedia of Race and Racism*, 4], year=2013, publisher= Macmillan Reference USA, pages=161–165, other=. Gale eBooks, link.gale.com/apps/doc/CX4190600434/GVRL?u=transylvaniausid=bookmark-GVRLxid=024fd767. Accessed 7 Jan. 2026.

Mark R Rosenzweig and T Paul Schultz. Estimating a household production function: Heterogeneity, the demand for health inputs, and their effects on birth weight. *Journal of political economy*, 91(5): 723–746, 1983.

Aliya Saperstein and Andrew M Penner. Racial fluidity and inequality in the united states. *American journal of sociology*, 118(3):676–727, 2012.

Sandra Scarr, Andrew J. Pakstis, Solomon H. Katz, and William B. Barker. Absence of a relationship between degree of white ancestry and intellectual skills within a black population. *Human genetics*, 39(1):69–86, 1977.

Eui Hang Shin. Black-white differentials in infant mortality in the south, 1940–1970. *Demography*, 12(1):1–19, 1975.

Gopal K Singh and M Yu Stella. Infant mortality in the united states, 1915-2017: large social inequalities have persisted for over a century. *International Journal of Maternal and Child Health and AIDS*, 8(1): 19, 2019.

Joedrecka S. Brown Speights, Samantha Sittig Goldfarb, Brittny A. Wells, Leslie Beitsch, Robert S. Levine, and George Rust. State-level progress in reducing the black–white infant mortality gap, united states, 1999–2013. *American Journal of Public Health*, 107(5):775–782, 2017.

Sabine M Staufenbiel, Brenda WJH Penninx, Anne T Spijker, Bernet M Elzinga, and Elisabeth FC van Rossum. Hair cortisol, stress exposure, and mental health in humans: a systematic review. *Psychoneuroendocrinology*, 38(8):1220–1235, 2013.

Seth Stephens-Davidowitz. The cost of racial animus on a black candidate: Evidence using google search data. *Journal of Public Economics*, 118:26–40, 2014.

Carl Veller and Graham M Coop. Interpreting population-and family-based genome-wide association studies in the presence of confounding. *PLoS biology*, 22(4):e3002511, 2024.

Siqi Wang, Jaime Berumen, Alejandra Vergara-Lope, Paulina Baca, Elizabeth Barrera, Fernando Rivas, Diego Aguilar-Ramirez, Rory Collins, Jonathan R Emberson, Michael Hill, et al. Direct effect of genetic ancestry on complex traits in a mexican population. *medRxiv*, pages 2025–09, 2025.

Alexander I Young, Seyed Moeen Nehzati, Stefania Benonisdottir, Aysu Okbay, Hariharan Jayashankar, Chanwook Lee, David Cesarini, Daniel J Benjamin, Patrick Turley, and Augustine Kong. Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nature genetics*, 54(6):897–905, 2022.

A Additional Study Background

Summary of responses.

Summary of Responses	
Stage	Count
Initial responses	153
Dropped responses:	
DNA analysis not reported	15
Excluded based on cutoffs	12
No pregnancy reported	6
Multiples	6
No data on length of pregnancy	1
Main dataset	113
No data on weight	14
No data on age at pregnancy	1
Birthweight dataset	98

Table 6: Summary of responses gathered May-June 2020 and May-September 2021.

Summary of Responses				
	Study		US 2022	
	Black	White	Black	White
Less than 2,000g	0	4.69	5.74	2.42
2,000-2,499	5.88	10.94	9.01	4.70
2,500-2,999	38.24	9.38	26.40	16.84
3,000-3,499	32.35	21.88	37.23	38.38
3,500-3,999	14.71	37.50	17.57	28.91
4,000g and greater	8.82	15.63	4.05	8.74

Table 7: Comparison of study birthweights with US data from the National Center for Health Statistics for 2022 Osterman et al. (2024).

B Additional Detail on Self-Identification

For the analysis in the main paper we use "Any self-identification as Black" as the key variable for self-identification. In the questionnaire, respondents were encouraged to note any of the standard U.S. Census racial categories ("White", "Black or African-American", "American Indian or Alaska Native", "Asian" or "Other"). Any self-identification then corresponded to those who checked "Black or African-American" regardless of what other categories they checked. We have a fuller measure of self-identification which notes if they only checked "Black", if they checked "Black" and another category, or if they did not check "Black." The following data summarizes some of the patterns by this measure.

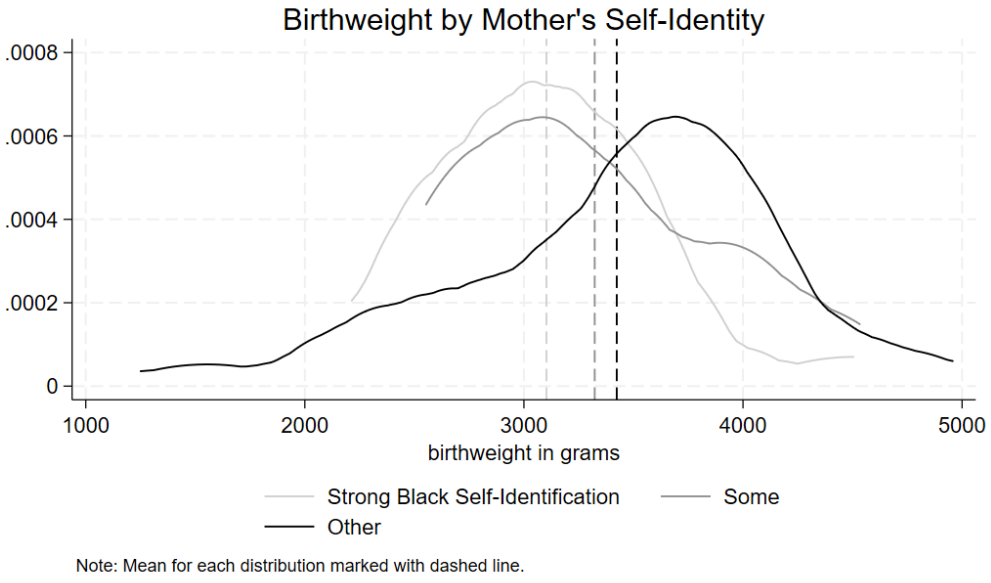


Figure 4: The above graph summarizes the distribution of birthweights for the three categories of Black self-identification.

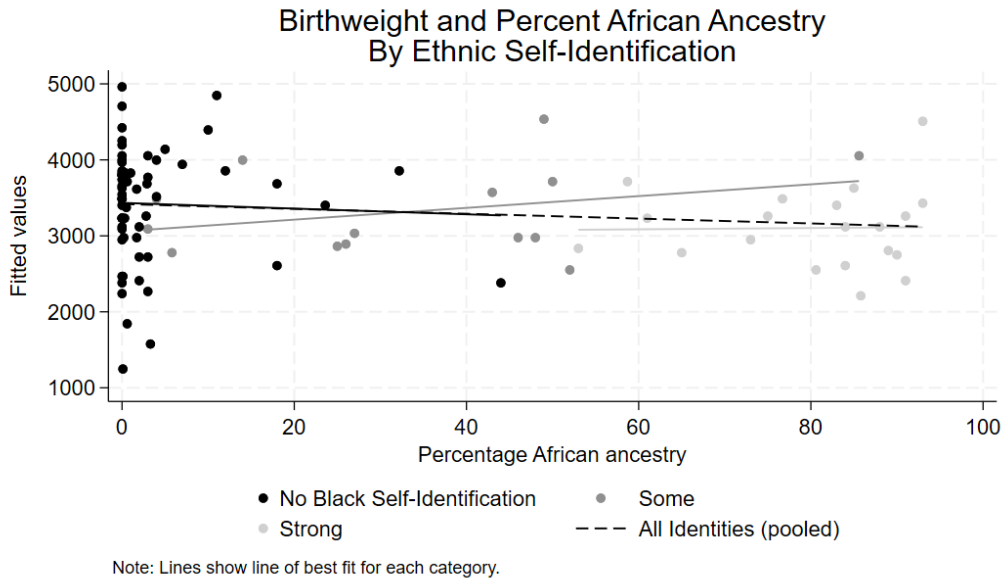


Figure 5: The above graph summarizes the relationship between percentage African ancestry and birthweight for the three categories of Black self-identification. The dashed line is the line of best fit for all respondents (pooled).

Black Self-Identification: Breakdown of Income and Education Covariates					
			Income		
	$< \$25{,}000$	\$25-50,000	\$50-100,000	$> \$100{,}000$	Total
Black self-identity..					
strong	3	5	8	4	20
some	2	5	7	0	14
none	9	12	27	16	64
Total	14	22	42	20	98
N	98				

			Education		
	< high school	High school	Bachelor's degree	Graduate degree	Total
Black self-identity..					
strong	0	7	6	7	20
some	1	6	3	4	14
none	3	26	14	21	64
Total	4	39	23	32	98
N	98				

Table 8: Summary of income and education by racial self-identity.

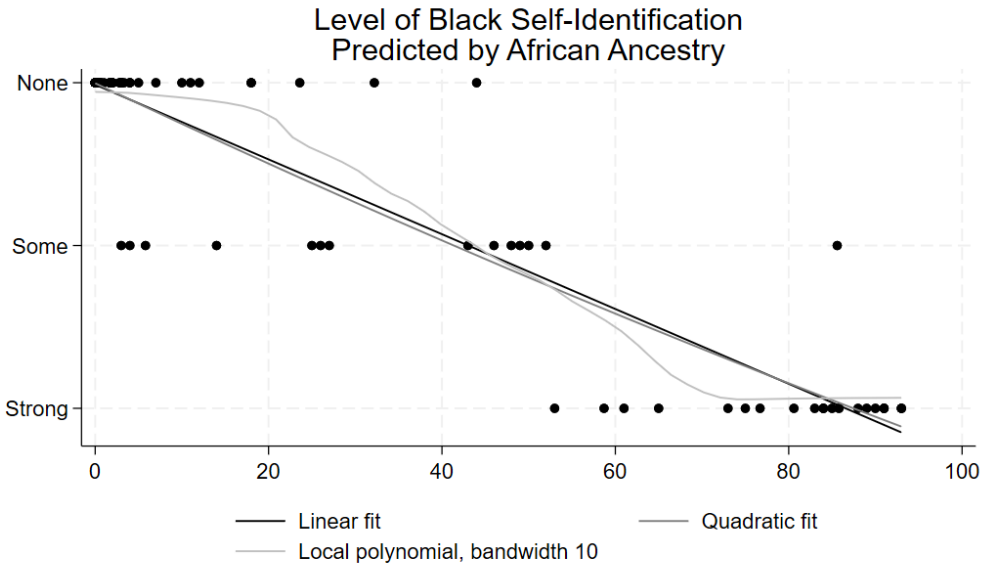


Figure 6: The above graph uses percentage of African ancestry to predict where respondents align in the three categories of Black self-identification.

C Additional Detail on Perceived Ancestry

For the analysis in the main paper we use "Any perceived African ancestry" as the key variable for perceived ancestry. In the questionnaire, respondents were asked to note how often strangers perceived them to have African, European, Native American/Pacific Islander, Asian and Other ancestries (they could respond "Always", "Sometimes", "Never" for any combination of these). Any perceived African ancestry then corresponded to those who checked either "Always" or "Sometimes" for African ancestry. We have a fuller measure of perceived African ancestry which notes if for African ancestry (a) they checked "Always", (b) "Sometimes," or (c) "Never." The following data summarizes some of the patterns by this measure.

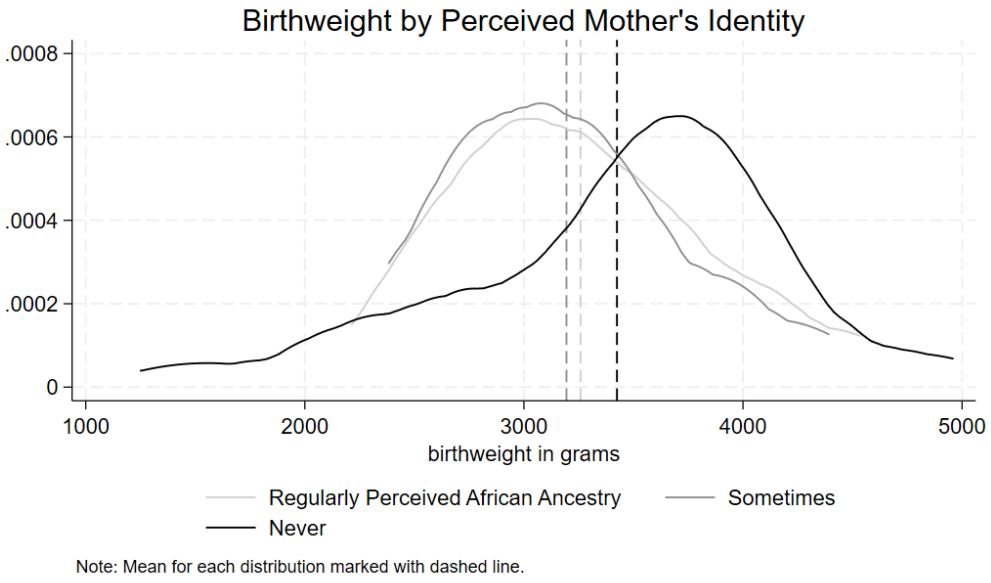


Figure 7: The above graph summarizes the distribution of birthweights for the three categories of perceived African ancestry.

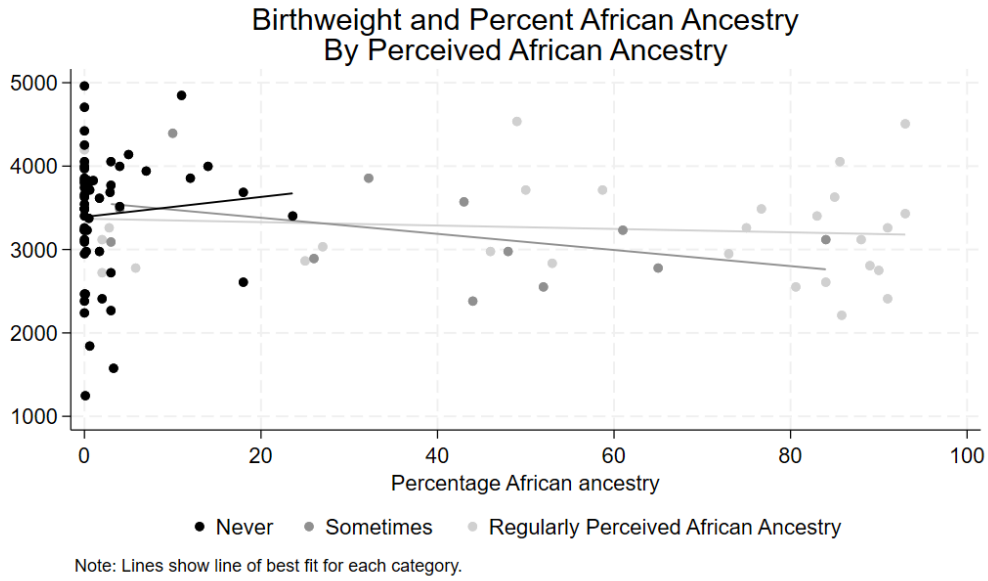


Figure 8: The above graph summarizes the relationship between percentage African ancestry and birthweight for the three categories of perceived African ancestry.

Perceived African Ancestry: Breakdown of Income and Education Covariates					
			Income		
	$< \$25{,}000$	\$25-50,000	\$50-100,000	$> \$100{,}000$	Total
African ancestry..					
always perceived	6	9	12	4	31
sometimes perceived	2	5	5	0	12
never perceived	6	8	25	16	55
Total	14	22	42	20	98
N	98				

			Education		
	$<$ high school	High school	Bachelor's degree	Graduate degree	Total
African ancestry..					
always perceived	1	15	4	11	31
sometimes perceived	0	5	4	3	12
never perceived	3	19	15	18	55
Total	4	39	23	32	98
N	98				

Table 9: Summary of income and education by perceived ancestry.

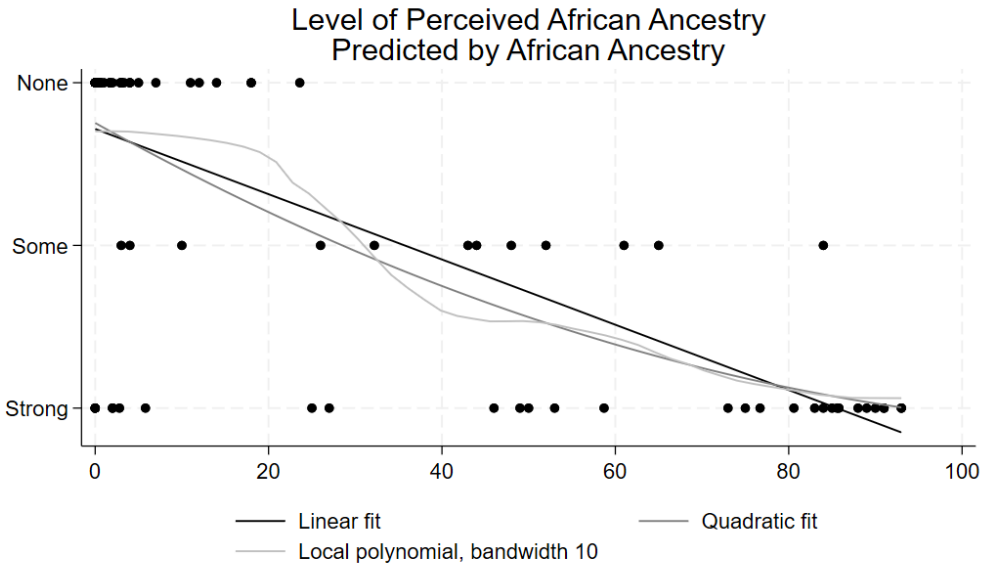


Figure 9: The above graph uses percentage of African ancestry to predict where respondents align in the three categories of perceived African ancestry.