sociological science

A Roadmap for Inequality Research: Transparency, Intersectionality, and Multiple Measures of Race

Emma Williams-Baron, Aliya Saperstein

Stanford University

Abstract: Most quantitative studies of U.S. inequality rely on single measures of race and do not transparently describe them. However, inconsistencies between measures can yield conclusions that differ both substantively and statistically. We ask: when faced with multiple ways to categorize respondents, how should researchers choose? We conduct intersectional analyses of five inequality outcomes, using the 1979 National Longitudinal Survey of Youth, which offers several measures of self-identification and external classification. Strikingly, we find the survey's screener race variable, ubiquitous in prior research, is never empirically preferred based on model fit across outcomes spanning the labor market (wages, salary, and unemployment), health (depression), and education (school discipline). Instead, the top-performing measure varies by gender, outcome, and fit statistic. The range of potential researcher decisions and the absence of a clear gold-standard highlights the need for greater transparency and more thoughtful decision-making when researchers operationalize race—whether racial categorization is central to the analysis or included primarily as a control variable. To that end, we offer a roadmap of key considerations inequality researchers can consult when designing their approach.

Keywords: racial classification; data disaggregation; survey measurement; transparency; reproducibility; intersectionality

Reproducibility Package: Data and code for reproducing the results presented in this article are publicly available in an Open Science Framework repository here: https://doi.org/10.17605/OSF.IO/K3RZT. Data may also be accessed through the NLSY Investigator site at: https://www.nlsinfo.org/investigator.

Citation: Williams-Baron, Emma, Aliya Saperstein. 2026. "A Roadmap for Inequality Research: Transparency, Intersectionality, and Multiple Measures of Race" Sociological Science 13: 825-863.

Received: September 30, 2025

Accepted: May 18, 2026

Published: July 9, 2026

Editor(s): Arnout van de Rijt, Kristian B. Karlson

DOI: 10.15195/v13.a32

In recent years, quantitative scholars have joined qualitative scholars in critiquing the status quo of demographic and survey research as being reductive and harmful in its understanding of inequality in general, and racial inequality, in particular. These concerns include inordinate attention to "group gap" and "deficit" perspectives (e.g., Leicht 2008; Pattillo 2021) and a lack of attention to individual intersectional experiences as well as structure and context (e.g., Brown and Hargrove 2013; Homan, Brown, and King 2021). Scholars across disciplines are also increasingly calling for researchers to interpret results that indicate "race differences" as originating not from individual characteristics but rather as speaking to relational processes of racism and discrimination (e.g., Cross, Fomby, and Letiecq 2022; Zuberi 2001).

Among these growing critiques of standard statistical practice is a concern about researchers' overreliance on a single, static measure to represent the complexity of how individuals identify themselves and are perceived by others in racial terms (Wilson, Beachy, and Schumm 2025). Different measures of race produce both different population counts and different conclusions about racial inequality (e.g.,

López and Hogan 2021; Saperstein 2006; Telles and Lim 1998; Vargas and Kingsbury 2016). Measures of race for the same individual can also vary over time, across contexts, and by social status (e.g., Liebler et al. 2017; Roth 2016; Sosina and Saperstein 2022; Villarreal and Bailey 2020). These results imply that inequality researchers must be particularly careful not only in interpreting their findings but also in choosing which measure, or measures, of race to use in the first place.

Yet consideration and use of multiple measures of race and/or ethnicity in research remains rare, especially outside of specialty journals, such as *Sociology of Race and Ethnicity*.[1] These omissions may result, in part, from inequality researchers either not knowing that multiple measures are available in major surveys, not being exposed to accessible examples of how to choose between these multiple options, or both. We aim to close this knowledge gap in several ways: first, we identify and demonstrate problems with the status quo of how race/ethnicity measures are currently selected and operationalized; second, we outline a set of practices that (a) makes the numerous operationalization decision points explicit and (b) encourages researchers to transparently justify and explain their decisions; overall, we seek to advance conversation within the discipline about how to put existing best-practice guidelines into more widespread use among all inequality scholars, including novice and expert researchers alike.

To demonstrate limitations of the status quo and illustrate how different measures of race generate different conclusions, we extend previous research using the 1979 National Longitudinal Survey of Youth (Light and Nandi 2007), by considering the amount of variation explained in each of five outcomes across three domains of inequality, separately for women and men. We consider four different measures of race, each operationalized using a range of approaches, yielding 29 different specifications. We focus on the NLSY79 in part because of its influence on social science research: it is one of the longest-running nationally representative longitudinal surveys in the United States, and it has been cited in over 6,700 research articles (U.S. Bureau of Labor Statistics n.d.-a). However, the broader approach we outline is relevant to other datasets that offer multiple measures of race, including the NLSY 1997 cohort, the National Longitudinal Study of Adolescent to Adult Health (Add Health), the Portraits of American Life Study, and the New York City Longitudinal Survey of Wellbeing. The differences in results we demonstrate here may also inspire additional data collection using multiple measures of race/ethnicity in the future.

In line with research on other surveys (e.g., Howell and Emerson 2017; Shiao 2019), we find there is no single best-performing measure or operationalization of race across all of our inequality outcomes: the measures and classification schemes that optimize the balance between model fit and parsimony differ by gender and the fit statistic being used. There is, however, a clearly poor performing measure: the NLSY79's screener race variable, which is frequently used in prior literature, but is never the best-performing option in any of our comparisons. Contrary to common practice, we also find operationalization approaches that aggregate all multiracial responses into a "two or more" category or that aggregate all Hispanic responses into a single category (often overriding race responses) are rarely the best fitting across our comparisons.

Our results—both the lack of a clear "gold-standard" approach and the weaknesses of many existing practices—underscore the importance of transparency in how researchers select, operationalize, and interpret their results in studies of inequality. Calls for more transparency and reproducibility in social science research have increased (Damian, Meuleman, and van Oorschot 2022; Freese 2007; Moody, Keister, and Ramos 2022; Steegen et al. 2016) alongside calls for more critical approaches to measuring race/ethnicity. We unite these two conversations by identifying the many sources of "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn 2011) when operationalizing race/ethnicity and demonstrating how those decisions can have consequences for inequality research.

In highlighting these challenges, our findings also point to important opportunities for future inequality research. When we incorporate multiple measures of race simultaneously, our results emphasize the importance of including observed race, which suggests that how people are perceived by others plays a key role in perpetuating racial inequality in the United States. Our findings also support calls for more intersectional approaches in quantitative inequality research (e.g., McCall 2005; Scott and Siltanen 2017). Although not all studies will be able to incorporate each of these insights, we argue they are key considerations to weigh during both study design and when interpreting results. Indeed, even if the differences in results generated by different measures or operationalization approaches were small, the value of establishing a strong theoretical justification for selecting measures, and the value of a robustly transparent description of them that promotes replication, offer important advances over the status quo on their own. However, in many cases, the empirical differences we find are quite large, as measured not only by improvements in model fit statistics but also by changes in substantive interpretations. Altogether, our results point to the need for improving how researchers justify their use of racial/ethnic measures, increasing transparency and generalizability, smoothing replication, and ultimately strengthening our understanding of inequality producing processes.

Measuring Race: Transparency, Theory, and Empirical Rigor

When reviewing papers that drew on NLSY79 data to study racial inequality, we often found it difficult to determine which of the four available race/ethnicity measures researchers used in their analyses. The NLSY79 is one of only a few major U.S. surveys that includes multiple measures of race both at the same point in time and across time: one calculated variable compiled from information gathered during a 1978 screening interview, two self-identified race/ethnicity measures from 1979 and 2002, and the interviewer's classification of the respondent across 17 years of the survey until 1998. The NLSY79 User's Guide recommends the initial screener variable, in part because it was used to calculate the survey weights (U.S. Bureau of Labor Statistics n.d.-b), and it is the only race/ethnicity variable included by default in all data extracts. We presume many studies that draw on the NLSY79 and incorporate race/ethnicity in their analyses follow suit and use this variable, though

we could often only guess this information based on how many racial categories the authors presented in their results. (As we describe in more detail hereafter, the screener has just three categories, while the self-identification measures have 29 and six response options, respectively.) Even then, we found the three screener race categories were often misleadingly labeled, with the "non-Black, non-Hispanic" residual category renamed, simply, as "White." The overall lack of transparency was true regardless of whether racial disparities were the explicit focus of analysis or whether race/ethnicity was treated simply as a control or confounding factor in an analysis focused on something else.[2]

Practices such as omitting or eliding how race/ethnicity data were collected and coded have been highlighted as a problem for several decades by academic journals in genetic and biomedical fields (e.g., Kaplan and Bennett 2003; Mauro et al. 2022). Many journals and professional associations in those disciplines have adopted explicit guidelines for the reporting of race/ethnicity data that stress the need for more precise language and transparency regarding data provenance, though the presence of such guidelines on its own has not consistently translated into improved published research (see e.g., Flanagin et al. 2023; Martinez et al. 2023). A recent consensus report on the use of race/ethnicity in biomedical research published by the National Academies of Sciences, Engineering, and Medicine further urges both journal editors and funders to not only provide explicit guidelines for researchers but also to generate more accountability to ensure compliance (Wilson et al. 2025). Interestingly, even with the recent push for transparency and reproducibility in the social sciences, similar efforts to codify standards for race/ethnicity reporting have not been forthcoming.

To help fill this gap and provide best-practice guidance for social scientists studying racial inequality, some quantitative researchers are beginning to unite under the banner of "QuantCrit" (see e.g., Castillo and Gilborn 2023; Garcia, López, and Vélez 2018). For scholars in this emerging tradition, best practice recommendations go beyond basic calls for precision or transparency to encompass many of the existing critiques of quantitative research on inequality noted previously. Our analyses build on two QuantCrit principles, in particular: 1) race is not an inherent individual characteristic but a proxy for the social experience of racialization and racism, and 2) racial categories are "neither 'natural' nor given" (Gillborn, Warmington, and Demack 2018:158). Following the latter principle, we carefully attend to decisions about category aggregation and specificity, aiming to balance potential concerns about using either too many categories in analysis (and losing statistical power) or too few (and obscuring meaningful heterogeneity).

A related line of scholarship extends from the first principle and highlights the importance of conceptualizing race as multidimensional, with different aspects or measures of race providing different results, in part because they are connected to different inequality-producing processes (see, e.g., Roth 2016; Saperstein 2012). For example, measures of ancestry or descent point to a different discriminatory mechanism than measures of skin tone or how someone chooses to identify themselves. Multiple measures of race can also provide insight into the predictors and consequences of racial identity contestation, or when one's self-identification is not validated by others (e.g., Vargas and Kingsbury 2016). This body of work highlights

concerns about content or overall construct validity, stressing the need for carefully theorizing which measure is most relevant for a given research question and, whenever possible, leveraging multiple measures in the same analysis to better explore empirical complexity or test hypotheses about the role of race in (re)producing inequality (e.g., Bailey, Saperstein, and Penner 2014; Monk 2016; Saperstein, Kizer, and Penner 2016; Stepanikova and Oates 2016).

Scholars also have begun to consider how to choose the most appropriate classification scheme for a given measure of race. These choices can be motivated by theoretical questions, such as the nature of the racial hierarchy in the United States (e.g., Guluma and Saperstein 2022), or empirical ones, such as how best to incorporate Hispanic respondents, multiracial respondents, people who select "other" and write-in responses, or people who change their self-identification over time (e.g., Howell and Emerson 2017; Shiao 2019, 2023; Wong et al. 2024). Collectively, this research highlights criterion evidence for validity (using both concurrent and predictive approaches) and suggests not only that it "matters how you measure" but also that the best classification scheme can differ by the outcome of interest. In combination with the research on race as multidimensional noted earlier, this growing body of work implies there is not a single gold-standard measure or operationalization of race that researchers can rely on for all purposes.

The recent emphasis on careful measure selection in race scholarship echoes more general calls for attention to researcher decision-making in quantitative social science research. Multiverse analysis, springing out of the open science movement towards transparency and reproducibility, recognizes the "garden of forking paths" that confronts researchers with many plausible choices in data processing and analysis, including the operationalization of variables and specification of models (Steegen et al. 2016).[3] Together, these sets of choices produce a theoretical "multiverse" of data sets, analyses, and outcomes. Given that many alternative specifications may all be reasonable, choosing just one can be arbitrary. To evaluate robustness, multiverse analysis recommends constructing multiple versions of the dataset under study by implementing all versions of reasonable and defensible data processing choices (e.g., Muñoz and Young 2018b). We share the goal of bringing transparency to the many "researcher degrees of freedom," in our case those underlying race/ethnicity measures. We also aim to connect the concerns of many race scholars with those of various proponents of multiverse analysis (e.g., Engzell and Mood 2023), revealing both existing blind spots and fruitful areas of future research.

Only one other published paper explicitly considers the relative performance of different measures of race in the NLSY79. We follow Light and Nandi (2007) and, like them, we examine racial wage disparities. However, Light and Nandi (2007) limited their comparisons of different classification schemes to the screener race variable and the two measures of self-identification. We extend their research by including the interviewer's racial classification of the respondent, testing combinations among the measures, and considering model performance based on penalized model fit statistics, which account for parsimony. We also examine other outcomes beyond wages and add an intersectional lens (see Collins 2015), splitting the sample by binary gender to examine whether the best measures differ between women and men. These additions prove to be important for our conclusions.

Data and Methods

The National Longitudinal Survey of Youth 1979 (NLSY79) was fielded annually from 1979 to 1996, then every other year through the present, yielding 29 waves of data to date. The NLSY79 covers a wide range of topics, including employment and assets, education, health, family background, attitudes, and more, making it broadly relevant to studies across the social sciences. The sample comprises 12,686 respondents aged 14–22 at the first interview and aged 41–50 in 2006, which was the last year of data in our analyses. Sample sizes for our analyses, which we conduct separately by gender, range from 3,554 to 6,403 depending on the outcome of interest.[4]

Outcome Variables: Three Domains of Inequality

We analyze how well each racial/ethnic measure performs on five outcomes that span three domains of inequality: the labor market (wages, salary, and unemployment), health (depression), and education (school discipline). We chose these outcomes to give both breadth (three different inequality domains) and depth (three labor market measures) of insight into racial inequality. The NLSY79 was originally devised to study labor market transitions, so we draw the majority of our outcomes from that domain; however, each of these outcomes is featured in at least one published article using NLSY data to study racial inequality (e.g., Chen and Tung 2024; Cheng 2016; Ramey 2018; Ritter and Taylor 2011; Weisshaar and Cabello-Hutt 2020).

Much like the race/ethnicity measures, there are many ways to operationalize the outcome variables. We provide brief summaries of our approaches here, with full details available in the replication package. Wages are measured as log mean hourly wages for 2002–2006; salary is measured as log mean annual salary for 2002–2006.[5] The unemployment outcome is measured as the mean annual percent of weeks unemployed from 2002–2006.[6] Depression is measured as respondents' scores on the seven-item Center for Epidemiologic Studies Depression Scale (CES-D), measured at age 40; higher scores indicate more and/or more frequent depression symptoms (Radloff 1977; Ross and Mirowsky 1989). School discipline is a dichotomous variable indicating whether the respondent had ever been suspended or expelled from school by 1980.

Our analysis of these outcomes can be seen as providing one type of evidence, specifically criterion evidence, toward more holistic assessments of measurement validity for the available race/ethnicity measures. We do not see criterion evidence as the only or even the most important consideration for determining overall construct validity, in line with contemporary validity theory (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014; Sireci 1998). Rather, we present criterion evidence—both concurrent (i.e., estimating outcomes at the same point in time) and predictive (i.e., estimating associations with future outcomes)—to illustrate that it cannot be the primary consideration researchers weigh when designing their studies, in part because our multi-approach and multi-outcome analyses do not point to a single "gold standard" measure. Relative to self-identification recorded in

2002, our analysis provides roughly concurrent criterion evidence, in part because choosing outcomes for later years would result in smaller sample sizes due to attrition. Relative to the measures of race recorded earlier in the survey (e.g., the 1978 screener), our analysis provides predictive criterion evidence that speaks to overall construct validity.

Measuring Race and Ethnicity in the NLSY79

The NLSY79 offers four measures to describe respondents' race/ethnicity. Here, we describe each of the variables available to researchers; in the next section, we outline the many decisions researchers must make to operationalize them and how we approached these choices.

The first race/ethnicity measure is the 1978 screener variable recommended in the user's guide. This variable has three categories: Black, Hispanic, and non-Black, non-Hispanic, which were assigned to respondents based on a combination of factors (Bureau of Labor Statistics n.d.). These factors included coding race "by observation," father's race, inquiring about the ethnicity of household members aged 14 and above, language spoken at home, and whether the family surname appeared on a Census Bureau list of Spanish surnames. This mix of criteria produces a race/ethnicity variable that lacks content or face validity, making it theoretically problematic to interpret, and many researchers may not be aware of this context. As an example of the idiosyncratic choices built into this variable, interviewers were directed to code Filipino descent as Hispanic, while the U.S. census has categorized Filipino Americans as Asian since 1977. Further, many researchers treat the "non-Black, non-Hispanic" category as "White," which is misleading given the diversity of the sample.

The second race/ethnicity measure is a self-identification measure collected in 1979. Respondents were shown a list of 28 responses and asked to identify their "origin or descent," with "American" and "None" coded by interviewers if the respondent volunteered those answers. Respondents could select up to six responses from these 30 categories.

The third race/ethnicity measure is a self-identification measure collected in 2002. Respondents selected all that applied from options that mirror the 2000 Census categories: White, Black or African American, Asian, Native Hawaiian or Other Pacific Islander, American Indian or Alaska Native, and some other race. A separate question collected data on self-identified Hispanic, Latino, or Spanish origins.

The fourth measure is based on interviewer observation. At the end of each interview from 1979–1986 and 1988–1998 (17 survey waves), interviewers marked their observation of whether the respondent was Black, Other, or White. Interviewers were given no instructions on how to record these categories. From a theoretical standpoint, these classifications reflect externally appraised race after an in-person interaction, offering a different type of information compared to the self-identification questions (Roth 2016). They also account for fluctuation in observed race over time, as a respondent may be classified in one category in some survey waves, but a different category in others (see Saperstein and Penner 2012).

Table 1: Operationalization decisions for race/ethnicity measures.

Decision Point	Examples
Metric	Categorical Continuous
Hispanic responses	Combine with "race" variable Separate "ethnicity" variable Dichotomous flag for pan-Hispanic identity Categorical measure for separate national origins
Multiracial responses	Exclusive "multiracial" group Series of non-exclusive dichotomous variables With or without dichotomous flag for multiracial
Aggregation	To a minimum sample size per category Based on theoretical assumptions (e.g. geography) Until empirical tests detect group differences Cluster analysis

These four measures could be seen as proxies for the same underlying construct—e.g., a person's racialized relational position in a society founded on white supremacy (see Vincent 2026)—which capture the same concept with differing amounts of measurement error, or the different types of measures represented—i.e., self-identified origin or descent, self-identified race, and interviewer observed race—could be seen as each reflecting theoretically separate aspects of the "bundle of sticks" (Sen and Wasow 2016; cf. Roth 2016) that comprises an individual's racialized experiences. Our empirical analyses cannot speak directly to these debates, in part because a thorough assessment of measurement error that can distinguish between the presence of one or more constructs requires multiple measures of race that are each measured at three or more points in time (see Blalock 1970 for a discussion). Nevertheless, we highlight that this remains an open question in the hope of inspiring more comprehensive data collection in the future.

Analytical Approach

In addition to deciding which of the four race/ethnicity measures to use, and whether to use just one or multiple, each measure requires researchers to make numerous operationalization decisions. Table 1 summarizes key decision points confronting researchers with examples of possible choices. Researchers must decide on a metric for the measure: for example, the observed race measures can be a continuous count or percentage of times a respondent is classified by interviewers as Black, or a categorical variable for "always, ever, or never" classified as Black, or a dichotomous variable for "ever" classified as Black. For the self-identification measures, researchers must choose whether to combine Hispanic responses with the race categories in various ways or to treat the Hispanic responses as independent from the race responses (e.g., they might use one dichotomous variable for

"Hispanic vs. non-Hispanic" and a different categorical variable representing the race categories). For the measures that allow for multiple responses, researchers must decide how to code respondents who selected multiple categories. For example, researchers could generate a series of dichotomous variables for each category, representing all the categories respondents selected in analyses, or researchers could recode these responses into one "multiracial" category (see Forthal 2025 for a review of multiracial categorization options). Finally, researchers must decide if and how to aggregate categories. This is a particularly pressing question for the self-identification 1979 measure, which contains 30 categories, many of which have very small sample sizes. In response, researchers could establish a minimum sample size per group and aggregate until reaching this minimum. They could aggregate based on factors like regional geography (e.g., "European," "Asian," and "Hispanic") or use cluster analysis methods to assign groups based on similarities on key outcome variables.

Operationalizing Race and Ethnicity Measures

Given the aforementioned considerations for operationalizing the race/ethnicity measures, we test multiple specifications for each measure, as summarized in Table 2 and described in detail in the supplementary materials (see Appendix A in the online supplement). To establish baselines, we also test three operationalizations of the screener variable with different degrees of freedom: the original three-category variable; a dichotomous variable comparing respondents coded as Black to everyone else; and a dichotomous variable comparing respondents coded as Hispanic to everyone else.

The specifications for the two self-identification measures constitute several "families" of approaches: three for the 1979 measure or two for the 2002 measure. Panel A of Table 2 lists each self-identification specification within the three families. We devised these sets of approaches with several key considerations in mind. First, we designed approaches that differ in their level of aggregation and thus the number of necessary model parameters, to allow for direct comparison between the measures. If we used them exactly as collected, the screener measure would require the fewest parameters (two) and the 1979 self-identification measure would require the most (29), making for an unequal comparison in model fit. Second, we designed different specifications to account for the other operationalization decisions highlighted in Table 1. In all, these considerations result in 10 specifications for the 1979 self-identification measure and seven specifications for the 2002 self-identification measure.

The first family of approaches aggregates the self-identification measures to correspond to the 1978 screener categories. This helps us determine whether, holding degrees of freedom constant, the two self-identification measures fit outcome data differently—i.e., provide better or worse criterion evidence for validity—compared to the screener variable. The second set of approaches approximates the 2000 Census categories, allowing for more direct comparison between the two self-identification

Table 2: Summary of race/ethnicity measure specifications.

Panel A. Summary of Self-Identification Specifications				
Family	Approach	Metric	Degrees of Freedom	Summary
Approximating the screener	1a	Categorical	2	Screener categories.
	1b	Dichotomous	1	Hispanic versus all others
	1c	Dichotomous	1	Black versus all others
Approximating Census 2000	2a	Categorical	5	Census 2000 categories.
	2b	Categorical and dichotomous	5	Census 2000 categories and Hispanic flag.
	2c	Non-exclusive dichotomies	5	Census 2000 categories.
	2d	Non-exclusive dichotomies	6	Census 2000 categories and multiracial flag.
Most disaggregated	3a	Categorical	17	*For '79 only.* Most disaggregated possible; small groups in "all remaining responses."
	3b	Non-exclusive dichotomies	17	*For '79 only.* Most disaggregated possible; small groups in "all remaining responses."
	3c	Non-exclusive dichotomies	18	*For '79 only.* Most disaggregated possible; small groups in "all remaining responses;" multiorigin flag.

Table 2: *(Continued)*

Panel B. Summary of Observed Race Specifications				
Race	Approach	Metric	Degrees of Freedom	Summary
Black	Percentage observed	Continuous	1	Percentage of waves in which the interviewer observed that the respondent was Black.
	Percentage observed in deciles	Categorical	10	Same as aforementioned, aggregated into deciles.
	Always, ever, never	Categorical	2	Whether interviewers observed that the respondent was Black in 100% of waves ("always"), in 0% of waves ("never"), or more than 0% but less than 100% ("ever").
Other	Percentage observed	Continuous	1	Percentage of waves in which the interviewer observed that the respondent was Other.
	Percentage observed in deciles	Categorical	10	Same as aforementioned, aggregated into deciles.
	Always, ever, never	Categorical	2	Whether interviewers observed that the respondent was Other in 100% of waves ("always"), in more than 0% but less than 100% ("ever"), or in 0% ("never").
White	Percentage observed	Continuous	1	Percentage of waves in which the interviewer observed that the respondent was White.
	Percentage observed in deciles	Categorical	10	Same as aforementioned, aggregated into deciles.
	Always, ever, never	Categorical	2	Whether interviewers observed that the respondent was White in 100% of waves ("always"), in 0% of waves ("never"), or more than 0% but less than 100% ("ever").

measures, which were collected at different times and with different question wording. The third set is the most disaggregated version possible. This family is only relevant for the 1979 self-identification measure, as the second family of approaches already provides the most disaggregated version for the 2002 measure. Across all approaches, categories with sample sizes below 25 for either women or men are aggregated, as needed, into a residual category of all remaining race/ethnicity responses.

Additional considerations for our specifications of the self-identification measures are highlighted in Table 1. Within each family of approaches, we compare: different metrics, approaches to Hispanic responses, and approaches to multiracial responses. We evaluate how mutually exclusive categorical versions compare to non-mutually exclusive versions using a series of dichotomous variables. We also test how including Hispanic origin responses in the mutually exclusive categorical specifications performs compared to including Hispanic responses as a dichotomous variable separate from the race variable. The latter allows respondents to be identified with both a race category and a Hispanic origin simultaneously. The operationalization of "multiracial" or "multi-origin" responses also varies across approaches: in the mutually exclusive categorical specifications, respondents who report multiple categories are grouped together into one category, while in the non-mutually exclusive specifications, respondents are represented in every category they report, with a value of one on the dichotomous variables of each category selected.[7] In specifications structured around a categorical variable, a "missing" or "noninterview" category is included. In specifications structured around a series of dichotomous variables, "missing/noninterview" is the implied reference category (i.e., respondents with zeros on all dichotomous variables).

Finally, for the interviewer-observed race measures, we devised three approaches which result in nine total specifications (i.e., three for each of the response categories: Black, Other, and White). For each observed race response, we constructed: 1) a set of continuous variables for the percent of survey waves in which respondents were coded in each category; 2) a set of categorical variables for the percent of survey waves coded in each race category divided into deciles; and 3) a set of categorical variables that indicates whether respondents were always, ever, or never coded in a given category. Panel B of Table 2 lists the observed race specifications, and Appendix A in the online supplement provides additional detail. All three observed race approaches combine information across multiple years of data, reducing potential measurement error from relying on any particular year (and thus decreasing attenuation bias in our regression estimates). The second and third approaches also allow for nonlinear relationships between observed race and the outcome(s) of interest.

Single- and Dual-Measure Models

We begin the analyses by assessing each specification individually; fit statistics from these analyses are available in the supplementary materials (see Appendix B in the online supplement). Next, we analyze dual-measure models, using two race/ethnicity measures concurrently. The analyses include the best-scoring

operationalization of interviewer-observed race paired with the best-scoring self-identification measure (from either 1979 or 2002). We chose this approach because observed race and self-identification are theoretically distinct in terms of their associated inequality-producing mechanisms, suggesting they could jointly yield complementary, rather than duplicative, information.

For all analyses, we take an intersectional perspective that examines whether the empirically preferred race/ethnicity measures differ for women and men, and how coefficients differ in direction, magnitude, and statistical significance across gender. Existing research demonstrates the importance of accounting for intersectional inequalities in the labor market (e.g., Browne and Misra 2003), health (e.g., (Hankivsky 2012), and education (e.g., McDaniel et al. 2011), but there are ongoing debates about how best to incorporate an intersectional approach with quantitative methods (see Mahendran, Lizotte, and Bauer 2022 for a recent review). At their most basic, these strategies can include using interaction effects, a series of intersectional dummy variables, or split-sample analysis (see e.g., Ragin and Fiss 2024; Scott and Siltanen 2017; Wilkes and Karimi 2024). We use split samples, stratifying our regression models by binary gender, to allow for estimating model fit statistics separately for women and men.

Evaluating Empirical Differences

Just as there are multiple decisions to make regarding how to operationalize a given measure, there are multiple ways to demonstrate and/or assess the influence of those decisions on empirical results. In addition to considering substantive differences, and their theoretical implications, researchers can choose from a range of statistical metrics, including differences in the amount of explained variation, predictive accuracy, variability of results, and distribution of residuals. Statistical packages and how-to guides for conducting "multiverse analysis," in particular, are increasingly available (Short et al. 2025; Young and Cumberworth 2025). We focus here on common measures of model fit, or explained variance, for their relative simplicity and accessibility to a wide range of audiences.[8] We use these fit statistics for illustrative purposes rather than implying that model fit alone is the most appropriate measure for researchers to consider in their decision-making.

We focus on two model fit statistics, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), which allow us to compare the amount of variance in the outcome of interest that is associated with each measure or operationalization. Generally, AIC is interpreted as selecting the model that can be expected to best predict future data using the existing data, while BIC is described as choosing the true model from a set of alternatives (Aho, Derryberry, and Peterson 2014; Chakrabarti and Ghosh 2011; Vrieze 2012). AIC may be particularly well suited to social science research, given that few social dynamics will be fully explained by a set of variables in a regression model, undermining BIC's assumption that the true model is present in the set. Social scientists often use one or the other criterion without an explicit rationale for their choice or use both without discussing how they adjudicate between them.

AIC and BIC scores are not directly interpretable; rather, they are useful for comparisons across models. A model with lower values on AIC or BIC is understood to perform better compared to models with higher values. A difference of two or fewer points is commonly interpreted as "weak" evidence of model difference (AIC and BIC), four to seven (AIC) or two to six (BIC) is "positive" evidence, and a difference of ten or more points (AIC and BIC) is very strong evidence that the lower-scoring model is better, in terms of variance explained, than the higher-scoring model (Burnham and Anderson 2004; Raftery 1995). Thus, when we find that a particular approach is the "best-performing" model as measured by a particular fit statistic, we interpret this as evidence that the race/ethnicity specification featured in that model has greater criterion evidence for validity than the alternatives, for the specific outcome and gender being analyzed.

When AIC and BIC disagree in our analyses (i.e., they indicate different models perform best), we choose the model preferred by AIC, because our objectives better match the theoretical aim of AIC. We do not expect that the "true" model is present among our set of models, because we are only including race/ethnicity measures as independent variables, so BIC is less well suited to our purposes. We also are interested in predicting how well each model would perform in other datasets, which aligns best with AIC's principles. However, we present both statistics for consideration, in part because, even when there is disagreement on the best-performing model, AIC and BIC may agree on which are the worst, thereby narrowing the set of empirically preferable options. As a robustness check, we conducted K-fold cross-validation with K = 5 (Dräger, Pforr, and Müller 2023; Verhagen 2022) for all analyses and compared results to a constant-only benchmark using the Root Mean Squared Error (RMSE). For interested readers, we also show the R^2 and adjusted R^2 (for models with continuous outcomes) and McFadden's pseudo R^2 (for school discipline, which is coded as a binary outcome) alongside the other fit statistics in the supplementary materials (see Appendix C in the online supplement).[9]

Transparency and Reproducibility

Our analysis demonstrates tools to aid in the movement toward transparency and reproducibility in the social sciences. At its core, the process we model here encourages scholars to think carefully about which measures they use, how the data were generated, how the measures map onto theoretical concepts, and how to operationalize them. As we described previously, and detailed in Appendix A in the online supplement, even after choosing a given measure or measures, scholars must make many operationalization decisions in order to analyze race/ethnicity data. We aim to make these considerations visible in our own work, encouraging more transparency in how such decisions are communicated in the future. Increased transparency would enhance generalizability in two ways: 1) by clarifying to which broader populations results can be applied and 2) by making replication easier, thus improving comparability across studies.

The transparency and reproducibility literature also highlights concerns about questionable research practices (QRPs), including model selection strategies, like

p-hacking, that aim to maximize statistically significant results. These practices thrive in secrecy and depend on concealment (Moody et al. 2022). In contrast, the process we use for illustration focuses on model fit statistics, rather than p-values, and better-scoring models do not reliably produce more statistically significant coefficients. For example, in our analyses, the best model for men's depression is self-identification 2002 Approach 2c. That model yields two statistically significant regression coefficients, while four alternative approaches yield three or more statistically significant coefficients but are nevertheless lower performing overall, according to AIC.[10] To further increase transparency, we share code to reproduce all the approaches and analyses, which can be adapted for a range of research purposes. We encourage other scholars to transparently report which race/ethnicity measures they use, how they operationalize them, and why.

Results

We begin by examining how the NLSY79 screener measure performs compared to the alternative measures in each of our outcome-by-gender analyses, to assess whether the frequently used default offers better criterion evidence for validity than either its content or face validity might imply. We next turn to the results for the single-measure models, which reveal whether certain approaches consistently produce greater criterion evidence for validity, across outcome and gender. Finally, we discuss the results of the dual-measure models and how they compare to the single-measure models. We use these empirical analyses to demonstrate limitations of the status quo, and how alternative approaches can improve both analytical leverage and model fit. We also argue our results imply that researchers cannot rely on empirics alone to determine preferred race/ethnicity measures and operationalization schemes. We return to this point in the conclusion.

Performance of the 1978 Default Screener Measure

Table 3 summarizes the race/ethnicity measures that perform best for each inequality outcome separately for women and men, showing the difference in model fit statistics between the screener variable and the best-fitting measure. First, we find that the 1978 default screener variable never emerges as the best approach for either women or men for any outcome. The fit differences range from 11 to 71 points better than the screener across gender, outcome, and fit statistic, with a mean difference of 32 points, indicating a large empirical penalty from using the screener. In fact, all three screener-based approaches (the default screener categories and the dichotomized versions for Hispanic and non-Hispanic Black respondents) perform among the worst measures on five of the ten outcome-by-gender AIC analyses (see Appendix B in the online supplement for detailed model fit statistics).[11] Even the most parsimonious dichotomous versions are at least 10 points worse than the best-performing measure in 37 out of 40 possible comparisons (across the two approaches, gender, outcome, and fit statistic),[12] and they never come within three points. The five-fold cross-validation results yield similar patterns: no screener measure ever produces the lowest RMSE, and screener measures produce RMSE metrics

Table 3: Summary of best-fitting race/ethnicity measures, by gender, outcome, and fit statistic.

	AIC		BIC	
Outcome	*Women*	*Men*	*Women*	*Men*
Wages	Self-id '79 3c –52	Self-id '79 3c –56	Always/ever/never Black –14	Always/ever/never White –21
Salary	Self-id '79 3c –18	Self-id '79 3c –42	Always/ever/never Black –16	Always/ever/never White –28
Unemployment	Self-id '02 2c Self-id '02 2d –52	Self-id '02 2c –71	Self-id '02 2c Self-id '02 2b –31	Self-id '02 2c –50
Depression	Self-id '79 3b –27	Self-id '02 2c –28	Self-id '02 1c –18	Percentage observed White –15
School discipline	Self-id '79 3b Self-id '79 3c –25	Self-id '79 3c –50	Percentage observed Black –11	Percentage observed Black –15

Note: Numbers indicate the difference between the default screener measure and the best-fitting measure on AIC or BIC. Negative values indicate the best-fitting measure has an improved fit compared to the screener measure.

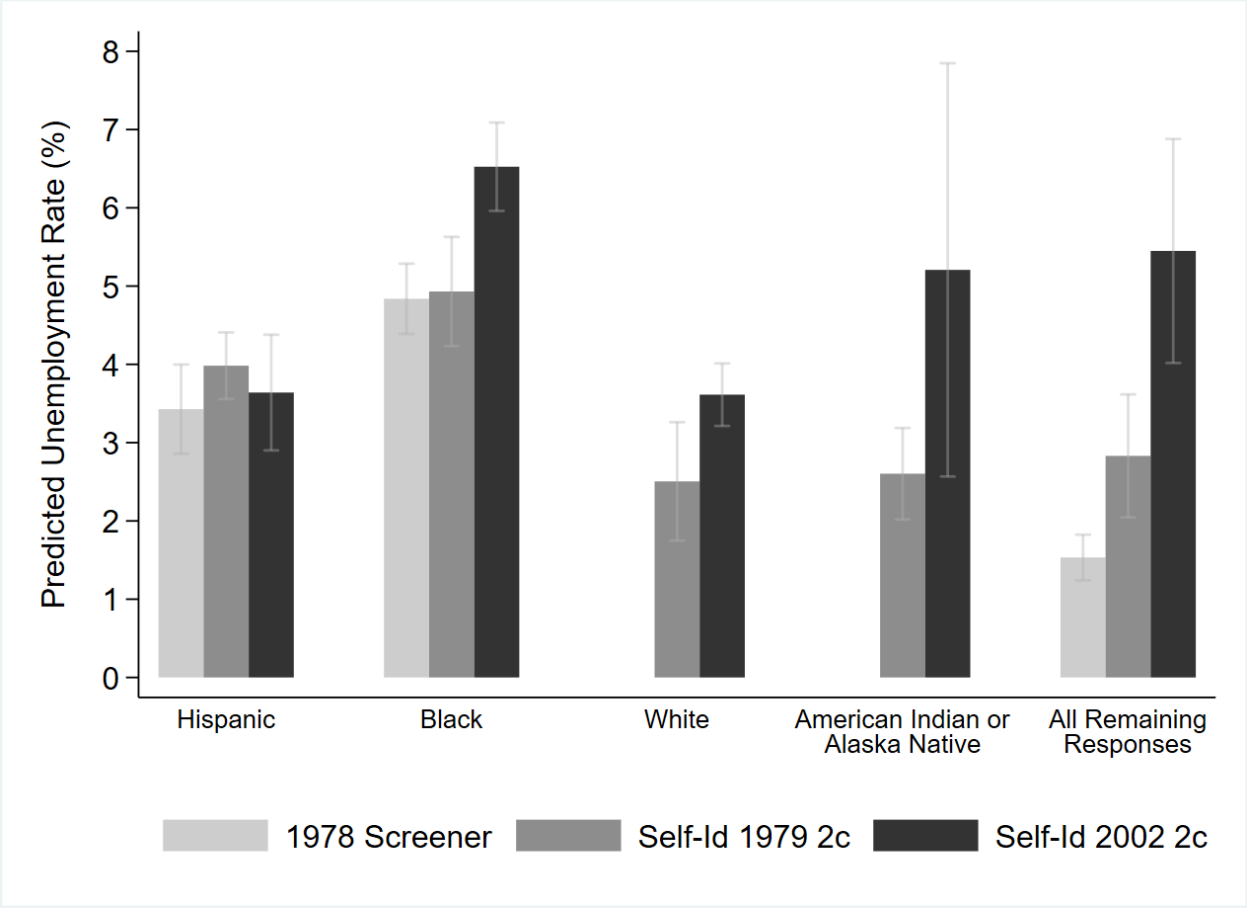

Figure 1: Men's predicted unemployment rates. *Source*: National Longitudinal Survey of Youth 1979. *Note*: $N = 6,403$. 95 percent confidence intervals. Unemployment rates are calculated based on the mean annual percent of weeks unemployed from 2002–2006.

higher than the average in 15 of 30 gender-by-outcome tests.[13] This is a primary contribution of our analysis, highlighting that scholars would do well to consider whether the other race/ethnicity measures, or a combination of them, would better suit their theoretical aims and better approximate the empirical patterns in the data than does the screener race classification.

To illustrate the potential for improved interpretation with alternative race/ethnicity measures, we compare results using the three-category screener measure and the approach that offered the greatest improvement over the screener across all models: self-identification 2002, Approach 2c for men's unemployment. On this outcome, 2002 Approach 2c performs better than the default screener model by 71 points on AIC and 50 points on BIC, despite using two more parameters. It also outperforms the analogous Approach 2c with the same number of parameters, drawing on 1979 self-identification data (90 points on AIC and 89 points on BIC). Figure 1 plots the predicted unemployment rate for men using each of these approaches.

Comparing results suggests little variation in the predicted unemployment rate for men classified as Hispanic across these approaches, but important differences for everyone else. For example, the average unemployment rate for men who self-identified as Black in 2002 is about 26 percent higher than estimates for Black men using either of the other two measures.[14] Approach 2c for 2002 also shows that rates of unemployment are not consistent for men who were aggregated together in the screener's residual category. Men who self-identified as White in 2002 (alone or in combination with other responses) have lower levels of unemployment than men who identified as American Indian or Alaska Native, Asian, Native Hawaiian, or "some other race" (the latter three categories are aggregated in "All remaining responses"). Importantly, unemployment rates for men who self-identified as White in 2002 are similar to men who self-identified as Hispanic in 2002 (3.61 percent and 3.64 percent, respectively), suggesting that this potential axis of stratification is less crucial for understanding unemployment disparities than implied by either the screener measure or the 1979 approach.

There may be several reasons why the self-identification 2002 Approach 2c specification outperforms the screener: it is a self-identification measure; it is measured closer in time to the unemployment outcome; it disaggregates respondents to a greater degree; and it allows for multiracial respondents to be represented in multiple categories. Comparing with self-identification 1979 Approach 2c helps to narrow down which of these reasons is driving the difference in results. Doing so suggests that the improvement in model fit from 2002 Approach 2c is likely not driven primarily by being a self-identification measure or allowing for multiple responses, which the 1979 and 2002 approaches have in common. Instead, it may be driven by measuring race/ethnicity closer to the time of unemployment combined with differences in how the two questions (and their answer options) were worded. Whatever the reason, we find that using a different race/ethnicity measure not only statistically improves model performance, it also changes our empirical understanding of how racial/ethnic inequality in unemployment is patterned among men, subsequently changing the conclusions researchers might draw for how best to target interventions.

Comparing Single-Measure Approaches

Although the default screener never performs best, there is also no one "winner" that emerges across all analyses. We see substantial variation in which type of measure—and which specification of each measure—is empirically best across gender, outcome, and fit statistic (see Table 3 and Appendix B in the online supplement). Only for unemployment is the same specification (2002 Approach 2c) consistently preferred, in terms of criterion evidence, for both women and men, and by both AIC and BIC.

Across outcomes, self-identification measures tend to perform best on AIC, while observed race measures are often—but not always—rewarded for parsimony by BIC. Among the self-identification specifications, the 1979 measure is preferred over the 2002 measure in seven of the 10 outcome-by-gender AIC comparisons and always in its most disaggregated and non-mutually exclusive forms (Approaches

3b and 3c). Among the observed race specifications, the categorical "always-ever-never" schemes are favored in four instances (wages and salary for women and men) and the continuous percentage approaches are also favored in four instances (depression for men and school discipline for women and men). The five-fold cross-validation yields nearly identical results to the AIC findings for the best-fitting measures and approaches.[15] Thus, we find that models with fewer parameters do not consistently perform better, even according to BIC, and even when comparing more or less aggregated versions of the same measure.

In most cases, the best-performing single measure is the same specification for women and men (on AIC: wages, salary, unemployment, and school discipline; on BIC: unemployment and school discipline). However, there are also outcomes where the same "always, ever, never" approach for observed race performs best but using different categories for women (Black) and men (White), and where entirely different measures perform best for women and men (on AIC: depression). We return to consider the importance of accounting for intersectionality in these results, in more detail, in the discussion.

Comparing Dual-Measure Approaches

Next, we uncover that incorporating a self-identification measure and an observed race measure simultaneously often outperforms any single measure alone, despite adding further parameters to the models. In eight out of the 10 gender-by-outcome tests, dual-measure models strongly outperform all single-measure models, by five to 31 points on AIC (Table 4). In the final two tests, men's depression and school discipline, dual measures outperform the best single measures by one and two points, respectively, indicating weak evidence of improvement. The multiple-measure models even outperform single-measure models on BIC in three instances, which is quite notable given how strictly BIC penalizes adding additional parameters to models. In all three cases (women's unemployment, men's unemployment, and men's depression), the magnitudes of BIC-score differences are high, with seven, 60, and 33 points of improvement, respectively. In the five-fold cross-validation, dual-measure models also produce lower RMSEs than single-measure models in about half (nine of 20) of gender-by-outcome tests. All together, these findings underscore that self-identification and observed race frequently provide different types of information, suggesting that scholars can sharpen both their theoretical conclusions about the state of racial/ethnic inequality and the empirical performance of their models by analyzing these complementary measures simultaneously.

To give one example of how pairing self-identified race/ethnicity with observed race may hone theoretical conclusions, as well as empirical model performance, consider women's experiences of school discipline (see Figure 2). The dual-measure analysis combines the best-performing self-identification measure for this outcome, Approach 3b for self-identification in 1979, with the best-performing observed race measure, always, ever, never observed Black. By adding the observed race measure, we see not only improved model performance (a five-point difference on AIC), but also that differences in the predicted rates of school discipline cannot be attributed to differences in self-identification alone. In the dual-measure model, the predicted rate of having ever experienced school discipline for women

Table 4: Summary of dual-measure analyses, by gender, outcome, and fit statistic.

Gender	Outcome	Approaches	AIC	BIC	Difference from Best Single-Measure Model AIC	Difference from Best Single-Measure Model BIC
Women	Wages	Self-id '79 3c and always/ever/never Black	5,702	5,826	–28	39
	Salary	Self-id '79 3c and always/ever/never Black	8,360	8,484	–25	78
	Unemployment	Self-id '02 2c and always/ever/never White	43,999	44,053	–20	–7
	Depression	Self-id '79 3b and Percentage observed White	24,796	24,910	–12	74
	School discipline	Self-id '79 3b and always/ever/never Black	5,563	5,690	–5	88
Men	Wages	Self-id '79 3c and always/ever/never White	5,653	5,778	–22	49
	Salary	Self-id '79 3c and always/ever/never White	8,358	8,482	–31	60
	Unemployment	Self-id '02 2c and Percentage observed White	46,486	46,486	–19	–60
	Depression	Self-id '02 2c and Percentage observed White	22,719	22,719	–1	–33
	School discipline	Self-id '79 3c and Percentage observed Black	7,526	7,654	–2	71

Note: Difference from best single-measure model is calculated by subtracting the single-measure models' lowest AIC from the dual-measure model's AIC, and likewise for BIC. A negative value indicates that the dual-measures model fits better than any of the single-measure models. Full regression output for dual-measure analyses is available upon request.

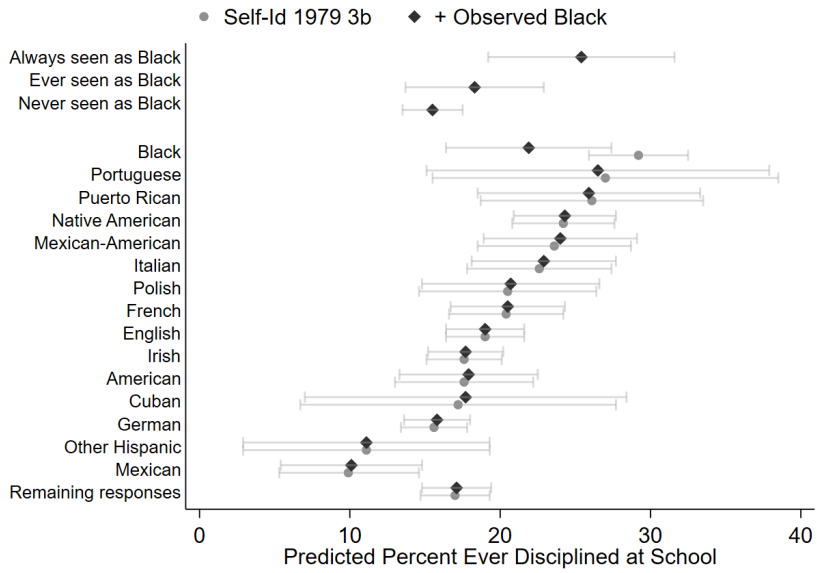

Figure 2: Single- and dual-measure analyses of women's school discipline. *Source*: National Longitudinal Survey of Youth 1979. *Note*: $N = 6,049$. 95 percent confidence intervals. School discipline is measured as whether or not respondents were ever suspended or expelled from school. See Appendix Table A7 in the online supplement for the percent of respondents in each self-identification 1979 Approach 3b category that were always, ever, and never seen as Black by interviewers.

self-identifying as Black shrinks by a third, from 29.2 percent to 21.9 percent, while the estimate for women who are always seen as Black yields an additional estimate of 25.4 percent.[16] This suggests an important role for teachers' and other school agents' interpretation of a student's race in shaping school discipline disparities. Thus, the dual-measure analysis adds another type of information about the source of observed inequality, which can aid our understanding of why and how school suspensions and expulsions are unequally distributed.

Discussion and Implications

Our finding that the best-fitting approach to measuring race/ethnicity varies by gender, by outcome of interest, and by model fit statistic suggests there is no one-size-fits-all empirical solution for researchers. We also show that researchers' operational decisions have consequences for the study's substantive conclusions and implications beyond basic statistical assessments like model fit. These results support key tenets of "QuantCrit" and are in line with the recent turn toward "multiverse analysis" and broader attention to "researcher degrees of freedom" in shaping research results.

In uniting these perspectives, our study not only offers a cautionary tale in model robustness but also provides clear guidance for operationalizing race/ethnicity measures in a widely used dataset for social science research. Our criterion evidence points to the importance of including multiple measures of race and of accounting for observed race, in particular, in studies of inequality in the U.S. The results also

leave no doubt as to which measure is an ill-fitting tool for understanding racial inequality: the NLSY79 screener variable. Relying on the default screener variable has long been problematic from the standpoint of face validity for two reasons: because it was generated based on a range of factors from father's race and surname to language spoken at home, and because its "non-Black, non-Hispanic" category should not be interpreted simply as "White," as many researchers do. We show that the screener race measure also comes up short empirically (i.e., in terms of criterion evidence) in accounting for racial inequality across domains spanning the labor market, health, and education.

Our results also have implications for best practice in several additional areas: conducting intersectional analysis in quantitative studies of inequality and how to incorporate respondents who identify as Hispanic or multiracial. Beyond carefully selecting a single- or multiple-measures approach for their research, scholars can further improve their modelling strategies and the strength of their interpretations by considering these additional factors.

Intersectional Patterns

Our findings highlight the utility of considering women's and men's experiences of racial inequality separately. Across our five outcome variables, we find several cases where the measurement approach that performs best varies by gender. In other cases, we find that even if the same measurement approach is preferred, the interpretation of coefficients' statistical significance, direction, and magnitude can still vary for women and men. Both types of outcomes validate the need for an intersectional approach in studies of inequality.

For the depression outcome, the best-performing single race/ethnicity measure for women and men was not just different specifications or approaches, but fully different measures (1979 versus 2002 self-identification). We use this case to illustrate the usefulness of splitting analyses by gender. We estimated pooled models for both the women's and men's preferred measures to compare with the intersectional results. Table 5 highlights how differences between the women's and men's results are lost in the pooled data. Models 1–3 show results for Approach 3b, the best-performing model for women, and Models 4–6 show the results for Approach 2c for self-identification 2002, the best-performing model for men. Were we to run only the pooled models, we would conclude that Approach 3b performed better overall, with a 19-point lower AIC. However, either set of pooled results seems to overstate some relationships and understate others that are better reflected in the gender-specific models. For example, the pooled results in Model 1 overstate depression scores for men with Black ancestry in 1979 (relative to Model 3), while the pooled results in Model 4 understate the depression scores of men who identified as American Indian or Alaska Native in 2002 (relative to Model 6). Similarly, applying the best-fitting approach for men to the data for women homogenizes the experiences of women who identified as White in 2002 (see Model 5), despite the existence of statistically significant coefficients of opposite signs among women who reported European origins in 1979 (see Model 2). These different patterns have practical implications. For instance, if a researcher were using this observational analysis to inform the design of a mental health intervention, the intersectional

Table 5: Depression models: Best-performing approaches, pooled and by gender.

	Model 1 Approach 3b	Model 2 Approach 3b	Model 3 Approach 3b
	Pooled	Women	Men
Series of dichotomous variables			
Black	0.65**	0.82**	0.33
	(0.15)	(0.22)	(0.20)
Cuban	–0.48	–0.49	–0.45
	(0.42)	(0.66)	(0.51)
Mexican	–0.26	–0.44	–0.20
	(0.25)	(0.36)	(0.33)
Mexican-American	0.15	0.17	–0.02
	(0.20)	(0.30)	(0.27)
Puerto Rican	1.46**	1.39**	1.41**
	(0.28)	(0.42)	(0.36)
Other Hispanic	–0.11	–0.57	0.23
	(0.41)	(0.61)	(0.54)
English	0.11	0.06	0.12
	(0.14)	(0.20)	(0.18)
French	0.30	0.63*	–0.25
	(0.19)	(0.27)	(0.27)
German	–0.39**	–0.44*	–0.44*
	(0.13)	(0.19)	(0.17)
Native American	1.05**	1.16**	0.74**
	(0.17)	(0.24)	(0.23)
Irish	0.07	0.19	–0.14
	(0.14)	(0.20)	(0.18)
Italian	–0.16	–0.23	–0.14
	(0.21)	(0.32)	(0.28)
Polish	0.07	0.43	–0.27
	(0.26)	(0.39)	(0.33)
Portuguese	–0.83	–0.39	–1.12
	(0.69)	(1.06)	(0.89)
All remaining responses	–0.11	–0.05	–0.26
	(0.13)	(0.19)	(0.17)
American	0.43	0.80*	–0.07
	(0.23)	(0.35)	(0.30)
Constant	3.09**	3.46**	2.86**
	(0.13)	(0.19)	(0.17)
Observations	8,364	4,255	4,109
AIC	47,740	24,808	22,739
BIC	47,859	24,916	22,847
R^2	0.015	0.018	0.015
Adjusted R^2	0.013	0.014	0.011
Five-fold CV RMSE	–	4.460	3.848

Table 5: *(Continued)*

	Model 4 Approach 2c ('02)	Model 5 Approach 2c ('02)	Model 6 Approach 2c ('02)
	Pooled	Women	Men
Series of dichotomous variables			
Black	0.61**	0.86**	0.27
	(0.15)	(0.23)	(0.20)
Hispanic	0.18	0.22	0.07
	(0.14)	(0.21)	(0.18)
White	–0.26	–0.10	–0.50**
	(0.13)	(0.20)	(0.17)
American Indian or Alaska Native	1.71**	0.99	2.53**
	(0.41)	(0.58)	(0.57)
All remaining responses	0.19	–0.15	0.52
	(0.25)	(0.36)	(0.33)
Constant	3.28**	3.63**	2.99**
	(0.12)	(0.19)	(0.16)
Observations	8,364	4,255	4,109
AIC	47,759	24,825	22,720
BIC	47,802	24,863	22,758
R^2	0.010	0.009	0.014
Adjusted R^2	0.009	0.008	0.013
Five-fold CV RMSE	–	4.475	3.840

Source: National Longitudinal Survey of Youth 1979.
Note: Standard errors in parentheses. $^{*}p < 0.05$ and $^{**}p < 0.01$ (two-tailed tests). Best-fitting approaches for women and men were determined using AIC.

models point to the importance of considering a targeted approach for men who identify racially as American Indian and Alaska Native (see Models 5–6). Similarly, designing a mental health intervention specifically for Black women could be useful, as they report statistically significantly greater midlife depression symptoms, while Black men do not (Models 2–3 and Models 5–6). In contrast, using the pooled analyses would not reveal these differences by gender, leading to more generic and potentially less effective interventions.

Applying the best-performing women's measure to the men's data does indicate that men who reported Puerto Rican origin in 1979 have much higher later-life depression scores than their Hispanic counterparts (see Model 3), which is obscured in the men's best-fitting measure (see Model 6). However, we interpret this result with caution, given both concerns about p-hacking and the potential for a single coefficient to be statistically significant by chance in a model with so many parameters. Instead, we consider this collective modeling exercise as further evidence of the value of considering multiple measures of race in analyses of inequality and of not assuming the same race/ethnicity measure will perform equally well across both the outcomes of interest and across gender.

For most of the outcome variables we examined, though, women's and men's best-performing measures were the same, and it would be tempting to conclude that

intersectional analysis is unnecessary in these cases. However, splitting the analyses by gender still revealed important differences, as the substantive conclusions differ even when the best-performing measure did not. For example, the top-performing approach for both women's and men's wage outcomes is Approach 3c for self-identification in 1979 (see Table 6). The model for women highlights intragroup heterogeneity: women reporting Cuban origin are at the top of the wage distribution, followed by women with Puerto Rican origin, while women reporting Mexican origin are at the bottom, with Native American women just above them (see also Ferraro et al. 2025). For men, those reporting Cuban origin are also at the top, but men with Black ancestry are at the bottom, with Mexican–American origin men just above them. Thus, the rank ordering of wage outcomes by race/ethnicity category differs meaningfully by gender. This example demonstrates how useful an intersectional analysis is, even in cases when the same race/ethnicity measure performs best. That is, scholars should not default to pooling the data, even if they consider multiple approaches and find the same specification statistically performs best for women and men, as an intersectional approach also can reveal substantively meaningful patterns and be motivated on theoretical grounds.

Incorporating Hispanic Respondents

Our findings do not establish one single best-fitting measure, in terms of criterion evidence, for operationalizing Hispanic analytic categories, but they do point to several takeaways that call into question current standard practices. First, we find that non-mutually exclusive approaches to incorporating Hispanic origin responses tend to fit better than their mutually exclusive counterparts. For example, when we consider the 1979 self-identification approaches that allow "Hispanic" responses to count alongside other racial/ethnic responses (Approaches 2b, 2c, 2d, 3b, and 3c), we find at least one such specification outperforms both specifications that assume Hispanic origin responses supersede other racial/ethnic responses (Approaches 2a and 3a) in seven of 10 gender-by-outcome AIC analyses.[17] This runs counter to how Hispanic origins have been reported by many researchers, including in U.S. Census Bureau publications (Grieco and Cassidy 2001; Humes, Jones, and Ramirez 2011).

Second, we note that the most disaggregated approaches, which account for different subgroups and origins, often fit the data best, rather than a classification scheme that uses an aggregate Hispanic category. For example, when comparing measurement strategies for using the 1979 self-identification data, the specifications with disaggregated Hispanic categories (3a, 3b, and 3c) outperform the specifications with a single aggregated Hispanic category (2a, 2b, 2c, and 2d) in eight out of 10 gender-by-outcome AIC analyses—despite using more than three times as many parameters.[18] Further, our cross-validation exercise shows that using a binary contrast to compare Hispanic respondents versus all other respondents (Approaches 1b for 1979 self-identification, 2002 self-identification, and the screener) performs worse than the constant-only benchmark at least half the time. That is, in 16 of 30 gender-by-outcome tests, modeling inequality by simply assigning all women or men their respective average outcome produces less error than comparing "Hispanic" and "Non-Hispanic" respondents. Together, these results point to the risks

Table 6: Wage regression analyses using 1979 self-identification approach 3c, by gender.

	Model 1: Women	Model 2: Men
Series of dichotomous variables		
Black	–0.08**	–0.25**
	(0.03)	(0.03)
Cuban	0.27**	0.18*
	(0.08)	(0.08)
Mexican	–0.14**	–0.14**
	(0.05)	(0.05)
Mexican—American	–0.01	–0.15**
	(0.04)	(0.04)
Puerto Rican	0.11*	–0.10*
	(0.06)	(0.05)
Other Hispanic	0.00	0.01
	(0.08)	(0.08)
English	–0.04	–0.00
	(0.03)	(0.03)
French	–0.05	0.02
	(0.04)	(0.04)
German	0.09**	0.07*
	(0.03)	(0.03)
Native American	–0.10**	–0.17**
	(0.03)	(0.04)
Irish	–0.01	0.03
	(0.03)	(0.03)
Italian	0.07	0.13**
	(0.04)	(0.04)
Polish	0.07	0.06
	(0.05)	(0.05)
Portuguese	0.21	0.01
	(0.13)	(0.12)
All remaining responses	0.08**	0.05
	(0.03)	(0.03)
American	–0.03	–0.04
	(0.05)	(0.04)
Multi-origin	0.06	0.07
	(0.03)	(0.04)
Constant	2.56**	2.84**
	(0.02)	(0.03)
Observations	3,680	3,726
AIC	5,730	5,675
BIC	5,842	5,787
R^2	0.039	0.097
Adjusted R^2	0.035	0.093
Five-fold CV RMSE	0.527	0.518

Source: National Longitudinal Survey of Youth 1979.
Note: Standard errors in parentheses. $^{*}p < 0.05$ and $^{**}p < 0.01$ (two-tailed tests).

in assuming a monolithic Hispanic experience, in line with other research (Alvero, Giebel, and Pearman 2024; Ferraro et al. 2025; Jiménez, Fields, and Schachter 2015).

Sample size and data availability may prohibit disaggregating Hispanic responses for many purposes, but our results underscore this is an important limitation that should be acknowledged. Our findings also echo recent guidance from the National Academies of Sciences, Engineering, and Medicine (Wilson et al. 2025) about carefully justifying all racial and ethnic category aggregation decisions, given their potential substantive influence on the results. Although it may have been politically advantageous to emphasize similarities among Mexicans, Puerto Ricans, and Cubans in advocating for greater attention to "Hispanic" people (see Mora 2014), the extent of shared inequality experiences remains an open empirical question. Indeed, intracategorical inequality should be kept in mind when working with all ethnoracial categories, including "White" (Read and Fairfax 2025).

Incorporating Multiracial Respondents

Similarly, while our results do not establish one "right" way to account for multiple racial/ethnic responses, we do see two features of approaches that offer stronger criterion evidence than others. First, despite the common practice of reporting results for a combined "two or more races" category, we find that aggregating people who select multiple racial/ethnic responses into one mutually exclusive group (Approaches 2a and 3a) often does not perform well. Instead, approaches that use a series of dichotomous variables for each race/ethnicity category (Approaches 2c, 2d, 3b, and 3c), which allow respondents to be represented in each category they indicate, tend to perform better. In all 10 gender-by-outcome tests, based on AIC, at least one of the specifications using a series of dichotomous variables outperforms those with a categorical, mutually exclusive approach. Although the non-mutually exclusive strategy may better represent multiracial identities in the abstract, as being the sum of multiple parts, it does visually erase multiraciality from the model results (see, e.g., Figure 1) and requires additional calculation to produce specific estimates (e.g., for respondents who identified as Black and White or Asian and White).

Second, we find that including a separate indicator for people who selected multiple responses alongside the series of dichotomous variables sometimes performs better than not including it, despite adding an additional parameter. Comparing Approaches 3c and 2d (both of which include a multiracial flag) to Approaches 3b and 2c (which do not), we find that at least one of the specifications that accounts for a shared multiracial experience provides a better fit among the non-mutually exclusive approaches in five of the 10 gender-by-outcome AIC analyses, along with two ties.[19] As Forthal (2025) notes, depending on the outcome of interest, there could be unique experiences for multiracial people with different racial backgrounds, shared experiences of being multiracial (such as monoracism), a combination of the two, or neither. This points again to the importance of weighing theoretical mechanisms and how they align with different measures' content or face validity, alongside criterion evidence and statistical performance, when producing empirical research on inequality.

Our findings regarding multiracial classification also offer an important update to the conclusions of Light and Nandi (2007), who argued, "multiracial respondents can simply be reassigned to *one* of their reported race categories [. . .] reassignment has virtually no effect on the race distribution or on the explanatory power of race variables" (2007, p.141, emphasis added). In contrast, we find that analyzing all available race/ethnicity data generally yields empirically better models and represents a more respectful stance towards multiracial respondents (Giebel 2023). Carefully considering all these potential pros and cons is important when assessing different approaches to including multiracial people in research (see Lam-Hine et al. 2024 for further discussion).

Conclusion

Combining the principles of transparency and reproducibility with the theoretical insights of "QuantCrit," we shed light on an overlooked issue: the need to transparently describe which racial/ethnic measures are used, how they are operationalized, and why. We hope that by highlighting how many operationalization decisions must be made to use common race/ethnicity measures, like those in the NLSY79, scholars will better understand the theoretical and empirical stakes inherent in their use of these data. By sharing the code to reproduce the approaches and analyses described in this paper, we also aim to encourage scholars to go beyond using the default screener measure in their analyses, instead adopting one or multiple alternate measures better suited to their needs. To formalize the process we describe, we conclude by offering a set of five key research design considerations that emerge from our analyses and are echoed by the recent NASEM guidance for biomedical researchers (see Box 1). Although the roadmap we outline here relies on existing best-practice standards, there is value in re-iterating and synthesizing them in one place for the benefit of newer inequality researchers as well as subject-matter experts who may be familiar with some, but not all, of these tenets.

We recommend researchers begin by carefully theorizing about the mechanisms they expect will be most relevant to their study, including evaluating whether some measures have greater validity than others depending on the inequality processes of interest. The theory guiding one research question may point to the relevance of how others perceive a person's race/ethnicity, while the theory shaping another project may be more concerned with how a person self-identifies with detailed racial/ethnic groups, and a third project might find both types of data useful frames for analysis concurrently (see Saperstein et al. 2016 for examples of how to leverage multiple measures of race in analysis). Another project might find fluctuations over time to be key to its aims—for either theoretical reasons (e.g., acknowledging fluidity in how people self-identify or are perceived by others) or methodological reasons (e.g., to address concerns about measurement error)—making data from the same measure at multiple timepoints important. Each of these considerations should be weighed in conjunction with any multiverse-inspired analysis and before the type of model-fit assessment we illustrated here. That said, for studies using NLSY79, we do have difficulty imagining theoretical interests that would be best matched to the default screener variable, given how it was constructed and the

Box 1: Key Considerations for Operationalizing Race and Ethnicity.

1. **Consider the relevance of different measures.** Theorize about the mechanisms involved in (re)producing racial/ethnic inequality in your study's outcomes of interest. For example, does it matter how your respondents are perceived by others? If so, starting with observed race variables may be most useful. Does it seem likely that exposure to inherited advantage or disadvantage will play a role in the dynamics under study? Do you theorize that people's contemporary understanding of themselves is a key factor? If so, measures that reflect racial ancestry or current racial identity may be the best tool. Could fluctuations over time be important, either for substantive reasons or due to concerns about measurement error? If so, drawing from measures at multiple points of time could be helpful, when such data are available. When relevant measures are not available, the limitation should be acknowledged, and substantive interpretations should be adjusted accordingly.
2. **Consider different classification schemes within the family of measures that aligns your theoretical framework.** Theorize about how different specifications within your answers to the first consideration might matter for your study (see Table 1). Does your research question require categories that are more (or less) detailed? Do you expect multiracial respondents have unique experiences based on their specific racial background, or are they likely to have a shared experience (e.g., of monoracism) that may be relevant to your outcome, or both? Similarly, should Hispanic origin be treated as a separate but shared experience or as one among many possible racial/ethnic categories? In the absence of clear theoretical guidance, empirically test different sets of specifications (e.g., corresponding to different theories) and consider criterion evidence for validity.
3. **Consider multiple measures of race/ethnicity.** After conducting the above exercises, consider whether adding one or more additional measures of race/ethnicity would enhance your theoretical leverage. Calculate the variance inflation factor to check that multiple measures do not introduce multicollinearity. Empirical tests of model fit could then offer criterion evidence of validity for multiple measures, which should be weighed alongside the substantive benefit.
4. **Consider the intersectional implications.** Are other axes of inequality, like gender, also likely to matter? When cell sizes allow, and either theory or previous research point to important potential variation, conduct analyses informed by intersectional perspectives.
5. **Strive for transparency and reproducibility.** At a minimum, be clear about which measure(s) of race/ethnicity you use and how you operationalize them, and ideally, justify your choice(s) with theory, criterion evidence of empirical performance, or both. In all the above considerations, do not base decisions on the number of statistically significant coefficients. Instead, use relevance to theoretical expectations, along with assessments of model robustness or fit, to guide decision-making. When using measures of model fit, specifically, consider which corresponds best to your theoretical and empirical aims: are you looking for the most parsimonious model, the model with the most variance explained, or does out-of-sample prediction or some other criterion metric of validity matter most?

availability of other measures. Further, as the screener example suggests, concerns about content or face validity should take precedence over solely criterion-based evidence when these perspectives offer competing assessments.[20]

Next, we encourage researchers to consider the range of empirical specifications that are possible for any given race/ethnicity variable. As we have demonstrated, even after selecting which measure(s) best match a study's theoretical interests, there are numerous additional decision points that are necessary to generate variables for analysis (see Table 1) but for which researchers rarely provide explicit justification. In some cases, theoretical reasoning may help narrow the set of operational approaches best suited to a project. In other cases, when no clear theoretical guidance resolves all decision points, researchers could empirically test the consequences of implementing different choices or err on the side of including more disaggregated categories (see e.g., Schwabish and Feng 2021). This step in descriptive or observational analysis can potentially help to improve theory abductively by highlighting which distinctions seem to be most relevant for particular outcomes (see Engzell and Mood 2023) and necessarily precedes efforts to provide more cutting-edge causal analysis of potential interventions (e.g., Lundberg 2024).

If researchers have focused on a single measure of race/ethnicity to this point, we suggest they consider the theoretical and empirical leverage that could be gained by including one or more additional measures of race/ethnicity. For example, if researchers have thus far focused on a self-identification measure, they might consider how an observed race measure could enhance the analyses, adding information about how racial inequality emerges through different processes. In doing so, researchers should carefully weigh concerns about multicollinearity (O'Brien 2007).

We also encourage researchers to consider whether other axes of inequality might matter in their analyses. A large body of prior research documents that intersectionality is important for understanding the overall structure of social inequality. In this study, we have shown that the best-performing race/ethnicity measures often differ across gender, offering further evidence that racial inequality is patterned differently for women and men. Thus, if researchers develop theoretical reasons why their analyses could vary along multiple axes of inequality, and such analyses are empirically possible (e.g., if cell sizes allow), we recommend researchers conduct intersectional analyses. See Mahendran, Lizotte, and Bauer (2022) and Spierings (2023) for reviews of incorporating intersectional perspectives in quantitative research.

Finally, we echo calls for researchers to implement open science principles of transparency and reproducibility. As we have shown, working with race/ethnicity measures is not straightforward, even in a publicly available dataset like the NLSY79. Researchers confront many decision points, and many different approaches are possible. Thus, it is essential to clearly describe the chosen approach, so that readers can fully understand the details—which are important to interpreting and generalizing results—and replicate the analyses. Ideally, researchers will not only document their choices but also justify their decisions, drawing on theory and empirical evaluations, as appropriate. Similarly, researchers should justify their approach to model selection, including considering cross-validation exercises to evaluate robustness to over-fitting (Verhagen 2022), as well as approaches that more

explicitly account for theory and/or model uncertainty (see, e.g., Muñoz and Young 2018a; Schultz 2018).

In sum, we draw on data from the widely used NLSY79 to highlight the range of decision points confronting inequality researchers who work with race/ethnicity data, decisions that are rarely detailed in published articles, regardless of whether racial disparities are the explicit focus of research. Although we demonstrate that there is no consistent "gold-standard" approach to operationalizing race/ethnicity across inequality domains and gender, we also find that the NLSY79's default screener variable, which has questionable theoretical value, is never the empirically best choice in our analyses. This underscores the need to revisit current taken-for-granted approaches, and we draw on existing best-practice recommendations for the use of race/ethnicity in research to guide inequality researchers in navigating this complex terrain.

We hope future research will not only explore alternative strategies, including analyzing multiple measures of race/ethnicity simultaneously and accounting for fluidity in categorization, but also devote greater attention to the inequality-producing mechanisms that best account for their results. As our analyses underscore, what may seem like a technical statistical exercise and a simple call for more transparency also has consequences for how we understand and attempt to address inequality in the United States.

Notes

1 We treat the terms race and ethnicity interchangeably in this paper, in part because the survey questions we are studying do not always clearly distinguish between the two concepts. We also use the terms origin or ancestry interchangeably when referring to measures that ask explicitly about geographic origin or descent. When measures use only categories such as "Black" and "White," we refer to them as measures of race but, collectively, we refer to this question type as measuring race/ethnicity for simplicity.

2 Because this practice is widespread, we feel it would be inappropriate to single out a particular paper or set of authors for illustration. Further, the list of studies would be too long to cite in its entirety.

3 The multiverse analysis approach has been extended to data collection (Harder 2020), but our focus is on data processing and analysis of preexisting datasets.

4 We use the term "gender" and the categories "women" and "men" when interpreting our analyses but, in line with our treatment of race/ethnicity, it is important to note how these data were collected and acknowledge their limitations. In the NLSY79 data, respondents are recorded as "female" or "male" by interviewer observation. We use the measure from the 1978 screening interview, in which interviewers only asked respondents to self-identify if the answer was "not obvious" to them (National Longitudinal Surveys User Services n.d.). This method of data collection generates both constraints on our analyses and limitations in their interpretation: (1) the concepts of sex, as a distinction based on biological criteria, and gender, how people present and interact socially, are conflated; (2) the available response categories are binary, despite greater diversity in both sex and gender; and (3) interviewer observations may differ from other methods of categorization, including respondents' self-identification (see Westbrook and Saperstein 2015 for further discussion of these issues). Indeed, screener categorizations for at least 48

respondents were later edited because of inconsistencies between the originally recorded data and answers to fertility questions in 1982 (National Longitudinal Surveys User Services n.d.).

5 We drop wage and salary outliers below the second percentile and above the 98th percentile, calculated by year, before taking the mean and applying the logarithmic transformation.

6 We calculate the percent of weeks unemployed using only those weeks that have data on labor force status. For example, if labor force status is available for only 10 weeks out of a year, and the respondent was unemployed for five of those weeks, their unemployment rate is calculated as 50 percent.

7 What constitutes a "multiracial" or "multi-origin" response is determined before we establish which categories fall below the sample size criteria for inclusion as a separate estimate. See Appendix A in the online supplement for further details.

8 Measures of model fit are also designed to assess whether adding more parameters offers more explanatory power, a consideration that is particularly relevant in our analysis of single- vs. multiple-measure models.

9 We depart from Light and Nandi (2007) and do not use R^2 statistics as our primary measure of model fit for several reasons. First, R^2 does not penalize model complexity, as AIC and BIC do, and our race/ethnicity specifications vary in their number of parameters (see Table 2). Second, the pseudo R^2 shown for school discipline is not interpretable on the same scale as R^2 for continuous outcomes. Finally, our objectives are explicit model comparison and predictive model performance with future data rather than summarizing the explanatory power of a single model using existing data.

10 Full regression results are available in Appendix C in the online supplement. The models yielding three or more statistically significant coefficients, but higher AIC scores, are shown in Appendix Table C9a (Approach 2a) and Appendix Table C9b (Approaches 3a, 3b, and 3c) in the online supplement.

11 These include women and men's wages; women's salary; men's depression; men's school discipline.

12 The three exceptions are BIC scores for the dichotomous Black screener specification for women's salary and school discipline for men and women.

13 For a summary of best-performing measures by gender and outcome, as evaluated by the K-fold cross-validation RMSE results compared to a constant-only benchmark, see Appendix Tables B6–B10 in the online supplement. Actual RMSEs are reported with the full regression results in Appendix C in the online supplement.

14 The predicted unemployment rate for Black men is 6.53 percent when measured using the 2002 self-identification measure, compared to 4.84 percent when using the screener and 4.86 percent when using the 1979 self-identification measure (Figure 1). The gaps between the 2002 self-identification measure compared to the other measures are about twice as large as the gap between White and Black men's employment rates during the first few months of the coronavirus pandemic (Dias 2021).

15 The sole divergence between the AIC and cross-validation best-performing measures is for men's unemployment. However, the two approaches identified are in the same conceptual "family" (2c and 2d) and, overall, differ very little in terms of fit for this outcome (e.g., on AIC, Approach 2d is just two points worse than 2c, see Appendix Table B3; on RMSE, see Table B8 in the online supplement).

16 Importantly, our dual measure models treat the relationship between observed and self-identified race as additive. In a study dedicated to understanding patterns of school

discipline it would be important to further test whether these relationships are better characterized as interactive or multiplicative, with the magnitude of the association with being seen as Black also varying depending on one's self-identification.

17 The approaches tie for men's unemployment, and either Approach 2a or 3a outperforms the others on women's unemployment and men's depression.

18 The two exceptions are women's and men's unemployment.

19 The exceptions are women's school discipline and men's depression.

20 Indeed, a history of validity theory shows that researchers have moved away from solely criterion-based approaches toward more comprehensive and theory-driven assessments of overall validity over time (see Sireci 1998; Strauss and Smith 2009).

References

Aho, Ken, DeWayne Derryberry, and Teri Peterson. 2014. "Model Selection for Ecologists: The Worldviews of AIC and BIC." *Ecology* 95(3):631–36. https://doi.org/10.1890/13-1452.1

Alvero, A. J., Sonia Giebel, and Francis A. Pearman II. 2024. "Income and Campus Application Disparities among European and Non-European Heritage Hispanic Undergraduate Applicants." *PNAS Nexus* 3:1–4. https://doi.org/10.1093/pnasnexus/pgae337

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bailey, Stanley R., Aliya Saperstein, and Andrew Penner. 2014. "Race, Color, and Income Inequality across the Americas." *Demographic Research* 31:735–56. https://doi.org/10.4054/DemRes.2014.31.24

Blalock, H. M. 1970. "Estimating Measurement Error Using Multiple Indicators and Several Points in Time." *American Sociological Review* 35(1):101–11. https://doi.org/10.2307/2093857

Brown, Tyson H., and Taylor W. Hargrove. 2013. "Multidimensional Approaches to Examining Gender and Racial/Ethnic Stratification in Health." *Women, Gender, and Families of Color* 1(2):180–206. https://doi.org/10.5406/womgenfamcol.1.2.0180

Browne, Irene and Joya Misra. 2003. "The Intersection of Gender and Race in the Labor Market." *Annual Review of Sociology* 29:487–513. https://doi.org/10.1146/annurev.soc.29.010202.100016

Burnham, Kenneth P. and David R. Anderson. 2004. "Multimodel Inference: Understanding AIC and BIC in Model Selection." *Sociological Methods & Research* 33(2):261–304. https://doi.org/10.1177/0049124104268644

Castillo, Wendy and David Gilborn. 2023. "How to 'QuantCrit:' Practices and Questions for Education Data Researchers and Users." *Annenberg Institute at Brown University(EdWorkingPaper: 22-546).* https://doi.org/10.26300/V5KH-DD65

Chakrabarti, Arijit and Jayanta K. Ghosh. 2011. "AIC, BIC and Recent Advances in Model Selection." Pp. 583–605 in *Philosophy of Statistics. Vol. 7, Handbook of the Philosophy of Science*, edited by P. S. Bandyopadhyay and M. R. Forster. Amsterdam: North-Holland.

Chen, Eva Yi-Ju and Eli Yi-Liang Tung. 2024. "Similarities and Differences in the Longitudinal Trajectories of Depressive Symptoms from Mid-Adolescence to Young Adulthood: The Intersectionality of Gender, Race/Ethnicity, and Levels of Depressive Symptoms."

Journal of Racial and Ethnic Health Disparities 11(3):1541–56. https://doi.org/10.1007/s40615-023-01630-5

Cheng, Siwei. 2016. "The Accumulation of (Dis)Advantage: The Intersection of Gender and Race in the Long-Term Wage Effect of Marriage." *American Sociological Review* 81(1):29–56. https://doi.org/10.1177/0003122415621263

Collins, Patricia Hill. 2015. "Intersectionality's Definitional Dilemmas." *Annual Review of Sociology* 41(1):1–20. https://doi.org/10.1146/annurev-soc-073014-112142

Cross, Christina J., Paula Fomby, and Bethany Letiecq. 2022. "Interlinking Structural Racism and Heteropatriarchy: Rethinking Family Structure's Effects on Child Outcomes in a Racialized, Unequal Society." *Journal of Family Theory & Review* 14(3):482–501. https://doi.org/10.1111/jftr.12458

Damian, Elena, Bart Meuleman, and Wim van Oorschot. 2022. "Transparency and Replication in Cross-National Survey Research: Identification of Problems and Possible Solutions." *Sociological Methods & Research* 51(2):499–526. https://doi.org/10.1177/0049124119882452

Dias, Felipe A. 2021. "The Racial Gap in Employment and Layoffs during COVID-19 in the United States: A Visualization." *Socius* 7:1–3. https://doi.org/10.1177/2378023120988397

Dräger, Jascha, Klaus Pforr, and Nora Müller. 2023. "Why Net Worth Misrepresents Wealth Effects and What to Do About It." *Sociological Science* 10:534–58. https://doi.org/10.15195/v10.a19

Engzell, Per and Carina Mood. 2023. "Understanding Patterns and Trends in Income Mobility through Multiverse Analysis." *American Sociological Review* 88(4):600–626. https://doi.org/10.1177/00031224231180607

Ferraro, A. C., Abby Young, Briceira Bernal, Tiffanie Vo, and Cyrus Schleifer. 2025. "Hispanic Intersectional Pay in the United States: Measuring the Intra- and Inter-Ethnoracial and Gender Income Differences by Hispanic Groups." *Sociological Forum* 40(3):319–41. https://doi.org/10.1111/socf.13046

Flanagin, Annette, Miriam Y. Cintron, Stacy L. Christiansen, Tracy Frey, Timothy Gray, Iris Y. Lo, and Roger J. Lewis. 2023. "Comparison of Reporting Race and Ethnicity in Medical Journals Before and After Implementation of Reporting Guidance, 2019-2022." *JAMA Network Open* 6(3):e231706. https://doi.org/10.1001/jamanetworkopen.2023.1706

Forthal, Sarah. 2025. "Appendix C: Multiracial Analysis Schemes: A Mechanism-Driven Approach to Categorizing Multiracial Participants in Biomedical Research." in *Rethinking Race and Ethnicity in Biomedical Research*, *Consensus Study Report*. Washington, D.C.: National Academies Press.

Freese, Jeremy. 2007. "Replication Standards for Quantitative Social Science: Why Not Sociology?" *Sociological Methods & Research* 36(2):153–72. https://doi.org/10.1177/0049124107306659

Garcia, Nichole M., Nancy López, and Verónica N. Vélez. 2018. "QuantCrit: Rectifying Quantitative Methods through Critical Race Theory." *Race Ethnicity and Education* 21(2):149–57. https://doi.org/10.1080/13613324.2017.1377675

Giebel, Sonia. 2023. "'As Diverse as Possible': How Universities Compromise Multiracial Identities." *Sociology of Education* 96(1):1–18. https://doi.org/10.1177/00380407221139180

Gillborn, David, Paul Warmington, and Sean Demack. 2018. "QuantCrit: Education, Policy, 'Big Data' and Principles for a Critical Race Theory of Statistics." *Race Ethnicity and Education* 21(2):158–79. https://doi.org/10.1080/13613324.2017.1377417

Grieco, Elizabeth M. and Rachel C. Cassidy. 2001. *Census 2000 Brief: Overview of Race and Hispanic Origin*. Census Briefs. Accessed March 10, 2025 (https://www2.census.gov/library/publications/decennial/2000/briefs/c2kbr01-01.pdf).

Guluma, Beka and Aliya Saperstein. 2022. "Consistent Divisions or Methodological Decisions? Assessing the U.S. Racial Hierarchy Across Outcomes." *Race and Social Problems 1–19*. https://doi.org/10.1007/s12552-021-09351-2

Hankivsky, Olena. 2012. "Women's Health, Men's Health, and Gender and Health: Implications of Intersectionality." *Social Science & Medicine* 74(11):1712–20. https://doi.org/10.1016/j.socscimed.2011.11.029

Harder, Jenna A. 2020. "The Multiverse of Methods: Extending the Multiverse Analysis to Address Data-Collection Decisions." *Perspectives on Psychological Science* 15(5):1158–77. https://doi.org/10.1177/1745691620917678

Homan, Patricia, Tyson H. Brown, and Brittany King. 2021. "Structural Intersectionality as a New Direction for Health Disparities Research." *Journal of Health and Social Behavior* 62(3):350–70. https://doi.org/10.1177/00221465211032947

Howell, Junia and Michael O. Emerson. 2017. "So What 'Should' We Use? Evaluating the Impact of Five Racial Measures on Markers of Social Inequality." *Sociology of Race and Ethnicity* 3(1):14–30. https://doi.org/10.1177/2332649216648465

Humes, Karen R., Nicholas A. Jones, and Roberto R. Ramirez. 2011. *Overview of Race and Hispanic Origin: 2010*. Census Briefs. U.S. Census Bureau. Accessed November 22, 2024 (https://www.census.gov/content/dam/Census/library/publications/2011/dec/c2010br-02.pdf)

Jiménez, Tomás R., Corey D. Fields, and Ariela Schachter. 2015. "How Ethnoraciality Matters: Looking inside Ethnoracial 'Groups."' *Social Currents* 2(2):107–15. https://doi.org/10.1177/2329496515579765

Kaplan, Judith B. and Trude Bennett. 2003. "Use of Race and Ethnicity in Biomedical Publication." *JAMA* 289(20):2709–16. https://doi.org/10.1001/jama.289.20.2709

Lam-Hine, Tracy, Sarah Forthal, Candice Y. Johnson, and Helen B. Chin. 2024. "Asking MultiCrit Questions: A Reflexive and Critical Framework to Promote Health Data Equity for the Multiracial Population." *The Milbank Quarterly* 102(2):398–428. https://doi.org/10.1111/1468-0009.12696

Leicht, Kevin T. 2008. "Broken Down by Race and Gender? Sociological Explanations of New Sources of Earnings Inequality." *Annual Review of Sociology* 34(1):237–55. https://doi.org/10.1146/annurev.soc.34.040507.134627

Liebler, Carolyn A., Sonya R. Porter, Leticia E. Fernandez, James M. Noon, and Sharon R. Ennis. 2017. "America's Churning Races: Race and Ethnicity Response Changes Between Census 2000 and the 2010 Census." *Demography* 54(1):259–84. https://doi.org/10.1007/s13524-016-0544-0

Light, Audrey and Alita Nandi. 2007. "Identifying Race and Ethnicity in the 1979 National Longitudinal Survey of Youth." *Population Research and Policy Review* 26(2):125–44. https://doi.org/10.1007/s11113-007-9021-1

López, Nancy and Howard Hogan. 2021. "What's Your Street Race? The Urgency of Critical Race Theory and Intersectionality as Lenses for Revising the U.S. Office of Management and Budget Guidelines, Census and Administrative Data in Latinx Communities and Beyond." *Genealogy* 5(3):75. https://doi.org/10.3390/genealogy5030075

Lundberg, Ian. 2024. "The Gap-Closing Estimand: A Causal Approach to Study Interventions That Close Disparities Across Social Categories." *Sociological Methods & Research* 53(2):507–70. https://doi.org/10.1177/00491241211055769

Mahendran, Mayuri, Daniel Lizotte, and Greta R. Bauer. 2022. "Describing Intersectional Health Outcomes: An Evaluation of Data Analysis Methods." *Epidemiology* 33(3):395. https://doi.org/10.1097/EDE.0000000000001466

Martinez, Rae Anne M., Nafeesa Andrabi, Andrea N. Goodwin, Rachel E. Wilbur, Natalie R. Smith, and Paul N. Zivich. 2023. "Conceptualization, Operationalization, and Utilization of Race and Ethnicity in Major Epidemiology Journals, 1995-2018: A Systematic Review." *American Journal of Epidemiology* 192(3):483–96. https://doi.org/10.1093/aje/kwac146

Mauro, Madelyn, Danielle S. Allen, Bege Dauda, Santiago J. Molina, Benjamin M. Neale, and Anna C. F. Lewis. 2022. "A Scoping Review of Guidelines for the Use of Race, Ethnicity, and Ancestry Reveals Widespread Consensus but Also Points of Ongoing Disagreement." *American Journal of Human Genetics* 109(12):2110–25. https://doi.org/10.1016/j.ajhg.2022.11.001

McCall, Leslie. 2005. "The Complexity of Intersectionality." *Signs* 30(3):1771–1800. https://doi.org/10.1086/426800

McDaniel, Anne, Thomas A. DiPrete, Claudia Buchmann, and Uri Shwed. 2011. "The Black Gender Gap in Educational Attainment: Historical Trends and Racial Comparisons." *Demography* 48(3):889–914. https://doi.org/10.1007/s13524-011-0037-0

Monk, Ellis P., Jr. 2016. "The Consequences of 'Race and Color' in Brazil." *Social Problems* 63(3):413–30. https://doi.org/10.1093/socpro/spw014

Moody, James W., Lisa A. Keister, and Maria C. Ramos. 2022. "Reproducibility in the Social Sciences." *Annual Review of Sociology* 48:65–85. https://doi.org/10.1146/annurev-soc-090221-035954

Mora, G. Cristina. 2014. "Cross-Field Effects and Ethnic Classification: The Institutionalization of Hispanic Panethnicity, 1965 to 1990." *American Sociological Review* 79(2):183–210. https://doi.org/10.1177/0003122413509813

Muñoz, John and Cristobal Young. 2018a. "Rejoinder: Can We Weight Models by Their Probability of Being True?" *Sociological Methodology* 48(1):43–51. https://doi.org/10.1177/0081175018796841

Muñoz, John and Cristobal Young. 2018b. "We Ran 9 Billion Regressions: Eliminating False Positives through Computational Model Robustness." *Sociological Methodology* 48(1):1–33. https://doi.org/10.1177/0081175018777988

National Longitudinal Surveys User Services. n.d. "Topical Guide to the Data: Sex." Washington, D.C.: U.S. Bureau of Labor Statistics. Retrieved March 21, 2026. https://nlsinfo.org/content/cohorts/nlsy79/topical-guide/household/sex

O'Brien, Robert M. 2007. "A Caution Regarding Rules of Thumb for Variance Inflation Factors." *Quality & Quantity* 41(5):673–90. https://doi.org/10.1007/s11135-006-9018-6

Pattillo, Mary. 2021. "Black Advantage Vision: Flipping the Script on Racial Inequality Research." *Issues in Race & Society* 10:5–39.

Radloff, Lenore S. 1977. "The CES-D Scale: A Self-Report Depression Scale for Research in the General Population." *Applied Psychological Measurement* 1(3):385–401. https://doi.org/10.1177/014662167700100306

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111–63. https://doi.org/10.2307/271063

Ragin, Charles C., and Peer C. Fiss. 2024. "A Set-Analytic Approach to Intersectionality." *Social Science Research* 120:1–12. doi:10.1016/j.ssresearch.2024.103002

Ramey, David M. 2018. "The Social Construction of Child Social Control via Criminalization and Medicalization: Why Race Matters." *Sociological Forum* 33(1):139–64. https://doi.org/10.1111/socf.12403

Read, Jen'nan G. and Fatima G. Fairfax. 2025. "Hidden Heterogeneity: How the White Racial Category Masks Interethnic Health Inequality." *Demography* 62(1):237–61. https://doi.org/10.1215/00703370-11790429

Ritter, Joseph A. and Lowell J. Taylor. 2011. "Racial Disparity in Unemployment." *The Review of Economics and Statistics* 93(1):30–42. https://doi.org/10.1162/REST_a_00063

Ross, Catherine E. and John Mirowsky. 1989. "Explaining the Social Patterns of Depression: Control and Problem Solving–or Support and Talking?" *Journal of Health and Social Behavior* 30(2):206–19. https://doi.org/10.2307/2137014

Roth, Wendy D. 2016. "The Multiple Dimensions of Race." *Ethnic and Racial Studies* 39(8):1310–38. https://doi.org/10.1080/01419870.2016.1140793

Saperstein, Aliya. 2006. "Double-Checking the Race Box: Examining Inconsistency between Survey Measures of Observed and Self-Reported Race." *Social Forces* 85(1):57–74. https://doi.org/10.1353/sof.2006.0141

Saperstein, Aliya. 2012. "Capturing Complexity in the United States: Which Aspects of Race Matter and When?" *Ethnic and Racial Studies* 35(8):1484–1502. https://doi.org/10.1080/01419870.2011.607504

Saperstein, Aliya, Jessica M. Kizer, and Andrew M. Penner. 2016. "Making the Most of Multiple Measures: Disentangling the Effects of Different Dimensions of Race in Survey Research." *American Behavioral Scientist* 60(4):519–37. https://doi.org/10.1177/0002764215613399

Saperstein, Aliya, and Andrew M. Penner. 2012. "Racial Fluidity and Inequality in the United States." *American Journal of Sociology* 118(3):676–727. https://doi.org/10.1086/667722

Schultz, Michael. 2018. "The Problem of Underdetermination in Model Selection." *Sociological Methodology* 48(1):52–87. https://doi.org/10.1177/0081175018786762

Schwabish, Jonathan, and Alice Feng. 2021. Combining Racial Groups in Data Analysis Can Mask Important Differences in Communities. Urban Institute. Accessed September 9, 2025. (https://www.urban.org/urban-wire/combining-racial-groups-data-analysis-can-mask-important-differences-communities).

Scott, Nicholas A. and Janet Siltanen. 2017. "Intersectionality and Quantitative Methods: Assessing Regression from a Feminist Perspective." *International Journal of Social Research Methodology* 20(4):14. https://doi.org/10.1080/13645579.2016.1201328

Sen, Maya and Omar Wasow. 2016. "Race as a Bundle of Sticks: Designs That Estimate Effects of Seemingly Immutable Characteristics." *Annual Review of Political Science* 19(1):499–522. https://doi.org/10.1146/annurev-polisci-032015-010015

Shiao, Jiannbin Lee. 2019. "When (In)Consistency Matters: Racial Identification and Specification." *Socius: Sociological Research for a Dynamic World* 5:1–18. https://doi.org/10.1177/2378023119848268

Shiao, Jiannbin Lee. 2023. "Measuring Hispanics/Latinxs: Racial Heterogeneity and Its Consequences for Modeling Social Outcomes in U.S. Population Samples." *Socius* 9:23780231231174830. https://doi.org/10.1177/23780231231174830

Short, Cassie, Nate Breznau, Maria Bruntsch, Micha Burkhardt, Niko Busch, Elena Cesnaite, Maximilian Frank, Carsten Gießing, Daniel Krähmer, Daniel Kristanto, Tina Lonsdorf, Claudia Neuendorf, Hung Nguyen, Manuel Rausch, Xenia Schmalz, Andreas Schneck, Cem Tabakci, and Andrea Hildebrandt. 2025. "Multi-Curious: A Multi-Disciplinary Guide to Multiverse Analysis." https://doi.org/10.31222/osf.io/4yzeh_v2

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting

Anything as Significant." *Psychological Science* 22(11):1359–66. https://doi.org/10.1177/0956797611417632

Sireci, Stephen G. 1998. "The Construct of Content Validity." *Social Indicators Research* 45(1):83–117. https://doi.org/10.1023/A:1006985528729

Sosina, Victoria E. and Aliya Saperstein. 2022. "Reflecting Race and Status: The Dynamics of Material Hardship and How People Are Perceived." *Socius* 8:23780231221124578. https://doi.org/10.1177/23780231221124578

Spierings, Niels. 2023. "Quantitative Intersectional Research: Approaches, Practices, and Needs." Pp. 235–48 in *The Routledge International Handbook of Intersectionality Studies*, edited by K. Davis and H. Lutz. London: Routledge. https://doi.org/10.4324/9781003089520-22

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing Transparency Through a Multiverse Analysis." *Perspectives on Psychological Science* 11(5):702–12. https://doi.org/10.1177/1745691616658637

Stepanikova, Irena and Gabriela R. Oates. 2016. "Dimensions of Racial Identity and Perceived Discrimination in Health Care." *Ethnicity & Disease* 26(4):501–12. https://doi.org/10.18865/ed.26.4.501

Strauss, Milton E. and Gregory T. Smith. 2009. "Construct Validity: Advances in Theory and Methodology." *Annual Review of Clinical Psychology* 5:1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Telles, Edward E. and Nelson Lim. 1998. "Does It Matter Who Answers the Race Question? Racial Classification and Income Inequality in Brazil." *Demography* 35(4):465–74. https://doi.org/10.2307/3004014

U.S. Bureau of Labor Statistics. n.d.-a. "National Longitudinal Surveys Annotated Bibliography." Retrieved November 14, 2024. https://www.nlsinfo.org/bibliography-start

U.S. Bureau of Labor Statistics. n.d.-b. "Topical Guide: Race, Ethnicity & Immigration Data." Retrieved October 3, 2022. https://nlsinfo.org/content/cohorts/nlsy79/topical-guide/household/race-ethnicity-immigration-data

Vargas, Nicholas and Jared Kingsbury. 2016. "Racial Identity Contestation: Mapping and Measuring Racial Boundaries." *Sociology Compass* 10(8):718–29. https://doi.org/10.1111/soc4.12395

Verhagen, Mark D. 2022. "A Pragmatist's Guide to Using Prediction in the Social Sciences." *Socius* 8:1–17. https://doi.org/10.1177/23780231221081702

Villarreal, Andrés and Stanley R. Bailey. 2020. "The Endogeneity of Race: Black Racial Identification and Men's Earnings in Mexico." *Social Forces* 98(4):1744–72. https://doi.org/10.1093/sf/soz096

Vincent, Luna. 2026. "Specifying Race: The Colonial Constitution of Race in a Set-Theoretic Framework." *Sociological Theory* 44(1):1–24. https://doi.org/10.1177/07352751251395706

Vrieze, Scott I. 2012. "Model Selection and Psychological Theory: A Discussion of the Differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)." *Psychological Methods* 17(2):228–43. https://doi.org/10.1037/a0027127

Weisshaar, Katherine and Tania Cabello-Hutt. 2020. "Labor Force Participation Over the Life Course: The Long-Term Effects of Employment Trajectories on Wages and the Gendered Payoff to Employment." *Demography* 57(1):33–60. https://doi.org/10.1007/s13524-019-00845-8

Westbrook, Laurel, and Aliya Saperstein. 2015. "New Categories Are Not Enough: Rethinking the Measurement of Sex and Gender in Social Surveys." *Gender & Society* 29(4):534–60. https://doi.org/10.1177/0891243215584758

Wilkes, Rima and Aryan Karimi. 2024. "From Intersectional Invisibility to Visibility: Black Women in Health Disparity Data and Quantitative Intersectional Models." *Current Research in Behavioral Sciences* 6:100149. https://doi.org/10.1016/j.crbeha.2024.100149

Wilson, M. Roy, Sarah H. Beachy, and Samantha N. Schumm, eds. 2025. *Rethinking Race and Ethnicity in Biomedical Research*. Washington, D.C.: National Academies Press. https://doi.org/10.17226/27913

Wong, Jaclyn S., Lauren Valentino, Christina Pao, Katie Donnelly Moran, D'Lane Compton, and Gayle Kaufman. 2024. "What to Do with 'Other, Describe.'" *Sociological Methodology* 55(2):244–68. https://doi.org/10.1177/00811750241304774

Young, Cristobal and Erin Cumberworth. 2025. *Multiverse Analysis: Computational Methods for Robust Results*. Analytical Methods for Social Research. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009003391

Zuberi, Tukufu. 2001. *Thicker than Blood: How Racial Statistics Lie*. Minneapolis: University of Minnesota Press.

Acknowledgments: We are grateful to our colleagues in the gender and inequality workshops at Stanford University for their helpful comments and suggestions, and to Steve McClaskie for responding to inquiries about the NLSY. Previous versions of this paper were presented at the 2024 American Sociological Association annual meeting and at a 2023 conference on racial inequality in education research hosted by NWEA in Portland, OR. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1656518. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Emma Williams-Baron: Department of Sociology, Stanford University.
E-mail: emmajwb@stanford.edu.

Aliya Saperstein: Department of Sociology, Stanford University.
E-mail: asaper@stanford.edu.