Which Effect of Race? Causal Inference without Holding All Else Equal

Thomas Leavitt

Abstract

Empirical studies of racial discrimination vary race while holding nonracial traits fixed, a design the literature defends as what credible inference requires. This defense bundles two claims: which effect of race a study should target, and whether its design can recover that effect. I separate them. The same randomization that secures credible estimation and inference recovers a family of race estimands, from the all-else-equal effect to a within-race effect that lets associated traits vary with race. Every member of that family is causal rather than descriptive, and the choice among members is a claim about what a racial category is — a claim extending to ethnicity, religion, and other identity categories that index associated traits. I derive conditions, weaker than the literature's, for credible estimation and inference. Reanalyzing a Spanish-language campaign experiment, I find coethnic preference among Hispanic voters under the within-race effect and none under the all-else-equal effect.

1 Introduction

Whether people face discrimination on the basis of race — as well as on other bases such as religion and ethnolinguistic identity — is among the oldest and most contested questions in political life. Consider the long-running debate over discrimination against Muslims in France: Would a French Muslim receive a colder response than a French Christian to a job application or a request to a public office? As it figures in public debate, the comparison sets a Muslim of Maghrebi origin — along with associated traits such as Arabic language and others that characterize a "real" Muslim in popular French perception (Adida et al., 2010; Diop, 1988) — against a Christian of European origin who typically bears none of these traits.

The standard approach is to pull these features apart, treating geographic origin and its associated traits as confounders of what Adida et al. (2010, p. 22386) call the "Muslim effect." Their correspondence study in the French labor market holds the applicant's Senegalese origin fixed while varying the religion the résumé signals, and finds that fewer employers would respond positively to the ostensibly Muslim applicant. Yet isolating religion in this way excludes the groups at the center of French public debate — Maghrebi-origin Muslims and European-origin Christians — putting in their place a Senegalese Muslim and a Senegalese Christian, each atypical of its category as it figures in that debate. Which effect of religion, then, is the one to study: the effect of religious affiliation separable from origin, language, and other traits, or the effect of the category "Muslim" with the traits the category indexes?

The same all-else-equal move — erasing rather than preserving differences in attributes across categories — underwrites much of what we know about the effects of race and ethnicity. As Sen and Wasow (2016, p. 512) observe, "*all* studies employing exposure to a racial or ethnic signal share a common experimental design," one that "present[s] a subject with information that differs only with respect to signals or cues about race or ethnicity." This design ranges from the canonical audit and correspondence studies of hiring (Bertrand

and Mullainathan, 2004; Pager, 2003; Quillian et al., 2017) to the conjoint experiments now common in political science (Hainmueller and Hopkins, 2015). Because the literature defends the all-else-equal design as what credible inference requires (see, e.g., Butler and Crabtree, 2021; Heckman, 1998), the choice of estimand embedded in the design is rarely recognized as a choice at all.

This defense of the all-else-equal design informs a practice that genuinely removes confounding but, in the same act, silently fixes *which* estimand a study recovers (Section 6.1). The literature's recovery conditions thus bundle two claims this paper separates: an estimand commitment, which concerns the contrast a study should target, and a recovery condition, which concerns whether a design recovers that contrast. Once the two claims are separated, the all-else-equal contrast defines only one member of a broader family of counterfactual race effects — including members that preserve the race-conditional structure of nonracial attributes. A single randomization recovers every member of that family of estimands. The choice among that family's members is then substantive rather than statistical: Differences in nonracial attributes "may not be *threats to inference*, but, rather, an important part of the *story of racial discrimination*" (Crabtree et al., 2023, p. 2, emphasis added). What has been lacking is a framework that makes these claims precise.

This paper supplies one, consisting of three parts. First, I derive a family of precisely defined race estimands. Second, I develop substantive grounds for choosing among the estimands. Third, I provide conditions, weaker than the literature assumes, under which any member of the estimand family can be recovered with unbiased or consistent estimation and valid inference.

The framework applies in either of the two dominant regimes for studying discrimination, *exposure-based* and *perception-based* (Hu and Kohler-Hausmann, 2025, pp. 259–260). In the exposure-based regime, race is a fixed feature of what the decision-maker encounters — a person or a racial signal — regardless of what the decision-maker perceives (e.g., Piper, 1992). For example, an officer may encounter a Black or a White civilian whose race remains fixed even if the officer cannot perceive (or misperceives) the civilian's race,

as under a veil of darkness (Grogger and Ridgeway, 2006; Pierson et al., 2020). Likewise, an employer may encounter a résumé whose name the researcher conceives as a Black or a White signal, whether or not the signal registers with the employer. In the perception-based regime, race is instead the racial interpretation the decision-maker forms, so the regime requires a cue to induce the interpretation — the résumé's name, say — and the racial variable is the cue's uptake, not the cue.

The family of race estimands is bounded by two extremes, with mixtures between them. At one extreme, the All-Else-Equal Effect (AEE) — of which the average marginal component effect (AMCE) from conjoint experiments (Hainmueller et al., 2014) is a special case — forces the distribution of nonracial attributes to be identical across racial conditions. At the other, the All-Else-Within-Race Effect (AEWR) lets that distribution differ across racial conditions in ways that reflect the social structure under study. The mixtures hold some attributes fixed and let others vary by race. Every member of the estimand family is built from the same unit-level race effect — a counterfactual contrast for a fixed decision-maker, not a difference across people (Section 3) — and the members differ only in which features are held fixed across racial conditions and how the remaining features are weighted, specifications the researcher sets under the exposure-based regime. No member of the estimand family is privileged independently of the substantive question.

Under the perception-based regime, the cue's uptake, rather than the researcher, determines which nonracial perceptions stay fixed and which move with perceived race: A cue that would shift a decision-maker's perceived race may shift nonracial perceptions with it, so the effect the data recover is whichever member of the estimand family those joint movements compose. Insofar as moving those perceptions together is partly what it means for a cue to be racialized, as Hu and Kohler-Hausmann (2025) argue, the member the joint movements compose is not a failed manipulation but the effect of interest — the perception-based counterpart of the within-race effect. Although much research seeks cues that shift perceived race and nothing else (Dafoe et al., 2018; Elder and Hayes, 2023; Landgrave and Weller, 2022), this paper shows that an isolated shift in perceived race need be neither a

condition for credible inference nor a privileged target.

The framework as a whole enters a debate at the center of the empirical literature on discrimination: whether, and how, a counterfactual effect of race can be defined and recovered. Two prominent accounts, one explicitly and one by presupposition, proceed from a constructivist conception of race on which the category consists in nominal membership together with the nonracial attributes the category indexes. Both hold one further premise: A counterfactual effect of race must hold all else equal. Granting the premise, one response studies one racial cue at a time and sets aside the effect of the category as constituted (e.g., Sen and Wasow, 2016); the other retains the category and objects that an all-else-equal contrast, by holding the indexed attributes equal, cannot be an effect of the category as constituted (Hu, 2023; Hu and Kohler-Hausmann, 2025), an objection whose strongest form dispenses with the potential outcomes framework for detecting discrimination (Kohler-Hausmann, 2019). This paper's contribution to that debate is to show that the premise is a choice rather than a requirement of credible inference (Section 2): Because membership in a racial category is fixed independently of the attributes the category indexes, contrasts that hold nothing else equal are well-defined effects of race. One can therefore keep the category as constituted without abandoning causal inference, and keep causal inference without reducing the category to a cue.

I illustrate the framework by reanalyzing a candidate-evaluation experiment (Zárate et al., 2024) under both regimes, asking whether Hispanic voters reward a coethnic candidate — a form of discrimination that favors rather than penalizes. Read through the exposure-based regime, the all-else-equal effect is indistinguishable from zero, while the within-race effect is positive and distinguishable from zero. On one experiment, then, the choice of estimand is the difference between finding coethnic preference and finding none, a divergence that reflects the different questions the two estimands answer rather than any artifact of estimation. Read through the perception-based regime, which member of the estimand family the data recover is no longer the researcher's to choose; auxiliary name-rating data (Elder and Hayes, 2023) suggest the recovered effect is unlikely to be

the all-else-equal member, though the recovered effect remains a well-defined estimand of discrimination either way.

In Section 9, I close with three questions that emerge once the choice of estimand is recognized as a choice: (i) how normative theories of discrimination might name the contrast a study should target; (ii) how a study can report across the family of estimands or instead adopt and defend a single member; and (iii) what it means to represent a racial group when some individuals belong to both that group and another defined by, say, income, so that, unlike race and income on a résumé, which a study can pair freely, the groups' preferences cannot move independently. Proofs and additional formal material appear in the Supplementary Material.

2 Must an Effect of Race Hold All Else Equal?

Following arguments that an individual's race cannot be manipulated (DiNardo, 2007; Holland, 1986, 2001, 2003; Zuberi, 2001), studies of discrimination define the effect of race not on an individual conceived as someone who could be, say, either Black or White, but on a decision-maker who could encounter either an individual racialized as Black or a different individual racialized as White (following Greiner and Rubin, 2011). The effect is the difference in how the decision-maker would respond to each individual, whether the encounter is direct, through a résumé, or through another representation of the individual.

Defining the effect through a decision-maker's responses to two differently racialized individuals sidesteps the manipulation of race but still depends on a conception of race itself: who counts as Black or as White. Among the many conceptions of race and ethnicity, nearly all agree that the categories rest on membership classifications justified in terms of *descent*, whether grounded in biological essence or social convention (Andreasen, 1998, 2000; Appiah and Gutmann, 1998; Chandra, 2006, 2012; Cornell and Hartmann, 1997; Fearon, 2003; Hardimon, 2003, 2017; Haslanger, 2000; Mills, 1998; Spencer, 2014; Taylor, 2003; Weber, 1978).

On a classical conception of categories, membership exhausts what a category is (Monk,

Jr., 2022, pp. 9–10): A racial category consists only in nominal membership as the descent-based classifications assign it, so a comparison of a Black and a White individual identical in nonracial features can omit nothing the category is. By contrast, on what Hu (2023) calls a "thick constructivist" conception of race, a racial category consists in nominal membership together with the nonracial attributes that membership probabilistically indexes — the category's social content (Haslanger, 2000; Taylor, 2003). Race, so conceived, is not reducible to nominal membership alone.

Insofar as a racial category indexes a distribution of nonracial attributes, some configurations are more typical of the category than others (Rosch, 1973, 1975; Rosch and Mervis, 1975; Wittgenstein, 1953). Typicality increases toward a prototype at the center of the category-conditional distribution — a graded conception of race already present in empirical work (Eberhardt et al., 2006; Ma et al., 2015; Maddox, 2004; Ocampo-Roland, 2025). To call a configuration typical of the category is to describe the social world — the distribution of nonracial attributes within the category, or the expectations society attaches to the category's members — not to judge how authentically Black or White anyone is (Appiah, 2005; Appiah and Gutmann, 1998; Gooding-Williams, 1998).

Two prominent recent accounts of race as a cause proceed from such a constructivist conception, one explicitly and one by presupposition. Kohler-Hausmann (2019) embraces the conception explicitly, as do Hu (2023, p. 12), for whom the marking of racial categories runs through "phenotype and supposed ancestry," and Hu and Kohler-Hausmann (2025, p. 250), for whom racial and nonracial features "co-constitute the category of race." Sen and Wasow (2016) presuppose the conception: Their designs exploit within-group variation, and a single "stick" serves as a proxy that "conveys information about race" (Sen and Wasow, 2016, p. 509) — formulations that require a racial category distinct from, and fixed independently of, any one cue. On either account, the descent that grounds membership — actual or presumed — can stay fixed when neighborhood or record moves, so who counts as Black or White can hold still while the attributes vary. In the image of race as a bundle of sticks (Sen and Wasow, 2016), the attributes are the sticks, and nominal membership is

the string that ties them into a bundle (Section 5).

Yet from this shared conception a division emerges over what the two racialized individuals must differ in, and what they may share, for the contrast between them to count as an effect of race. The division's two sides move in opposite directions, yet rest on a single premise: A causal effect in the potential outcomes framework must be an all-else-equal contrast. Grant the premise, and the category as constituted admits no effect defined in terms of potential outcomes: An effect of the category would let the indexed attributes move with membership, and that movement is precisely not holding all else equal. As Kohler-Hausmann (2019, p. 1171) concludes, "race cannot be conceptualized as an isolated treatment in the counterfactual causal model, and accordingly, racial discrimination cannot be defined as the treatment effect of race."

One response — the cue-effects response — retains potential outcomes and studies one cue at a time, holding the remaining attributes fixed, while setting aside the effect of the category as constituted (Sen and Wasow, 2016). The other — the incompatibility response — retains the category as constituted and denies that a contrast holding the indexed attributes equal — the all-else-equal contrast — can be an effect of race (Hu, 2023; Hu and Kohler-Hausmann, 2025; Kohler-Hausmann, 2019; Kohler-Hausmann and Dembroff, 2022); Dembroff et al. (2020) make the same denial for sex. The denial has resulted in two conclusions. One dispenses with the framework for detecting discrimination on the grounds that "[t]he counterfactual causal model of discrimination is based on a flawed theory of what the category of race references" (Kohler-Hausmann, 2019, p. 1163), a conclusion Dembroff et al. (2020) also reach. The other holds that positive inquiry into the effects of race cannot be separated from normative inquiry, a claim the accounts advance with varying scope (Dembroff et al., 2020; Hu, 2026; Hu and Kohler-Hausmann, 2025; Kohler-Hausmann and Dembroff, 2022).

I challenge the premise that the cue-effects response and the incompatibility response share. As the sections below show, a contrast in which the nonracial attributes move with racial membership across the two individuals is a well-defined effect in the potential

outcomes framework. Hu and Kohler-Hausmann (2025, p. 253) come closest to relaxing the premise, expressing one such contrast within potential outcomes, though as the estimand the constructivist conception calls for rather than as one member of a family of estimands the researcher can choose among. Once the premise falls, the two responses are no longer exhaustive alternatives: A study can infer cue effects in the sense of the cue-effects response and effects of the category as constituted in the sense of the incompatibility response. Which effect the study targets becomes a substantive choice rather than a consequence of the premise.

3 Formal Setup

I now formalize the common structure of studies of discrimination and, within it, the two regimes introduced above. I take the exposure-based regime as the baseline, building through it the shared apparatus — the configuration space, the potential-outcome function on it, and the family of race effects; for that reason the exposure-based regime has no subsection of its own here. The perception-based regime, treated separately (Section 3.3) because it recasts the configuration as itself a potential outcome of a cue, borrows that apparatus intact. Later sections, which treat the two regimes in parallel, give each regime its own heading.

I develop the apparatus on a running example of prosecutorial charging — a setting the causal inference literature itself studies (Gaebler et al., 2022) and the case its critics treat as paradigmatic (Hu and Kohler-Hausmann, 2025). A prosecutor reads an arrest file and decides whether to file charges. The file records the arrestee's race — Black or White — or only cues of race, such as the arrestee's name, which may induce the prosecutor's perception of the arrestee as one or the other. Alongside the stated race or the cue, the file contains two nonracial features: the location of the arrest and the extent of the arrestee's prior record (Hu and Kohler-Hausmann, 2025, pp. 256–257). Neither feature is neutral background: Black, relative to White, indexes a greater likelihood of residing in — and so of being arrested in — neighborhoods like Brownsville, Brooklyn, a low-income, largely

Black neighborhood (Hu and Kohler-Hausmann, 2025, p. 256), and a greater likelihood of an extensive prior record, if only because such neighborhoods are subject to greater policing.

The example thus confronts the researcher with a choice. Holding location fixed compares a Black and a White arrestee both arrested in, say, Brownsville — a shared location that makes "the Black arrestee 'expected'" and "the white arrestee 'unexpected'" (Hu and Kohler-Hausmann, 2025, p. 257). Yet an "unexpected" White arrestee is still White. Who counts as White holds fixed while an attribute the category indexes varies. Letting location vary instead compares the arrestees at locations more likely within each arrestee's category: say, a Black arrestee in Brownsville and a White arrestee on the Upper East Side, a high-income, largely White neighborhood. An analogous choice arises for the prior record.

3.1 Decision-Makers, Decisions, and Configurations

There is a set of decision-makers, prosecutors in the running example, and each faces one or more decisions; indexing the resulting decision-maker–decision pairs as *units* $i \in \{1, \ldots, N\}$, a unit is one prosecutor reading one case file (Section S1.1 gives the explicit indexing by decision-makers and decisions). The pair, rather than the prosecutor, is the unit because the same prosecutor need not respond alike across decisions (Section 6.1). Under the exposure-based regime, unit i encounters a *configuration* $\boldsymbol{t}_i \coloneqq (z, \boldsymbol{v})$: the case file, formally. The feature $z \in \{0, 1\}$ is a *racially constituted attribute* — a feature specified in terms of race, such as the race the file states or a name conceived as a racial signal. The remaining *nonracial features* $\boldsymbol{v} \coloneqq (v_1, \ldots, v_L)$, such as the prior record and the neighborhood, take values in $\mathcal{V} \coloneqq \mathcal{V}_1 \times \cdots \times \mathcal{V}_L$, where each $\mathcal{V}_\ell$ is finite, and their joint distribution with z may itself be part of how race manifests in the social world. Each configuration $\boldsymbol{t}_i$ thus lies in the common space $\mathcal{T} \coloneqq \{0, 1\} \times \mathcal{V}$.

The set $\mathcal{T}$ is a product space — either racial value can be paired with any nonracial profile — and this pairing lets every member of the estimand family be defined. What licenses the pairing is the distinction between nominal membership and the attributes that

membership indexes: For the racial categories this paper treats, membership runs through descent, actual or presumed, a criterion outside the nonracial profile (Section 2). A file that pairs one racial category with a profile typical of the other is therefore atypical, not ill-defined. Which nonracial features compose $\boldsymbol{v}$ is a substantive choice the researcher makes in defining the estimands, not a claim about measurement. The framework takes the set as given and does not derive it.

3.2 Potential Outcomes and the Race Effect

In principle, a unit's potential outcome could depend on the configurations that other units encounter — including the other files the same prosecutor reads. A stable unit treatment value assumption (SUTVA), stated formally in Section S1.1, rules out such interference and carryover: The potential outcome for unit i depends only on unit i's own configuration, defining a function $y_i : \mathcal{T} \to \mathbb{R}$, where $\mathbb{R}$ is the set of real numbers, with $y_i(\boldsymbol{t}_i) = y_i(z, \boldsymbol{v})$ for $\boldsymbol{t}_i = (z, \boldsymbol{v}) \in \mathcal{T}$. In the running example, the function y_i is the charging response unit i's prosecutor would give to each possible file.

A unit-level effect is a contrast between two configurations:

$$\tau_i(\boldsymbol{t}, \boldsymbol{t}') \coloneqq y_i(\boldsymbol{t}) - y_i(\boldsymbol{t}'), \qquad \boldsymbol{t}, \boldsymbol{t}' \in \mathcal{T}.$$

For race effects, the two configurations differ in the value of z. Writing $\boldsymbol{t} = (1, \boldsymbol{v})$ and $\boldsymbol{t}' = (0, \boldsymbol{v}')$, the *race effect* is

$$\tau_i\left((1, \boldsymbol{v}), (0, \boldsymbol{v}')\right) = y_i(1, \boldsymbol{v}) - y_i(0, \boldsymbol{v}'), \qquad \boldsymbol{v}, \boldsymbol{v}' \in \mathcal{V}. \tag{1}$$

Because $\boldsymbol{v}$ and $\boldsymbol{v}'$ each range over $\mathcal{V}$, the quantity in (1) defines a family of effects, one for each pair $(\boldsymbol{v}, \boldsymbol{v}')$ of nonracial profiles. The all-else-equal contrast is the special case $\boldsymbol{v} = \boldsymbol{v}'$; the opposite extreme leaves $\boldsymbol{v}$ and $\boldsymbol{v}'$ free to vary across racial conditions. In the running example, setting $\boldsymbol{v} = \boldsymbol{v}'$ compares the Black and the White arrestee at one shared prior

record and neighborhood, while setting $\boldsymbol{v} \neq \boldsymbol{v}'$ lets the two files differ not only in race but also in prior record and neighborhood.

3.3 Perception-Based Regime

Return once more to the running example, and take the file in its other form, recording race only through the arrestee's name — DeShawn or Connor. Suppose the file contains nothing else, so that neighborhood and prior record, like race, are left for the prosecutor to infer. The arrestee's own race is fixed however anyone reads the file, but what now drives the decision is the race the prosecutor perceives, which need not match the arrestee's own, and the same cue may move nonracial perceptions too; for some files the name would move perception not at all. The perception-based regime formalizes this reading, treating the perceived race and features as potential outcomes of the name and confining the race contrast to the files whose perceived race the name would actually shift.

Formally, the manipulation is a *cue* $w_i \in \{0, 1\}$ — a name (Bertrand and Mullainathan, 2004), photo (Terkildsen, 1993), accent (Purnell et al., 1999), or other stimulus — not the configuration itself, with $w_i = 1$ intended to induce $z = 1$ and $w_i = 0$ intended to induce $z = 0$. The configuration is the perception the cue actually would induce, not necessarily the perception the cue is intended to induce. The racial feature, the nonracial features, and the response are therefore all potential outcomes of the cue.

As in the exposure-based regime, a cue-level SUTVA — stated formally in Section S1.1 — rules out interference, so unit i's response, racial perception, and nonracial perceptions depend only on unit i's own cue. Write $z_i(w)$ and $\boldsymbol{v}_i(w)$ for the racial and nonracial components of the perception that cue value w would induce for unit i, and define the cue-induced configuration

$$\boldsymbol{t}_i(w) \coloneqq (z_i(w), \boldsymbol{v}_i(w)) \in \mathcal{T}.$$

The component functions z_i and $\boldsymbol{v}_i$ take values in the same spaces as in the exposure-

based regime, but here the components are unit-level potential outcomes of the cue. As under the exposure-based regime, the researcher specifies which perceptions compose $\boldsymbol{v}$; the cue's uptake, not the researcher, determines the values $z_i(w)$ and $\boldsymbol{v}_i(w)$. The configuration realized for unit i is $\boldsymbol{t}_i(w_i)$.

The potential response to a cue may in principle depend on the cue through channels other than the cue-induced perception — through, say, an aesthetic or affective reaction to the cue that operates whatever race and nonracial features the decision-maker would read from the cue. To rule out such direct channels — channels that would make part of the cue effect an effect of the stimulus itself, a part no effect of perceived race in (1) could represent — I assume the potential outcome depends on the cue only through that unit's perceived configuration, an assumption that holds only if every channel through which the cue moves the outcome runs through the perceived race or the nonracial perceptions $\boldsymbol{v}$.

Assumption 1 (Perception sufficiency)**.** *For all units* $i \in \{1, \ldots, N\}$ *and all* $w, w' \in \{0, 1\}$*,*

$$\boldsymbol{t}_i(w) = \boldsymbol{t}_i(w') \implies y_i(w) = y_i(w').$$

Together, the cue-level SUTVA and perception sufficiency (Assumption 1) do for the perception-based regime what SUTVA alone did for the exposure-based one (Section 3.2): The two assumptions reduce the potential responses to the cue to the same function y_i, evaluated at the cue-induced configuration,

$$y_i(w) = y_i\left(\boldsymbol{t}_i(w)\right) = y_i\left(z_i(w), \boldsymbol{v}_i(w)\right), \qquad w \in \{0, 1\}.$$

The unit-level cue effect is thus the contrast between the two configurations induced by the alternative cue values:

$$\tau_i\left(\boldsymbol{t}_i(1), \boldsymbol{t}_i(0)\right) = y_i\left(z_i(1), \boldsymbol{v}_i(1)\right) - y_i\left(z_i(0), \boldsymbol{v}_i(0)\right). \tag{2}$$

Under the cue-level SUTVA and Assumption 1, whether the cue effect in (2) corresponds to a race effect in (1) depends on how the cue changes perceived race. Following Frangakis

and Rubin (2002), classify units into principal strata by $(z_i(0), z_i(1))$: *compliers* $(0,1)$, *always-takers* $(1,1)$, *never-takers* $(0,0)$, and *defiers* $(1,0)$. For always-takers and never-takers, the cue would induce no race contrast because z would not change across cue values. For defiers, the cue would move z opposite the intended direction, so the cue effect equals the negative of a race effect in (1), evaluated at the perceptions the cue would induce. The intended directions rest on shared conventions of racial signification — the coding that makes DeShawn read as Black and Connor as White — so a defier would read the coding in reverse, against the conventions the cue's design presupposes. For compliers, by contrast, the cue would induce the race contrast in the intended direction.

Let $\mathcal{C} \coloneqq \{i : z_i(0) = 0, z_i(1) = 1\}$ denote the set of compliers. On $\mathcal{C}$, the cue effect equals the race effect at the configurations $\boldsymbol{t}_i(0) = (0, \boldsymbol{v}_i(0))$ and $\boldsymbol{t}_i(1) = (1, \boldsymbol{v}_i(1))$ induced by the cue:

$$\tau_i\left(\boldsymbol{t}_i(1), \boldsymbol{t}_i(0)\right) = y_i\left(1, \boldsymbol{v}_i(1)\right) - y_i\left(0, \boldsymbol{v}_i(0)\right), \qquad i \in \mathcal{C}, \tag{3}$$

where $\boldsymbol{v}_i(0), \boldsymbol{v}_i(1) \in \mathcal{V}$ are the nonracial perceptions that would be induced under each cue value. Unlike in (1), where the researcher can in principle set the values of $\boldsymbol{v}, \boldsymbol{v}' \in \mathcal{V}$, the perception-based regime allows the researcher to set only the value of the cue: Perceived race moves from 0 to 1 by the definition of $\mathcal{C}$, while the nonracial perceptions $\boldsymbol{v}_i(0)$ and $\boldsymbol{v}_i(1)$ are governed by the decision-maker's potential responses to that cue.

4 Anatomy of an Estimand

An *estimand* collapses the family of unit-level race effects in (1) into a scalar through three components: A *contrast* holds certain nonracial features fixed across racial conditions and lets the others vary by race; a *weighting* on $\mathcal{V}$, either common across racial conditions or race-specific, weights the nonracial profiles; and a *subset* restricts which units enter the average. Each such estimand is therefore a linear functional of the potential outcomes — a difference-scale weighted average of unit-level race effects.

The weighting is either a distribution ρ on $\mathcal{V}$ common across racial conditions — which

weights nonracial features without regard to race — or race-specific distributions q_z on $\mathcal{V}$ — which weight nonracial features conditional on each racial condition $z \in \{0, 1\}$. Both can be derived from a single joint PMF on $\mathcal{T}$, $q(z, \boldsymbol{v}) \coloneqq \Pr((Z, \boldsymbol{V}) = (z, \boldsymbol{v}))$, via the marginal $\rho(\boldsymbol{v}) \coloneqq \sum_z q(z, \boldsymbol{v})$ and, where the denominator is positive, the conditional $q_z(\boldsymbol{v}) \coloneqq q(z, \boldsymbol{v}) / \sum_{\boldsymbol{v}' \in \mathcal{V}} q(z, \boldsymbol{v}')$. These distributions are part of the estimand's definition — quantities the researcher specifies, not a model of how the data were generated or an empirical distribution measured in the study. Because the researcher places these weights, a distribution may concentrate on any subset of $\mathcal{V}$, and the researcher can put little or no weight on a profile of no substantive interest. I take the weighting to be common across units (not to be confused with common across racial conditions); unit-specific distributions would require only added notation.

4.1 The Estimand: Contrast, Weighting, and Subset

4.1.1 Exposure-Based Regime

Specifying an estimand begins with the contrast. Contrasts can be ordered by how much of $\boldsymbol{v}$ they hold fixed across racial conditions, running between two extremes. The *all-else-equal* contrast holds every feature fixed and the *within-race* contrast holds none; *mixed* contrasts, holding some features fixed while leaving others free to vary by race, lie in between. Across all three contrasts, I average over all N units, though the framework allows narrower subsets.

All-else-equal (AEE) effect. The all-else-equal contrast sets $\boldsymbol{v} = \boldsymbol{v}'$, so the unit-level race effect at each $\boldsymbol{v} \in \mathcal{V}$ is $y_i(1, \boldsymbol{v}) - y_i(0, \boldsymbol{v})$. Because both conditions sit at that shared profile, a single distribution ρ, common across racial conditions, supplies the weights. Weighting these profile-specific effects by ρ and averaging over all N units yields the *All-Else-Equal*

Effect (AEE),

$$\text{AEE} \coloneqq \frac{1}{N} \sum_{i=1}^{N} \sum_{\boldsymbol{v} \in \mathcal{V}} \left[y_i(1, \boldsymbol{v}) - y_i(0, \boldsymbol{v}) \right] \rho(\boldsymbol{v}), \tag{4}$$

which, with the weighting common across units, coincides with the average marginal component effect (AMCE) of conjoint experiments (Hainmueller et al., 2014).

All-else-within-race (AEWR) effect. The within-race contrast leaves $\boldsymbol{v}$ and $\boldsymbol{v}'$ free to differ, so each side of the racial contrast is weighted by its own race-specific conditional q_z — the distribution of $\boldsymbol{v}$ within racial condition z — and averaging over all N units yields the *All-Else-Within-Race Effect* (AEWR),

$$\text{AEWR} \coloneqq \frac{1}{N} \sum_{i=1}^{N} \left[\sum_{\boldsymbol{v} \in \mathcal{V}} y_i(1, \boldsymbol{v})\, q_1(\boldsymbol{v}) - \sum_{\boldsymbol{v}' \in \mathcal{V}} y_i(0, \boldsymbol{v}')\, q_0(\boldsymbol{v}') \right]. \tag{5}$$

The AEE and the AEWR reverse the order in which they difference and average: The AEE differences the racial conditions at a shared profile and then averages the within-profile differences over the common ρ, while the AEWR averages each racial condition over its own q_z and then differences the race-specific averages, with no shared profile anchoring the comparison.

The two estimands differ except in special cases: The difference between the AEWR and the AEE vanishes when $q_1 = q_0$, so that each racial condition's own distribution is already the common one, or when the average potential outcome $\bar{y}(z, \boldsymbol{v}) \coloneqq (1/N) \sum_{i=1}^{N} y_i(z, \boldsymbol{v})$ is constant in $\boldsymbol{v}$ for each racial condition z, so that every distribution returns the same race-specific average. Proposition S1 in the Supplementary Material records the gap between any two members of the estimand family, mixed effects included, and shows that the same logic determines when the members differ.

Mixed effects. The choice of contrast need not be all-or-nothing. Consider racial discrimination in officer use of force (Fryer, Jr., 2019; Knox et al., 2020): Officers are the

decision-makers, civilians bear the configurations, and the researcher may want to use a legal standard to sort the features of $\boldsymbol{v}$. Conduct such as openly carrying a weapon bears on that standard and may be held fixed, while dress and neighborhood, no legally justified basis for force, may be left free to vary by race.

A mixed contrast divides the nonracial features into a fixed set and a free set. The fixed features are held equal across racial conditions, while the free features vary by race and are averaged over within each racial condition. The weighting follows the partition: A common distribution weights the fixed features, as ρ weights all of $\boldsymbol{v}$ for the AEE, and race-specific distributions weight the free features given the fixed profile, as the q_z do for the AEWR. In the use-of-force example, the two racial conditions share one weighting over whether the civilian carries a weapon, while dress and neighborhood follow each racial group's own distribution given that conduct. A mixed effect is thus all-else-equal on the fixed features and within-race on the free ones; Section S1.5 of the Supplementary Material gives the formal partition and the resulting partially marginalized average effect.

4.1.2 Perception-Based Regime

Perceived configurations lie in the same space $\mathcal{T}$ as in the exposure-based regime (Section 3.3), so the cue effect belongs to the same family of unit-level race effects. What the perception-based regime withholds is the choice among that family's members: Under the cue-level SUTVA (Section S1.1), each cue value induces a single perceived configuration $(z_i(w), \boldsymbol{v}_i(w))$ for each unit. For compliers, the cue fixes the nonracial perception at whatever value of $\boldsymbol{v}_i(w)$ the cue would induce. This degenerate distribution is common across racial conditions when $\boldsymbol{v}_i(0) = \boldsymbol{v}_i(1)$ and race-specific on at least some nonracial features otherwise.

The cue effect in (2) corresponds to the race effect in (1), now read as the effect of perceived race, only for the compliers $\mathcal{C}$ (Section 3.3). For compliers, and for compliers alone, the cue would move perceived race from 0 to 1 — the contrast in (1). The subset therefore narrows from all N units to those in $\mathcal{C}$, and under perception sufficiency (Assumption 1)

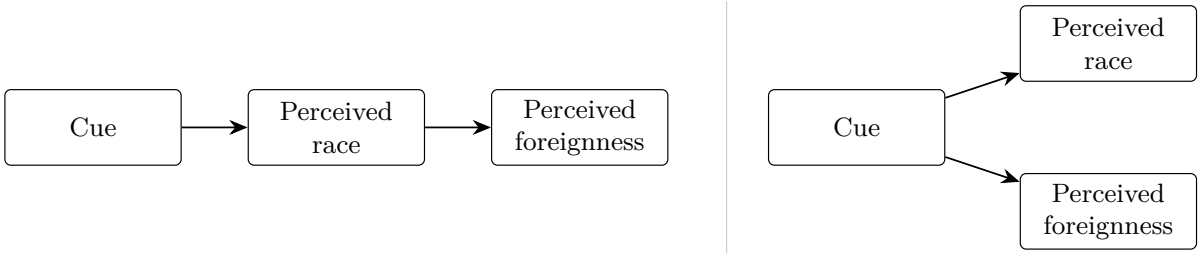
the average cue effect on $\mathcal{C}$ is what I call the *natural race effect among compliers* (NREC),

$$\text{NREC} \coloneqq \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \left[y_i \left(1, \boldsymbol{v}_i(1)\right) - y_i \left(0, \boldsymbol{v}_i(0)\right) \right], \tag{6}$$

where $|\mathcal{C}|$ is the number of compliers.

The NREC averages the mix of contrast types the cue induces among compliers. If $\boldsymbol{v}_i(0) = \boldsymbol{v}_i(1)$ for a complier, that unit contributes an all-else-equal contrast; if the cue changes every nonracial feature, a within-race contrast; and if the cue changes some nonracial features but not others, a mixed contrast. Only if $\boldsymbol{v}_i(0) = \boldsymbol{v}_i(1)$ for every $i \in \mathcal{C}$ does the NREC reduce to the all-else-equal estimand among compliers, the standard target in the perception-based literature (Dafoe et al., 2018; Elder and Hayes, 2023; Landgrave and Weller, 2022). As the introduction argues, the co-movement does not disqualify the NREC as an effect of race: Each induced contrast is a unit-level race effect in (1), and insofar as shifting nonracial perceptions together with perceived race is partly what it means for a cue to be racialized (Hu and Kohler-Hausmann, 2025), the NREC is not a failed manipulation but the effect of interest.

The label *natural* follows mediation usage (Pearl, 2001; Robins and Greenland, 1992), under which a mediator takes the value the treatment would induce. The simultaneous ordering matches how Hu and Kohler-Hausmann (2025, p. 251) describe reading a cue racially: activating the category together with the nonracial attributes the category indexes.



(a) sequential *(b) simultaneous*

Figure 1: Sequential versus simultaneous mediation, illustrated with perceived foreignness: In panel *(a)*, the cue would move the nonracial perception only through perceived race; in panel *(b)*, alongside perceived race.

Note that, although the simultaneous ordering is the one Hu and Kohler-Hausmann (2025) describe, the sequential ordering is no less compatible with the thick constructivist conception: Inference from perceived race to nonracial attribute can be one way the category's indexing operates (VanderWeele, 2015). The orderings differ instead in where the movement sits (Figure 1). Downstream of perceived race, the inferred attribute moves as part of the effect of race itself, so nothing co-moves that an all-else-equal contrast must hold fixed.

5 All-Else-Equal as a Substantive Commitment

Restricting the estimand family of Section 4 to the all-else-equal contrast involves two tradeoffs. The first is conceptual: On some conceptions of race, holding nonracial attributes fixed across racial categories may strip away part of the category's social content. The second is structural: Insofar as the distribution of nonracial attributes differs by racial category, no single weighting over profiles can simultaneously represent the distribution of those attributes within each racial group. Neither tradeoff condemns the all-else-equal contrast; together they make adopting it a substantive decision rather than a default.

5.1 The Conceptual Tradeoff

On the classical conception (Section 2), nominal membership exhausts the category, so a contrast that moves membership while holding the indexed attributes fixed compares the categories whole. On the thick constructivist conception, the same contrast compares the categories stripped of their social content. With the indexed attributes held fixed, what varies across the two individuals is nominal membership alone — the string without the bundle. The contrast can therefore refer to the effect of membership as such (Hu, 2023), but not to the effect of the category as constituted (Hu and Kohler-Hausmann, 2025). On either conception, the contrast is well-defined; the conceptions differ only in what the contrast gives up.

The conception sets what the contrast costs, not whether to adopt it. Even on the thick

constructivist conception, a study may hold the indexed attributes fixed for other reasons: because the study seeks the mechanisms through which race exerts an effect, or because differences in some attributes normatively justify differential treatment. The study then pays the cost deliberately rather than by default (Section 5.3).

5.2 The Structural Tradeoff

Even if the goal is only to represent each racial group well, the all-else-equal contrast requires one weighting over nonracial profiles applied to both racial categories. When the two groups' distributions of nonracial profiles overlap only slightly, as the arrestees' neighborhoods and prior records do in the running example, the common weighting faces a tradeoff. Confined to the small region of overlap, the weighting captures only a sliver of each group rather than the profiles where most group members lie. Spread out to represent both groups more faithfully, the weighting must place mass on profiles improbable under one group's distribution or the other's. The race-specific weightings q_z escape the tradeoff, evaluating each racial condition where its own mass lies, though the escape costs the common distribution and, with it, the all-else-equal contrast. Figure 2 depicts the two choices of common weighting, together with the race-specific weighting, in the attribute space of the running example; Proposition S2 in Section S1.6 formalizes the structural tradeoff.

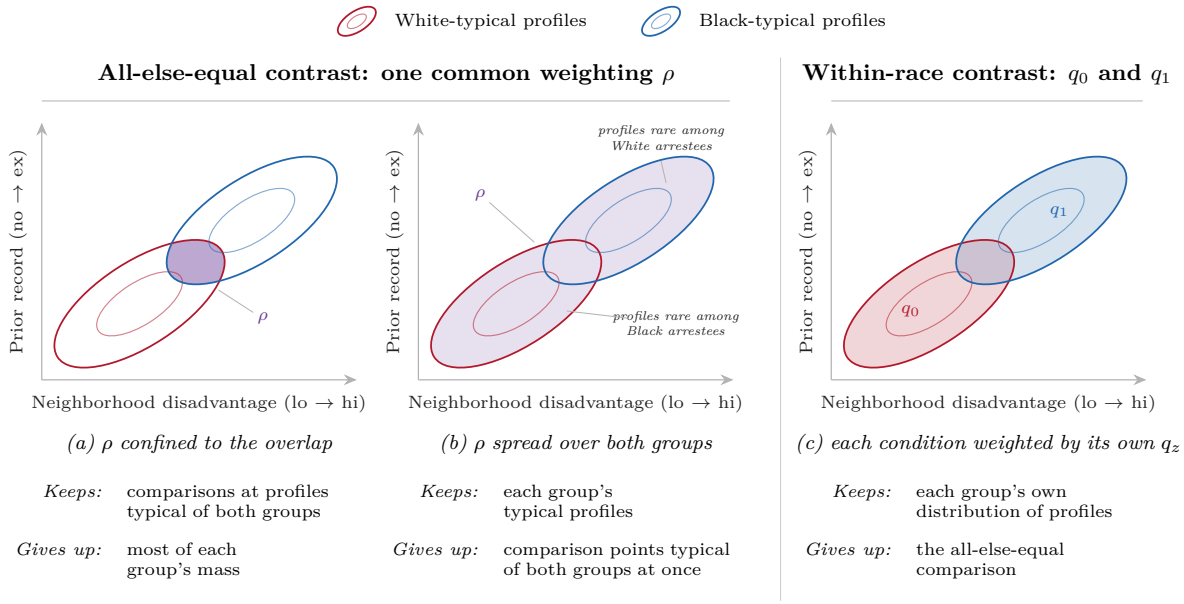


Figure 2: The structural tradeoff in the attribute space of the charging example. Each closed curve is a contour of a group's distribution of nonracial profiles (outer curve: at least a $1 - \varepsilon$ share of the group's mass, Proposition S2; inner curve: the denser core), and shading marks the support of the weighting. Panels *(a)* and *(b)* show the two choices of common weighting ρ; panel *(c)*, the race-specific weightings q_z.

5.3 Grounds for Choosing the Contrast

A classical conception of the racial category points to the all-else-equal effect, and a thick constructivist conception points to the within-race effect. Yet even on that conception, whether race indexes a feature does not alone settle whether to hold the feature fixed. The further grounds for the contrast sort the features on criteria of their own. A study that embraces the thick constructivist conception may therefore reasonably hold fixed attributes that race indexes, yielding a mixed or all-else-equal contrast (Section 4.1). The grounds that follow, for choosing first the contrast and then the weighting, are illustrative rather than exhaustive, and several may apply in one study; Table S2 in the Supplementary Material collects the grounds, ordered by which component of the estimand each ground bears on.

5.3.1 Analyzing Causal Mechanisms

Under the sequential ordering of Figure 1, panel *(a)*, decision-makers infer nonracial attributes from race, and a substantial literature asks whether such statistical, as opposed to taste-based, discrimination (Arrow, 1973; Becker, 1957; Phelps, 1972) is less morally objectionable (e.g., Alexander, 1992; Hellman, 2008; Lippert-Rasmussen, 2011). Apart from those normative debates, identifying the pathways through which racial effects operate is a positive question in its own right, an inquiry that may itself be motivated by a normative commitment to understanding how to reduce discrimination.

Holding an attribute fixed across the racial conditions can block a pathway through which race, or a signal of race, affects the outcome. An inferred attribute, however, is itself a potential outcome of race, so it cannot be held fixed directly; what a design controls is the information the configuration provides. Providing the attribute aims to preempt the inference the decision-maker could otherwise draw by fixing the corresponding belief at the provided value. If successful, the design achieves "perfect" — as opposed to imperfect — "manipulation of the mediator" (Acharya et al., 2018, p. 359). The racial contrast at that value then blocks the mediating path and isolates the remaining channels (VanderWeele and Robinson, 2014) — isolation that implies neither that the "true" effect of race is what remains once one blocks particular pathways nor that any one pathway is a privileged effect of race. For example, by stating a partisan signal alongside a racially distinctive name in constituency service requests to state legislators — who may favor copartisans (Fenno, 1978) — Butler and Broockman (2011) supply the party identification that legislators could otherwise infer from race (Mcdermott, 1998).

5.3.2 Normative and Legal Commitments

In much empirical work, normative commitments about when a decision-maker is justified in discriminating are often implicit (Hu, 2026; Hu and Kohler-Hausmann, 2025; Kohler-Hausmann, 2019). The example of racial discrimination in officer use of force (Section 4.1)

made one such commitment explicit by sorting the conduct that may justify force from the features that may not. The study of racial disparities in health care, an important area of inquiry in political science and related fields (Grogan and Park, 2017; Michener, 2018, 2020; Olvera et al., 2023), makes such commitments systematic, building them into the definition of disparity itself.

Drawing on a conception of certain disparities as "avoidable, unfair, and unjust" (Whitehead, 1992, p. 107), the landmark report of the Institute of Medicine (IOM) defined disparity in health care as "racial or ethnic differences in the quality of health care that are not due to access-related factors or clinical needs, preferences, and appropriateness of intervention" (Institute of Medicine, 2002, pp. 3–4). On this standard, differences in care between, say, Black and White patients are justified only insofar as those differences reflect needs or preferences (Cook et al., 2009a, 2012, 2009b; McGuire et al., 2006).

The IOM's needs-and-preferences criterion has generated work distinguishing "allowable" from "non-allowable" sources of racial differences in health care (Cook et al., 2012; Duan et al., 2008; Jackson, 2021). McGuire et al. (2006) implement that distinction by holding health status fixed across racial conditions while leaving race-specific distributions of non-allowable features, such as socioeconomic status, unchanged. The resulting estimand marginalizes over the allowable feature using a distribution common across racial groups and over the non-allowable features using each group's own distribution. In the vocabulary of the mixed contrast (Section 4.1), the normative distinction sorts allowable features into the held-fixed set and the non-allowable features into the free set.

5.4 Grounds for Choosing the Weighting

The choice of weighting is specific to the exposure-based regime since, as Section 4.1 explains, in the perception-based regime each complier's configurations are fixed potential responses to the cue, so no ρ or q_z enters the estimand. Like the choice of contrast, how to specify the conditionals q_z and the marginal ρ — set by the researcher, not estimated from the data (Section 4) — is a substantive choice that turns on what the comparison is

meant to represent. I give two motivations: external validity and typicality.

5.4.1 External Validity

A researcher who wants the estimand to speak to a specific target population can calibrate the joint distribution $q(z, \boldsymbol{v})$ to match that population's distribution of profiles. For the AMCE, de la Cuesta et al. (2022) calibrate the common marginal ρ to external benchmarks; because the calibrated distribution is common across racial conditions, the reweighting presupposes the all-else-equal contrast. The family of race estimands in Section 4 permits the same calibration on the race-conditional side, with each q_z matched to the target population's within-race distribution. Section S1.7 shows that calibrating each q_z requires reference information about how attributes co-vary within race, which single-attribute marginals do not supply.

5.4.2 Typicality

Section 2 described a graded conception of typicality, on which racial categories are organized around prototypes and typicality may vary continuously with perceived similarity to the prototype. The conditional distribution q_z can encode that gradation by assigning greater weight to profiles that are more typical of group z. Ma et al. (2015), for instance, operationalize such typicality with racial-prototypicality scores derived from respondents' photograph ratings.

One way, among many, to translate graded typicality into the conditional distribution $q_z(\boldsymbol{v})$ builds on the construction of Gärdenfors (2000, 2014). In this construction, how typical a profile is of group z — for example, how typically Black or White the profile is — corresponds to the profile's closeness to the group's prototype, with each nonracial attribute weighted by how diagnostic it is of group membership. Profiles near a Black prototype are therefore more probable under q_1, and profiles near a White prototype are more probable under q_0. In a name-based audit study, for example, the Black prototype might center on names like DeShawn or Jamal, with Connor or Greg at the periphery. The

rate at which typicality declines with distance from the prototype may differ across groups, so DeShawn may register as more atypically White than Greg does atypically Black.

6 Recovering the Estimand Family without Holding All Else Equal

Section 5 treated the all-else-equal contrast as a choice of estimand; whether data recover a chosen estimand is a separate question. The literature's answer to each recovery problem — confounding under exposure, excludability under perception — does double duty, securing recovery while silently restricting the estimand to the all-else-equal contrast. This section shows that, in each regime, a weaker condition secures recovery alone, so the choice of contrast and weighting remains substantive.

6.1 Confounding

A unit's race effect contrasts two potential outcomes — the responses to a Black-marked and to a White-marked file — and at most one can be observed; nor does the same prosecutor at a later decision supply the other. A docket that fills toward the end of a term, attention that flags over a day, or an intervening high-profile case may each change the prosecutor, so the response either file receives need not equal the response the same file would have received at the other decision. When chance alone determines which unit encounters which file, the units under different configurations differ only by chance, so systematic differences in their responses reflect the files. Studies of discrimination therefore remove confounding through the assignment of configurations to decision-makers. Experiments randomize that assignment directly (e.g., Bertrand and Mullainathan, 2004; Sen and Wasow, 2016), and observational studies are judged by how closely their assignment processes approximate such a randomization (Cochran, 1965; Rubin, 2008).

By two conventions of practice, the assignment also fixes the estimand. First, the association between race and the nonracial attributes it indexes is itself called confounding, and a problem so named invites a remedy in the assignment of configurations. Second, the

weights that define the estimand are read off the assignment rather than chosen in their own right.

The literature's guidance prescribes one form of assignment: The two racial conditions share one distribution over nonracial profiles — so that whatever neighborhood and record a file contains, its chance of being marked Black is the same — with identical probabilities for every unit. This one act of random assignment is chosen to satisfy two conditions in the name of removing confounding, and only one of the two matters for unconfounded inference.

The two conditions do two jobs — selecting the estimand and securing valid inference — and because both are properties of one mathematical object, the researcher's single choice of assignment settles both at once. Write the assignment as the array of unit-level probabilities $p_i(z, \boldsymbol{v}) \coloneqq \Pr(\boldsymbol{T}_i = (z, \boldsymbol{v}))$ that unit i encounters configuration $(z, \boldsymbol{v})$, and define the two conditions as two invariances of that array, in deliberately parallel form.

Definition 1 (Unit-blind and profile-blind assignments)**.** *An assignment with unit-level probabilities $p_i(z, \boldsymbol{v})$ is unit-blind if $p_i(z, \boldsymbol{v})$ does not vary with i: Every unit faces the same probabilities, so the assignment cannot track a unit's attributes or potential outcomes. The assignment is profile-blind if $\Pr(Z = z \mid \boldsymbol{V} = \boldsymbol{v})$ does not vary with $\boldsymbol{v}$: Race falls independently of the nonracial profile, so neither racial condition is tied to particular profiles.*

Unit-blindness is an invariance across units, holding when all rows of the array are identical, and profile-blindness is an invariance across profiles, holding when the share of probability on each racial condition is the same within every nonracial profile.

A stripped-down charging example makes the two conditions — unit-blindness and profile-blindness — concrete (Figure 3): Three prosecutors each read their own case file, and each file pairs the arrestee's race with a nonracial profile of neighborhood disadvantage (lo/hi) and prior record (no/ex), and the assignment's probabilities form a 3×8 array, one row per prosecutor and one column per configuration. The uniform default of Hainmueller et al. (2014) (panel *a*) and a population-calibrated design of de la Cuesta et al. (2022) (panel *b*) satisfy both invariances: The rows are identical, and the chance that a file is

marked Black, $\Pr(Z = 1 \mid \boldsymbol{V} = \boldsymbol{v})$, is constant across nonracial profiles; the only difference between the two panels is the weighting ρ common to the two racial conditions.

The two conditions can fail separately. Panel *(c)* keeps the rows identical but lets the chance that a file is marked Black vary across nonracial profiles, from 1/4 at (lo, no) to 3/4 at (hi, ex), so unit-blindness holds while profile-blindness fails — and because every configuration retains positive probability, panel *(c)* supports the same inferences as panels *(a)* and *(b)*. Unit-blindness would fail instead if some prosecutor's row differed from the rest. If, say, prosecutors in heavily policed boroughs were both more likely to receive Black-marked files and more likely to charge, the Black-file cell means would draw disproportionately on those prosecutors, mixing the effect of the file with differences among the prosecutors who read it.

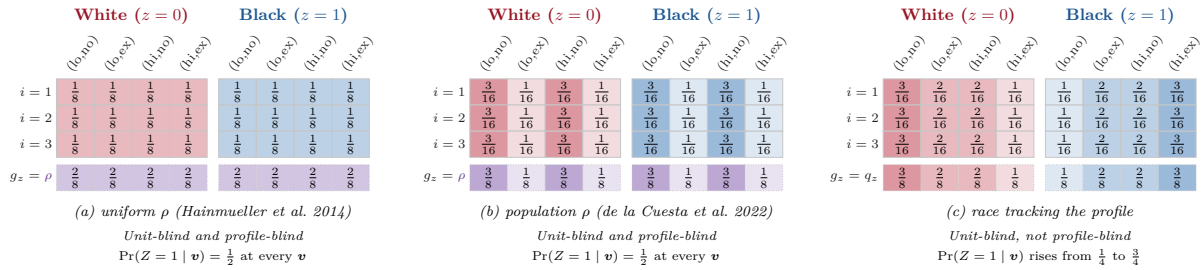


Figure 3: Three configuration assignments for the charging example: one row per prosecutor i, one column per configuration $(z, \boldsymbol{v})$, cell shading proportional to the cell's probability. The dotted bottom margin reports the weights read off each assignment, $\Pr(\boldsymbol{V} = \boldsymbol{v} | Z = z)$: one common weighting ρ in panels *(a)* and *(b)*, the race-specific q_z in panel *(c)*.

Unit-blindness settles the inferential problem by itself. Every member of the estimand family is a difference between two race-specific weighted averages of cell means (Section 4). Part *(i)* of Proposition 1 states that a unit-blind assignment recovers every member of the family without bias, so long as the assignment places positive probability on every profile with positive weight under the target (Section 7); the proof never invokes profile-blindness. Profile-blindness is neither necessary nor sufficient for recovering any member of the estimand family; profile-blindness contributes nothing to whether inference is confounded. If profile-blindness buys no inference, what is profile-blindness buying?

The answer is the second of the two conventions. The field reads the estimand's weights

off the assignment, so the distribution the design places on nonracial profiles within each racial condition becomes the distribution over which the estimand averages, $\Pr(\boldsymbol{V} = \boldsymbol{v} | Z = z)$. The convention is nearly universal: As de la Cuesta et al. (2022, pp. 20–21) report, 88% of reviewed conjoint articles randomize every factor with the uniform distribution, and fewer than 4% theoretically motivate the randomization distribution in the main text.

Proposition 1. *Suppose the configuration-level SUTVA (Assumption S1), so that the estimand family of Section 4 is well defined, and let the assignment be unit-blind, with fixed cell counts and $p(z, \boldsymbol{v}) > 0$ on every profile with positive weight under the target. (i) For every member of the estimand family, the difference in cell means reweighted to that member's weighting is unbiased for that member; profile-blindness plays no role. (ii) If instead the researcher reads the weighting off the assignment, so that nonracial profiles within each racial condition are weighted by $\Pr(\boldsymbol{V} = \boldsymbol{v} \mid Z = z)$, then the weighting is common across the two racial conditions if and only if the assignment is profile-blind.*

Section S1.3 gives the proof of part *(ii)*; part *(i)* is the unbiasedness of the plug-in estimator of the estimand family, developed in Section 7 and proved in Section S1.8. Part *(i)* is what randomization buys; part *(ii)* is what profile-blindness buys. A weighting common to the two racial conditions is the all-else-equal structure of (4), so profile-blindness, read for its weighting, is the selection of the all-else-equal member. Profile-blindness was never doing inference; profile-blindness was choosing the estimand.

Both concerns travel under one name, *confounding* — applied by the statistical literature on causal inference to dependence between the assignment and the potential outcomes, and by critiques of discrimination experiments (Heckman, 1998) to the association between race and the nonracial attributes it indexes. One name applied to two conditions presents their joint removal as a single requirement, which is how the choice of estimand came to arrive inside a credibility condition: To remove confounding, in the undifferentiated sense, was to run a clean experiment and, in the same act, to choose the all-else-equal effect.

Reserving *confounding* for failures of unit-blindness breaks the fusion and matches the word's formal definition: An assignment mechanism is unconfounded when its probabilities do not depend on units' potential outcomes (Imbens and Rubin, 2015; Rosenbaum and

Rubin, 1983). On this usage, the association between race and the nonracial attributes it indexes is never a confounder. Sorting those attributes into the held-fixed and the free is the choice of contrast, a choice made on substantive and often normative grounds (Section 5). Confounding remains a property of the assignment, so normative considerations enter the choice of estimand, not the inference that recovers the chosen estimand.

The distinction between choosing and recovering the estimand also explains the conclusion, noted in Section 2, that positive inquiry into the effects of race cannot be separated from normative inquiry (Hu, 2026; Hu and Kohler-Hausmann, 2025). When one act of assignment must both select the estimand and secure inference, the normative grounds of the estimand present themselves as normative grounds of inference. Separate the two jobs, and the appearance dissolves: The estimand can be, and often is, normative; inference, once the estimand is fixed, is not.

Two connections between assignment and estimand remain, and both run from the estimand to the design: support — which profiles the assignment can produce at all — since the assignment must place positive probability on every profile with positive weight under the chosen member of the estimand family, and precision, taken up in Section 7. The design secures the estimand's recovery; the design never selects the estimand. Lundberg et al. (2021) recommend the same order, estimand before design, for empirical research generally, and Blair et al. (2019) formalize the order for political science.

6.2 Excludability

In the perception-based regime, the researcher assigns the cue $w_i \in \{0, 1\}$; the perceptions the cue would induce, $z_i(w)$ and $\boldsymbol{v}_i(w)$, are potential outcomes of the cue (Section 3.3). The target is the NREC of (6), the average over the compliers $\mathcal{C}$ of the outcome contrasts the cue would induce. Complier status is counterfactual — defined by both potential racial perceptions while any realized cue reveals at most one — so the NREC cannot be formed by subsetting. Recovery proceeds instead through two averages over all N units, the cue's average effect on the outcome and the cue's average effect on perceived race;

under assumptions given below, the ratio of the first to the second equals the NREC:

$$\frac{\frac{1}{N}\sum_{i=1}^{N}\left[y_i\left(z_i(1),\boldsymbol{v}_i(1)\right)-y_i\left(z_i(0),\boldsymbol{v}_i(0)\right)\right]}{\frac{1}{N}\sum_{i=1}^{N}\left[z_i(1)-z_i(0)\right]}, \tag{7}$$

the cue's average effect on the outcome divided by the cue's average effect on perceived race.

Suppose that no unit would invert the cue, perceiving the Black cue as White and the White cue as Black, so that every unit is a complier, an always-taker, or a never-taker (Section 3.3). The two averages in the ratio then run over the compliers, whose contrasts the NREC averages, and the non-compliers, whose contrasts the NREC excludes. Any difference between the ratio in (7) and the NREC in (6) therefore comes from the non-compliers alone. For the ratio to equal the NREC, it thus suffices that the non-compliers contribute nothing, and the following assumption states that condition.

Assumption 2 (No isolated nonracial shift)**.** *For all units* $i \in \{1,\ldots,N\}$ *with* $z_i(0) = z_i(1)$, $\boldsymbol{v}_i(0) = \boldsymbol{v}_i(1)$.

Assumption 2 asserts that the cue would not move nonracial perceptions on its own: A decision-maker who would read "Lakisha Washington" and "Emily Walsh" (Bertrand and Mullainathan, 2004) as the same race would, under the assumption, pick up none of the names' nonracial content.

The assumption restricts non-compliers alone. A complier's perceptions remain entirely free: Racial perception may shift with nonracial perceptions in tow, or shift by itself. The NREC averages whatever joint movement the cue would induce among compliers.

Proposition 2. *Suppose the cue-level SUTVA (Assumption S2), no defiers, the existence of at least one complier — conditions stated formally in Section S1.4 — and perception sufficiency (Assumption 1). Under Assumption 2, the ratio in* (7) *equals the NREC in* (6)*.*

The literature secures the equality of the ratio and the NREC through conditions stronger than Assumption 2 and frames the stronger conditions in terms of excludability. Butler and Homola (2017, pp. 122–123) assume the outcome does not depend on

perceived nonracial features, $y_i(z, \boldsymbol{v}) = y_i(z, \boldsymbol{v}')$. Dafoe et al. (2018, eq. 6, p. 404) require that the cue move no unit's nonracial perceptions, $\boldsymbol{v}_i(0) = \boldsymbol{v}_i(1)$ for all i — *information equivalence* in their vocabulary, *cue stability* hereafter — and relate the requirement to the exclusion restriction in instrumental-variable analysis (Elder and Hayes, 2023; Landgrave and Weller, 2022). With defiers ruled out, every unit is a complier or a non-complier, so a condition imposed on every unit says two separate things: what the condition requires of non-compliers and what the condition requires of compliers.

Proposition 3 (Decomposition of cue stability)**.** *Suppose the cue-level SUTVA (Assumption S2), no defiers, the existence of at least one complier, and perception sufficiency (Assumption 1). Cue stability is the conjunction of (i) Assumption 2 and (ii) complier cue stability:* $\boldsymbol{v}_i(0) = \boldsymbol{v}_i(1)$ *for all* $i \in \mathcal{C}$*. Conjunct (i) is necessary and sufficient for the ratio in* (7) *to equal the NREC, and, given conjunct (i), conjunct (ii) is necessary and sufficient for the NREC to equal the AEE among compliers.*

The two conjuncts restrict disjoint sets of units and do different work (Figure 4). Given conjunct *(i)*, the ratio equals the NREC, so no recovery remains for conjunct *(ii)* to help; what conjunct *(ii)* determines is the estimand. Holding compliers' nonracial perceptions fixed across the cue values leaves each complier's induced contrast differing in perceived race alone, and the NREC collapses into the AEE among compliers (Section 4). I draw from the decomposition a distinction the literature leaves implicit: Conditions on the units outside the target subgroup bear on whether the ratio equals the NREC; conditions on the units inside the target subgroup bear on which member of the estimand family the ratio equals. The distinction is the perception-based counterpart of unit-blindness and profile-blindness in Section 6.1: Conjunct *(i)* does the recovery work of unit-blindness, and conjunct *(ii)* the estimand work of profile-blindness.

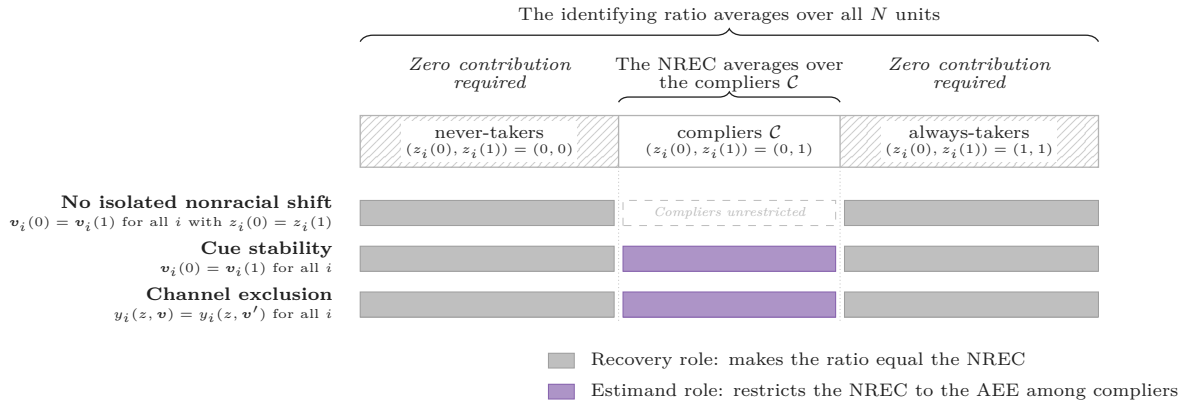


Figure 4: Excludability conditions as coverage over the compliance types (defiers ruled out; segment widths illustrative). Brackets: The ratio in (7) averages over all N units, while the NREC averages over the compliers only, so the hatched flanks must contribute zero. Rows: Each assumption covers the units it restricts; color records each block's role (Proposition 3).

Reserving *excludability* for conditions on the units the target excludes gives the word one target-invariant meaning: Conjunct *(i)* exhausts excludability's content whichever member a study seeks, while conjunct *(ii)* is a condition of a different kind — not recovery but the selection of the all-else-equal member. A researcher who wants the AEE among compliers may therefore design cues with conjunct *(ii)* as the aim; a researcher who wants the NREC needs conjunct *(i)* alone, and on either target the movement of compliers' nonracial perceptions does not violate excludability.

The ratio in (7) is the Wald ratio (Angrist and Pischke, 2008, pp. 127–128; Gerber and Green, 2012, p. 180), and the recovery of the NREC in Proposition 2 has the same structure as the identification of local average treatment effects (Angrist et al., 1996; Imbens and Angrist, 1994), with the cue as the instrument and the perceived configuration — racial and nonracial components together — as the treatment. Section S1.4 of the Supplementary Material develops the parallel, including why conjunct *(ii)* has no counterpart in the classical setting of instrumental variables.

In practice, many cue-based studies measure perceived race, if at all, in a separate sample rather than on the decision-makers who supply the outcome, so such studies can estimate only the numerator of (7); the numerator shares the NREC's sign and understates

the NREC's magnitude (Corollary S1 of the Supplementary Material), and a researcher can treat the unknown complier share as a sensitivity parameter, analogous to the sensitivity analyses in Leavitt and Rivera-Burgos (2024, 2026).

7 Estimation and Inference across the Estimand Family

7.1 Exposure-Based Regime

Under the exposure-based regime, the researcher's design choice is the set of cell counts — how many units encounter each profile — and random assignment allocates units to profiles given those counts.

Assumption 3 (Complete random assignment of configurations)**.** *Given cell counts $n_{z,\boldsymbol{v}}$, one positive integer per profile $(z, \boldsymbol{v}) \in \mathcal{T}$, the configuration assignment $\boldsymbol{T} = (\boldsymbol{T}_1, \ldots, \boldsymbol{T}_N)$ takes each value in the resulting set Ω with probability $1/|\Omega|$, so $\Pr(\boldsymbol{T}_i = (z, \boldsymbol{v})) = n_{z,\boldsymbol{v}}/N$.*

Because the probabilities $n_{z,\boldsymbol{v}}/N$ do not vary with the unit, the assignment is unit-blind (Definition 1), unable to track a unit's attributes or potential outcomes; the assumption leaves profile-blindness unconstrained. Unit-blindness suffices to recover whichever member of the estimand family the researcher targets (Section 6.1).

For each profile $(z, \boldsymbol{v}) \in \mathcal{T}$, the one quantity estimated from data is the cell mean

$$\hat{\mu}(z, \boldsymbol{v}) \coloneqq \frac{1}{n_{z,\boldsymbol{v}}} \sum_{i=1}^{N} Y_i \, \mathbb{1}\{\boldsymbol{T}_i = (z, \boldsymbol{v})\}, \tag{8}$$

the average outcome among units assigned to profile $(z, \boldsymbol{v})$, where Y_i is unit i's observed outcome and $\mathbb{1}\{\cdot\}$ is the indicator function, equal to 1 when the condition in braces holds and 0 otherwise. Every estimand in Section 4 is a difference between two race-specific weighted averages of these cell means, so one plug-in estimator covers the entire estimand family. Writing g_0 and g_1 for the weighting distributions the target estimand places on the

profiles with $z = 1$ and $z = 0$, the plug-in estimator is

$$\widehat{\theta} \coloneqq \sum_{\boldsymbol{v} \in \mathcal{V}} g_1(\boldsymbol{v})\, \hat{\mu}(1, \boldsymbol{v}) - \sum_{\boldsymbol{v} \in \mathcal{V}} g_0(\boldsymbol{v})\, \hat{\mu}(0, \boldsymbol{v}). \tag{9}$$

The AEE takes the common marginal, $g_0 = g_1 = \rho$; the AEWR takes the race-specific conditionals, $g_z = q_z$; and a mixed contrast takes a weight that is common on the held-fixed features and race-specific on the free ones (Section S1.5). These weights carry no hat: The researcher fixes them in advance, so (9) targets the chosen member of the estimand family, and changing the weights changes the estimand, not the estimator's ability to recover the estimand.

Under Assumption 3, $\widehat{\theta}$ is unbiased and consistent for its estimand, whatever the weights g_0 and g_1. A single construction yields valid confidence intervals for every member of the estimand family (Section S1.8). Unbiasedness holds under any design that assigns at least one unit to every profile cell: The estimator reweights the cell means to the target distribution whatever the cell counts $n_{z,\boldsymbol{v}}$, so the weights define the estimand and the counts do not.

The variance of $\widehat{\theta}$ grows as the weights $g_z(\boldsymbol{v})$ depart from the assigned shares $n_{z,\boldsymbol{v}}/N$, so a researcher who controls the design can increase precision by assigning profiles in proportion to the target weights — the strategy de la Cuesta et al. (2022) develop for the AMCE, which in the estimand family of Section 4 means assigning by the common ρ for an all-else-equal target and by the race-specific q_z for a within-race target. Neither weighting is inherently more precise. As in Section 6.1, the order runs from estimand to design: A researcher who specifies the member of the estimand family to target can then choose the assignment that recovers the member precisely.

7.2 Perception-Based Regime

In the perception-based regime the target is the NREC, and the researcher assigns only the cue.

Assumption 4 (Complete random assignment of cues)**.** *Given cue counts n_w, one positive integer per value $w \in \{0,1\}$, the cue assignment $\boldsymbol{W} = (W_1, \ldots, W_N)$ takes each value in the resulting set Ω_W with probability $1/|\Omega_W|$, so* $\Pr(W_i = w) = n_w/N$.

Randomizing the cue does not randomize the racial contrast, which is a potential outcome of the cue. Recovery therefore needs, beyond randomization, the conditions of Section 6.2: no defiers and at least one complier (Section S1.4), and Assumption 2; nothing stronger is required.

Under Assumption 4, unit i's cue is now the random variable W_i, and $Z_i \coloneqq z_i(W_i)$ is the perceived race that W_i induces (Section 3.3). Two difference-in-means estimators compare the outcome and perceived race across the two cue values — the reduced-form estimator $\widehat{\Delta}_Y$ and the first-stage estimator $\widehat{\Delta}_Z$:

$$\widehat{\Delta}_Y \coloneqq \frac{1}{n_1}\sum_{i=1}^{N} Y_i \, \mathbb{1}\{W_i = 1\} - \frac{1}{n_0}\sum_{i=1}^{N} Y_i \, \mathbb{1}\{W_i = 0\}, \tag{10}$$

$$\widehat{\Delta}_Z \coloneqq \frac{1}{n_1}\sum_{i=1}^{N} Z_i \, \mathbb{1}\{W_i = 1\} - \frac{1}{n_0}\sum_{i=1}^{N} Z_i \, \mathbb{1}\{W_i = 0\}. \tag{11}$$

When perceived race is measured on the same respondents who supply the outcome, $\widehat{\text{NREC}} \coloneqq \widehat{\Delta}_Y/\widehat{\Delta}_Z$ estimates the NREC: By Proposition 2, the ratio is the instrumental-variables estimator (Section 6.2), with the cue as the instrument and perceived race as the treatment taken up. Absent a first stage, as in most cue-based audits, only $\widehat{\Delta}_Y$ is available, and $\widehat{\Delta}_Y$ unbiasedly estimates the lower bound in magnitude of the NREC (Corollary S1 of the Supplementary Material). A confidence interval for the ratio can be derived from a delta-method approximation that, like the usual instrumental-variables variance, degrades when the cue moves perceived race only weakly; Section S1.9 gives the justification and alternatives for weak instruments.

8 Application: Race and Language in Candidate Evaluation

Would Hispanic voters evaluate a Hispanic candidate more favorably than an Anglo candidate? Coethnic preference is the expectation of the literature on linked fate — the belief that one's own life chances are tied to those of the group — and group consciousness (Dawson, 1994; G. R. Sanchez, 2008; G. R. Sanchez and Masuoka, 2010). I take the question to the candidate-evaluation experiment of Zárate et al. (2024), read under both regimes of Section 3. Across both regimes, which conclusion is right comes down to which effect of race the study targets, not to which estimand the data better recover.

In a 2×3 between-subjects experiment, Zárate et al. (2024) vary a candidate's race, Anglo ($z = 0$) or Hispanic ($z = 1$), and the language of a recorded campaign speech. The candidate, a hypothetical state representative seeking reelection, is named either Josh Martin or Josué Martínez, and the written vignette that introduces the recording gives the full name and states the candidate's race outright: "Representative [Martin/Martínez], who is [White/Latino]" (Zárate et al., 2024, p. 368). Language is the sole nonracial feature, so $\mathcal{V} = \{\text{English}, \text{Non-Native Spanish}, \text{Native Spanish}\}$: The speech is in English throughout or includes two sentences in Spanish, spoken with an American-English or a native accent, and both racial conditions use the same recordings, made by a single voice actor. Each respondent in a national sample of Hispanic adults listens to one randomly assigned recording and rates the candidate on an index combining vote intention, trust, liking, and feeling represented.

The design can be read under both regimes. The exposure reading takes the stated race as assigned, and the perception reading — since a respondent may perceive a race other than the one stated — takes the name and label together as a single cue. Zárate et al. (2024) also measure the race each respondent perceives — a measurement many cue-based studies lack — so the perception-based regime's target is directly estimable (Section 8.2).

Zárate et al. (2024) ask whether Spanish-language appeals raise Hispanic respondents'

evaluations and whether respondents evaluate the Hispanic candidate more favorably than the Anglo candidate. On the second question, the authors compare the two candidates at one language at a time and find coethnic preference under English but no detectable preference under either Spanish condition. Those comparisons hold language fixed; the reanalysis asks the same question without deciding in advance whether language is held fixed across the racial conditions (the AEE) or varies with race (the AEWR).

Whether the AEE or the AEWR answers the question of coethnic preference depends on whether one treats language as separable from the category Hispanic — the choice of Section 2 between the classical conception, on which a racial category consists in nominal membership alone, and the thick constructivist conception, on which the category also probabilistically indexes nonracial attributes. Commenting on the seminal Urban Institute audit of Hispanic and Anglo job applicants (Cross et al., 1990), Heckman and Siegelman (1993, pp. 217–218) exemplify the classical conception: They read the Hispanic accent as separable from the category, to be held fixed to isolate discrimination against "Hispanics per se."

For the most part, the literature on Latino politics instead treats language as part of what the category is, on both sides of the encounter — the candidate's and the respondent's. On the candidate's side, Hispanic candidates court Hispanic voters (Barreto, 2007) in Spanish (Abrajano, 2010), so the two racial conditions differ in their language distributions as part of the category's social content. On the respondent's side, race is perceived partly through speech — "sounding like a race" (Rosa, 2018) — and to be read as Hispanic is in part to be read as foreign (Chavez, 2013), among Latino perceivers as well as White ones (Zou and Cheryan, 2017).

Each side of the encounter matters under one regime. Under the exposure reading, the stated race is taken as assigned, and the open choice is the weighting over language; the weights come from the candidate's side. Weighting each racial condition by its own campaign distribution, the AEWR answers the constructivist question; weighting both by a common distribution, the AEE answers the classical one (Section 8.1). Under the percep-

tion reading, the candidate's race is what the respondent perceives, and the respondent's side determines what the cue induces. Speech shapes which race the respondent perceives, foreignness may move alongside perceived race, and the target is the NREC (Section 8.2).

Throughout the application, the estimands average over Hispanic respondents' heterogeneity in origin, generation, and language (Dowling, 2014; Masuoka and Junn, 2013; D. T. Sanchez et al., 2012). On the candidate's side, the campaign distributions represent how candidates speak, not how Hispanic voters do. Narrower subsets or different reference distributions would define other members of the estimand family (Section 4).

8.1 The Exposure-Based Reading: All-Else-Equal versus Within-Race over Language

The exposure reading turns on the contrast over language, held fixed (the AEE) or varying with race (the AEWR). The AEE weights language by a common distribution ρ. The AEWR instead weights language by each racial condition's own q_z.

I ground both distributions in how candidates campaign. Zárate et al. (2024) analyze every Spanish-language U.S. House advertisement aired between 2010 and 2018. Among the candidates who campaigned in Spanish — 49 Anglo and 77 Hispanic — 96 percent of Anglos used an American accent (Non-Native Spanish) and 83 percent of Hispanics a native one (Table 1). Being Hispanic rather than Anglo thus flips which accent is typical. The campaign frequencies are the race-specific distributions q_z; weighting them by each racial condition's share of the Spanish-language campaigners gives the common distribution ρ.

These weights enact both grounds of Section 5.4: The weights calibrate the comparison to the population of real candidacies (external validity) and evaluate each racial condition at the accent typical of its own campaigners (typicality). The two grounds coincide because typicality is here defined by the target population itself, though a researcher could define typicality against a different reference distribution. Because the same two race-conditional distributions barely overlap, the common weighting faces the structural tradeoff of Section 5: Any one distribution either confines the comparison to the sliver of overlap or

places weight on an accent rare under one of the two distributions (Figure 2).

	Non-Native Spanish	Native Spanish	Total	Share of candidates
Within Anglo, q_0	.96	.04	1.00	.39
Within Hispanic, q_1	.17	.83	1.00	.61
Common weight, ρ	.48	.52	1.00	

Table 1: Language weights among U.S. House candidates who aired a Spanish-language advertisement, 2010–2018 (Zárate et al., 2024). The common weight is the share-weighted average $\rho = .39\,q_0 + .61\,q_1$.

The plug-in estimator of Section 7 gives an AEWR of 0.10 (95% CI $[0.04, 0.15]$), distinguishable from zero, and an AEE of -0.01 (95% CI $[-0.05, 0.03]$), not (Figure 5). On the same experiment, then, the choice of estimand is the difference between finding coethnic preference and finding none.

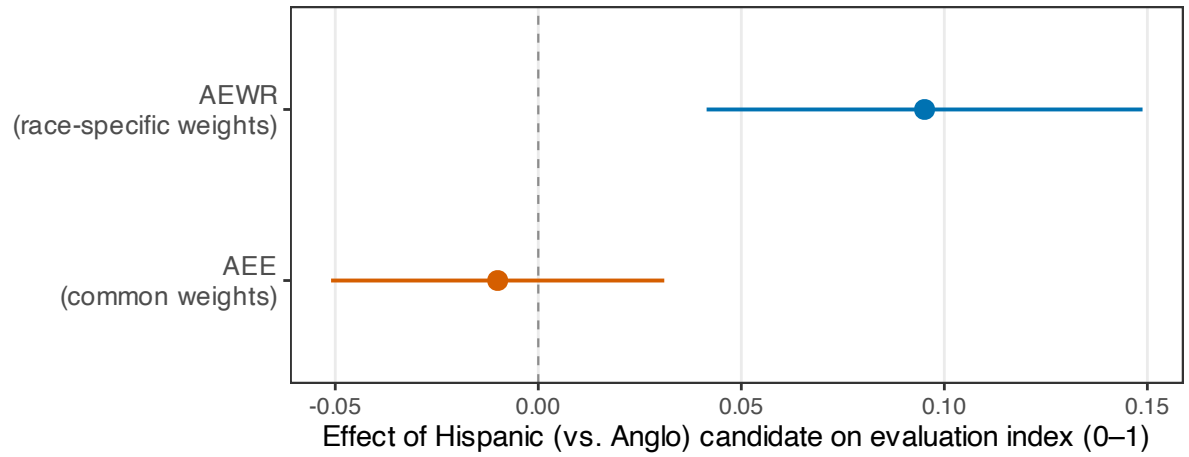


Figure 5: AEWR and AEE estimates of the effect of a Hispanic (versus Anglo) candidate on Hispanic respondents' evaluations, restricted to the two Spanish-language conditions. Bars are 95% confidence intervals based on consistent estimators of the variance upper bound and a Normal approximation.

The source of the divergence between the AEWR and AEE estimates is visible in the race-by-language cell means (Table 2). Within each language, the Anglo and Hispanic mean evaluations are nearly identical. However, across languages, respondents rate Native Spanish more positively than Non-Native Spanish, for candidates of both races.

Candidate	Non-Native Spanish	Native Spanish
Anglo	0.49	0.63
Hispanic	0.49	0.62

Table 2: Mean evaluation (0–1 index) among Hispanic respondents by candidate race and campaign language, Study 1.

The AEE averages the two near-zero within-language race contrasts and is essentially zero: The common weight ρ puts 0.52 on native Spanish for Anglo candidates, far above the 4 percent who campaign that way, crediting them with evaluations they rarely earn. Because both within-language race contrasts are near zero, any common weighting would leave the AEE near zero — whether the uniform default (Hainmueller et al., 2014, p. 12) or the target population's distribution for external validity (de la Cuesta et al., 2022), based here on the campaign data. The AEWR instead compares the typical Hispanic candidate (Native Spanish, 0.62) with the typical Anglo candidate (Non-Native, 0.49), and is positive.

If the category Anglo indexes language, so that being Anglo consists partly in rarely speaking native Spanish, then the AEE weights an atypical profile as though it were common, while the AEWR does not (Section 5). Both estimands are recovered from the same assignment under the same assumptions: The experiment randomizes race and language jointly, so every race-by-language cell is filled by random assignment, and each estimand is a different weighting of the same cell means (Section 6.1).

8.2 The Perception-Based Reading: The Natural Race Effect among Compliers

The name and label form the cue w, and the perceived-race item measures the cue's potential outcome, perceived race $z_i(w)$, at the assigned cue value (Section 3.3). The item directly follows the recording and precedes the evaluation items. Answering the item could raise the salience of race for the evaluations that follow, but only salience that the item raises more under one cue value than the other would enter the contrast. The Hispanic

name and label might raise the salience of race more than the White ones do, a possibility the design cannot rule out.

I estimate the NREC within each language condition. At a fixed language, respondents hear the same recording and differ only in the name and label, so every difference in their perceptions, racial or nonracial, is induced by the cue. Pooled across the language conditions, perceived race would move with the assigned language as well as with the cue, and the compliers would no longer be defined at a single language.

The instrumental-variables ratio of the evaluation on perceived race, with the cue as the instrument, estimates the NREC (6) (Section 7). Figure 6 reports, by language condition, estimates of the first stage — the cue's effect on the share perceiving the candidate as Hispanic — and of the NREC.

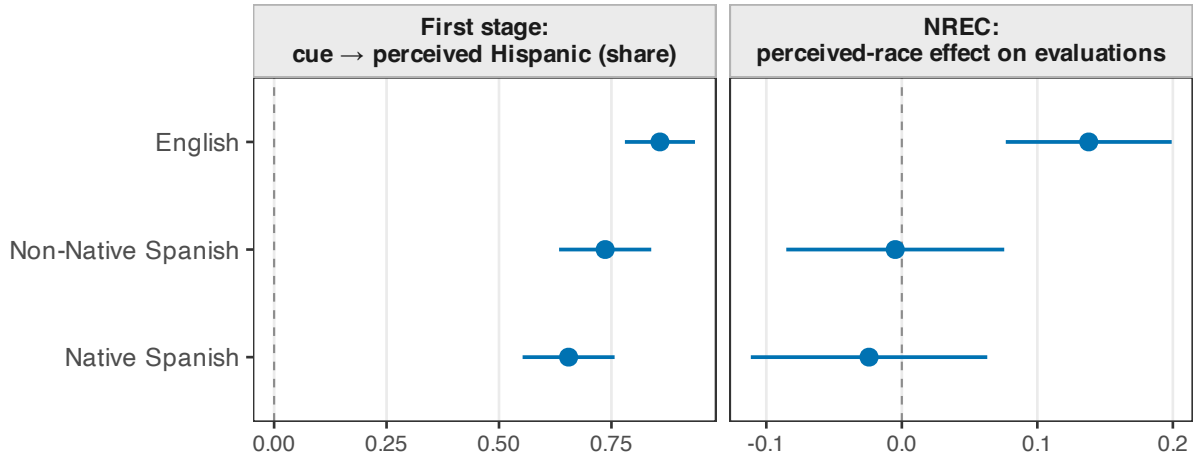


Figure 6: Perception-based estimates by language condition (Study 1, Hispanic respondents): the first stage (*left*) and the NREC (*right*). Bars are 95% confidence intervals; the NREC interval uses a delta-method standard error (Pashley, 2022) with a Normal approximation.

The NREC answers the question of coethnic preference in the perception reading's terms, and the answer depends on the language the candidate speaks. Under English, the cue moves perceived race sharply (estimated first stage 0.86) and the estimated NREC is positive (0.14, 95% CI $[0.08, 0.20]$): Respondents evaluate a candidate they perceive as Hispanic more favorably than one they perceive as White.

In both Spanish conditions — non-native and native — the estimated NREC is indis-

tinguishable from zero, and two patterns may together account for the null results. First, within a Spanish condition the race contrast in evaluations all but vanishes, as though speaking Spanish mutes the role of perceived race in respondents' judgments, consistent with accented language operating as an intergroup cue in its own right (Hopkins, 2015). Second, the accent appears to reshape perceived race itself, weakening the cue's first stage: Under native Spanish, 35 percent of respondents perceive the Anglo candidate as Hispanic, against 8 percent under English, while under non-native Spanish, 35 percent perceive the Hispanic candidate as White, against 6 percent. The two patterns together suggest that respondents perceive race partly through the speech that conveys it — "sounding like a race" (Rosa, 2018) registering in the first stage itself — and leave a clear perceived-race contrast only under English.

Whether that perceived-race effect is also the all-else-equal effect of race turns on the cue moving compliers' perceived race and nothing else, which a name bearing a race label may fail to satisfy. Figure 7 bears on whether the name signals more than race, with the first-name trait ratings of Elder and Hayes (2023): "Josué," the Hispanic name Zárate et al. (2024) use, tracks Hispanic names as a whole, rated markedly more foreign, modestly more working-class, and less Republican than White-coded names.

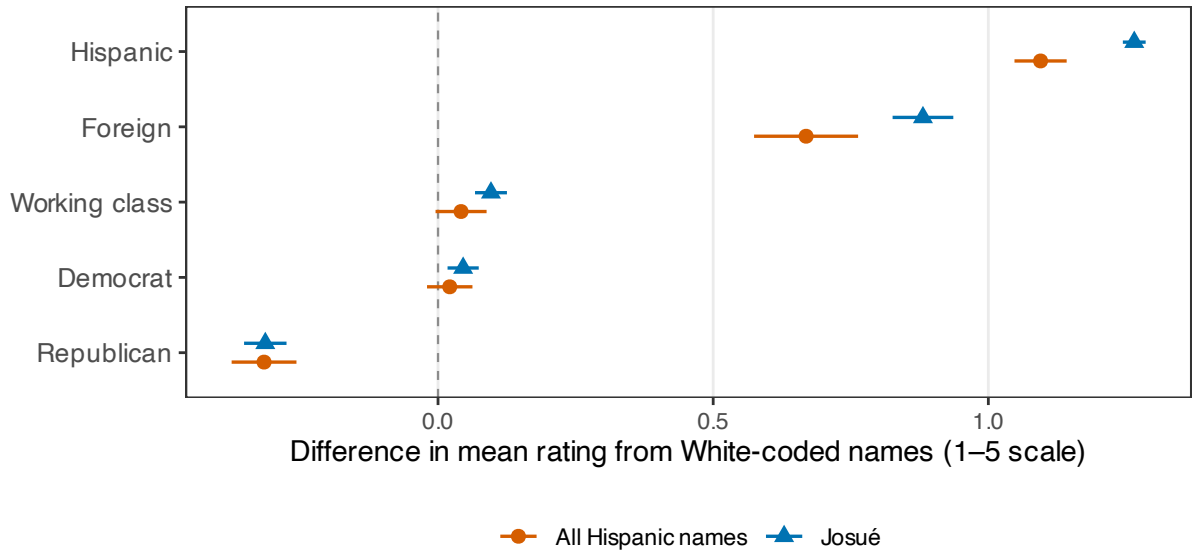


Figure 7: Mean first-name trait ratings from Elder and Hayes (2023) (1–5 scale), as contrasts against White-coded names: all Hispanic-coded names (orange) and "Josué," the name Zárate et al. (2024) use (blue). The ratings pool a separate respondent sample, include "Josué" but not the Anglo name, and compare across names, so the figure shows that the name signals these traits, not how a respondent combines them. Intervals reflect variation across names — for "Josué," the spread of the White-coded comparison names — rather than rater-level sampling.

That the name conveys these traits suggests the cue moves the perceptions rather than leaving them fixed, so the NREC would bundle the movement and depart from the all-else-equal effect. The departure presumes that the name marks foreignness directly rather than that respondents infer foreignness from perceived race (Section 3.3), and the two orderings cannot be told apart, not by ratings from a separate sample and not in principle, since the sequencing occurs inside the respondent's head (Robins and Richardson, 2010). Nor need the orderings be exclusive: The departure follows if any part of the foreignness moves directly with the name.

8.3 The Two Readings Together

The reanalysis returns an answer about coethnic preference that the all-else-equal default conceals. Hispanic respondents do reward the Hispanic candidate, but the reward attaches to the candidate who is Hispanic in the way Hispanic candidates actually are — speaking

native-accented Spanish when they campaign in the language — and the reward disappears when the comparison strips the language away. The common weighting evaluates the Anglo condition mainly at native Spanish, a profile few Spanish-campaigning Anglos fit; the Anglo candidate who speaks native Spanish is the counterpart of the Senegalese Muslim of the introduction, the atypical member of the category Anglo on whom an all-else-equal comparison must lean.

The two regimes reach the same lesson by different routes. What the audit literature might treat as a confound to hold fixed (Section 6.1), and what the cue-stability literature might treat as a failed manipulation (Section 6.2), may be the content of the category itself (Hu and Kohler-Hausmann, 2025). The reanalysis kept the category as constituted and kept causal inference (Section 2): The AEWR and the NREC let language move with the category, and both are well-defined causal estimands, with estimates and confidence intervals from the same experiment.

9 Conclusion

This paper has argued that the literature's defenses of the all-else-equal design bundle an estimand commitment with a recovery condition. Unbundled, the same randomization recovers an entire family of race estimands — each built from counterfactual effects at the unit level, not from disparities across people — and the choice among that family's members is a claim about what a racial category is, not a concession on rigor. The "Muslim effect" of the opening pages was such a choice all along: Isolating religion from origin and language decided, before any data arrived, that the comparison would sit at the categories' edges rather than their centers. The application shows the choice governing the answer: On one experiment, Hispanic voters' coethnic preference is absent under the all-else-equal weighting and clear under the within-race weighting, because the preference lives in the language the category indexes. Three questions concern what the framework asks of empirical research once the estimand is a substantive choice.

The first question is how to choose among the members of the estimand family. The

paper lays out grounds for the choice (Section 5) but does not adjudicate among them, and develops the normative grounds least: The choice of contrast can be grounded in what one takes discrimination to be and in which differences across groups are invidious rather than benign (Hu, 2026), a stance explicit in health-care policy research (Institute of Medicine, 2002; McGuire et al., 2006) but rare in political science. The open task is the pairing itself: which conception of discrimination names which member of the estimand family, so that normative theory sets the target, empirical design recovers it, and each constrains the other rather than proceeding apart.

The second question is what to do with a choice the data cannot settle. A study might report several members of the estimand family — as the application does, setting the AEE beside the AEWR — so readers see how a finding of discrimination depends on the conception of the category, much as a sensitivity analysis displays dependence on an unverifiable assumption. The study might instead commit to a single estimand and defend the conception that estimand embodies. Either path sets practical tasks: recovering the within-group distributions q_z that the within-race and mixed estimands require (Section S1.7), deciding which attributes a category indexes and so belong in $\boldsymbol{v}$ (Section 3), measuring perceived categories in the perception-based regime, and carrying the estimands to observational studies, where identification assumptions replace the randomization this paper assumes.

The third question reaches beyond discrimination to political representation (Mansbridge, 1999; Pitkin, 1967). When a representative is said to represent a racial group, is the group fixed by a category label or by the broader profile the label indexes? The question becomes concrete in defining the group's preference, for example the median opinion against which a legislator's votes are judged (Lax et al., 2019; Miller and Stokes, 1963; Rivera-Burgos, 2025): Is the median taken over the racial group with income and the rest held fixed, or over the group as constituted? No design can dissolve the choice. Some individuals, for example, are both Black and middle-class, so a counterfactual shift in Black opinion moves middle-class opinion with it; unlike race and income on a résumé, which a study can pair freely, the two opinions cannot be paired at will. The estimand family

gives the choice precise form, with legislators as the decision-makers, roll-call votes as the decisions, and the constituency as the configuration, where z encodes a position of the racial group of interest and $\boldsymbol{v}$ the analogous statistics for other groups.

This paper does not answer these questions. Its contribution is a framework for taking them up: The framework translates contested conceptions of race and other identity categories into distinct causal estimands that empirical studies can reliably recover. Each contested conception can then be combined with rigorous causal evidence rather than debated without it.

References

Abrajano, Marisa (2010). *Campaigning to the New American Electorate: Advertising to Latino Voters*. Stanford, CA: Stanford University Press.

Acharya, Avidit, Matthew Blackwell, and Maya Sen (2018). "Analyzing Causal Mechanisms in Survey Experiments". In: *Political Analysis* 26.4, pp. 357–378.

Adida, Claire L., David D. Laitin, and Marie-Anne Valfort (2010). "Identifying Barriers to Muslim Integration in France". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.52, pp. 22384–22390.

Alexander, Larry (1992). "What Makes Wrongful Discrimination Wrong? Biases, Preferences, Stereotypes, and Proxies". In: *University of Pennsylvania Law Review* 141, pp. 149–219.

Andreasen, Robin O. (1998). "A New Perspective on the Race Debate". In: *The British Journal for the Philosophy of Science* 49.2, pp. 199–225.

— (2000). "Race: Biological Reality or Social Construct?" In: *Philosophy of Science* 67.S3: Supplement, S653–S666.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables". In: *Journal of the American Statistical Association* 91.434, pp. 444–455.

Angrist, Joshua D. and Jörn-Steffen Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Appiah, Kwame Anthony (2005). *The Ethics of Identity*. Princeton, NJ: Princeton University Press.

Appiah, Kwame Anthony and Amy Gutmann (1998). *Color Conscious: The Political Morality of Race*. Princeton, NJ: Princeton University Press.

Arrow, Kenneth Joseph (1973). "The Theory of Discrimination". In: *Discrimination in Labor Markets*. Ed. by Orley Ashenfelter and Albert Rees. Princeton, NJ: Princeton University Press, pp. 3–33.

Barreto, Matt A. (2007). " ¡Sí Se Puede! Latino Candidates and the Mobilization of Latino Voters". In: *The American Political Science Review* 101.3, pp. 425–441.

Becker, Gary S. (1957). *The Economics of Discrimination*. Chicago, IL: The University of Chicago Press.

Bertrand, Marianne and Sendhil Mullainathan (2004). "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination". In: *American Economic Review* 94.4, pp. 991–1013.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys (2019). "Declaring and Diagnosing Research Designs". In: *The American Political Science Review* 113.3, pp. 838–859.

Butler, Daniel M. and David E. Broockman (2011). "Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators". In: *American Journal of Political Science* 55.3, pp. 463–477.

Butler, Daniel M. and Charles Crabtree (2021). "Audit Studies in Political Science". In: *Advances in Experimental Political Science*. Ed. by James N. Druckman and Donald P. Green. New York, NY: Cambridge University Press. Chap. 3, pp. 42–55.

Butler, Daniel M. and Jonathan Homola (2017). "An Empirical Justification for the Use of Racially Distinctive Names to Signal Race in Experiments". In: *Political Analysis* 25.1, pp. 122–130.

Chandra, Kanchan (2006). "What is Ethnic Identity and Does It Matter?" In: *Annual Review of Political Science* 9.1, pp. 397–424.

— (2012). "What Is Ethnic Identity? A Minimalist Definition". In: *Constructivist Theories of Ethnic Politics*. Ed. by Kanchan Chandra. New York, NY: Oxford University Press. Chap. 2, pp. 51–96.

Chavez, Leo R. (2013). *The Latino Threat: Constructing Immigrants, Citizens, and the Nation*. 2nd. Palo Alto, CA: Stanford University Press.

Cochran, William G. (1965). "The Planning of Observational Studies of Human Populations". In: *Journal of the Royal Statistical Society. Series A (General)* 128.2, pp. 234–266.

Cook, Benjamin Lê, Thomas G. McGuire, Ellen Meara, and Alan M. Zaslavsky (2009a). "Adjusting for Health Status in Non-linear Models of Health Care Disparities". In: *Health Services and Outcomes Research Methodology* 9.1, pp. 1–21.

Cook, Benjamin Lê, Thomas G. McGuire, and Alan M. Zaslavsky (2012). "Measuring Racial/Ethnic Disparities in Health Care: Methods and Practical Issues". In: *Health Services Research* 47.3, pt 2, pp. 1232–1254.

Cook, Benjamin Lê, Thomas G. McGuire, and Samuel H. Zuvekas (2009b). "Measuring Trends in Racial/ Ethnic Health Care Disparities". In: *Medical Care Research and Review* 66.1, pp. 23–48.

Cornell, Stephen E. and Douglas Hartmann (1997). *Ethnicity and Race: Making Identities in a Changing World*. Thousand Oaks, CA: SAGE Publications.

Crabtree, Charles et al. (2023). "Validated Names for Experimental Studies on Race and Ethnicity". In: *Scientific Data* 10.130.

Cross, Harry E., Genevieve Kenney, Jane Mell, and Wendy Zimmermann (1990). *Employer Hiring Practices: Differential Treatment of Hispanic and Anglo Job Seekers*. Washington, D. C.: Urban Institute Press.

Dafoe, Allan, Baobao Zhang, and Devin Caughey (2018). "Information Equivalence in Survey Experiments". In: *Political Analysis* 26.4, pp. 399–416.

Dawson, Michael C. (1994). *Behind the Mule: Race and Class in African-American Politics*. Princeton, NJ: Princeton University Press.

de la Cuesta, Brandon, Naoki Egami, and Kosuke Imai (2022). "Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution". In: *Political Analysis* 30.1, pp. 19–45.

Dembroff, Robin, Issa Kohler-Hausmann, and Elise Sugarman (2020). "What Taylor Swift and Beyoncé Teach Us About Sex and Causes". In: *University of Pennsylvania Law Review Online* 169.1.

DiNardo, John (2007). "Interesting Questions in Freakonomics". In: *Journal of Economic Literature* 45.4, pp. 973–1000.

Diop, A. Moustapha (1988). "Stéréotypes et stratégies dans la communauté musulmane de France". In: *Les musulmans dans la société française*. Ed. by Rémy Leveau and Gilles Kepel. Paris, France: Presses de la Fondation Nationale des Sciences Politiques, pp. 77–87.

Dowling, Julie A. (2014). *Mexican Americans and the Question of Race*. Austin: University of Texas Press.

Duan, Naihua et al. (2008). "Disparities in defining disparities: Statistical conceptual frameworks". In: *Statistics in Medicine* 27.20, pp. 3941–3956.

Eberhardt, Jennifer L., Paul G. Davies, Valerie J. Purdie-Vaughns, and Sheri Lynn Johnson (2006). "Looking Deathworthy: Perceived Stereotypicality of Black Defendants Predicts Capital-Sentencing Outcomes". In: *Psychological Science* 17.5, pp. 383–386.

Elder, Elizabeth Mitchell and Matthew Hayes (2023). "Signaling Race, Ethnicity, and Gender with Names: Challenges and Recommendations". In: *The Journal of Politics* 85.2, pp. 764–770.

Fearon, James D. (2003). "Ethnic and Cultural Diversity by Country". In: *Journal of Economic Growth* 8.2, pp. 195–222.

Fenno, Richard F. (1978). *Home Style: House Members in Their Districts*. Boston, MA: Little, Brown & Co.

Frangakis, Constantine E. and Donald B. Rubin (2002). "Principal Stratification in Causal Inference". In: *Biometrics* 58.1, pp. 21–29.

Fryer, Jr., Roland G. (2019). "An Empirical Analysis of Racial Differences in Police Use of Force". In: *The Journal of Political Economy* 127.3, pp. 1210–1261.

Gaebler, Johann et al. (2022). "A Causal Framework for Observational Studies of Discrimination". In: *Statistics and Public Policy* 9.1, pp. 26–48.

Gärdenfors, Peter (2000). *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: The MIT Press.

— (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. Cambridge, MA: The MIT Press.

Gerber, Alan S. and Donald P. Green (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York, NY: W.W. Norton.

Gooding-Williams, Robert (1998). "Race, Multiculturalism and Democracy". In: *Constellations* 5.1, pp. 18–41.

Greiner, D. James and Donald B. Rubin (2011). "Causal Effects of Perceived Immutable Characteristics". In: *The Review of Economics and Statistics* 93.3, pp. 775–785.

Grogan, Colleen M. and Sunggeun (Ethan) Park (2017). "The Racial Divide in State Medicaid Expansions". In: *Journal of Health Politics, Policy and Law* 42.3, pp. 539–572.

Grogger, Jeffrey and Greg Ridgeway (2006). "Testing for Racial Profiling in Traffic Stops From Behind a Veil of Darkness". In: *Journal of the American Statistical Association* 101.475, pp. 878–887.

Hainmueller, Jens and Daniel J. Hopkins (2015). "The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants". In: *American Journal of Political Science* 59.3, pp. 529–548.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto (2014). "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments". In: *Political Analysis* 22.1, pp. 1–30.

Hardimon, Michael O. (2003). "The Ordinary Concept of Race". In: *Journal of Philosophy* 100.9, pp. 437–455.

— (2017). *Rethinking Race: The Case for Deflationary Realism*. Cambridge, MA: Harvard University Press.

Haslanger, Sally (2000). "Gender and Race: (What) Are They? (What) Do We Want Them to Be?" In: *Noûs* 34.1, pp. 31–55.

Heckman, James J. (1998). "Detecting Discrimination". In: *Journal of Economic Perspectives* 12.2, pp. 101–116.

Heckman, James J. and Peter Siegelman (1993). "The Urban Institute Audit Studies: Their Methods and Findings". In: *Clear and Convincing Evidence: Measurement of Discrimination in America*. Ed. by Michael E. Fix and Raymond J. Struyk. Washington, D. C.: The Urban Institute Press. Chap. 5.

Hellman, Deborah (2008). *When Is Discrimination Wrong?* Cambridge, MA: Harvard University Press.

Holland, Paul W. (1986). "Statistics and Causal Inference". In: *Journal of the American Statistical Association* 81.396, pp. 945–960.

— (2001). "The False Linking of Race and Causality: Lessons from Standardized Testing". In: *Race and Society* 4.2, pp. 219–233.

— (2003). "Causation and Race". In: *ETS Research Report Series* 2003.1.

Hopkins, Daniel J. (2015). "The Upside of Accents: Language, Inter-group Difference, and Attitudes toward Immigration". In: *British Journal of Political Science* 45.3, pp. 531–557.

Hu, Lily (2023). "What is 'Race' in Algorithmic Discrimination on the Basis of Race?" In: *Journal of Moral Philosophy* 21.1-2, pp. 1–26.

— (2026). "Normative Facts and Causal Structure". In: *The Journal of Philosophy*.

Hu, Lily and Issa Kohler-Hausmann (2025). "What is Perceived When Race is Perceived and Why It Matters for Causal Inference and Discrimination Studies". In: *Law & Society Review* 59.2, pp. 239–264.

Imbens, Guido W. and Joshua D. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects". In: *Econometrica* 62.2, pp. 467–475.

Imbens, Guido W. and Donald B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY: Cambridge University Press.

Institute of Medicine (2002). *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. Ed. by Brian D. Smedley, Adrienne Y. Stith, and Alan R. Nelson. Washington, D. C.: The National Academies Press.

Jackson, John W. (2021). "Meaningful Causal Decompositions in Health Equity Research: Definition, Identification, and Estimation Through a Weighting Framework". In: *Epidemiology* 32.2, pp. 282–290.

Knox, Dean, Will Lowe, and Jonathan Mummolo (2020). "Administrative Records Mask Racially Biased Policing". In: *The American Political Science Review* 114.3, pp. 619–637.

Kohler-Hausmann, Issa (2019). "Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination". In: *Northwestern University Law Review* 113.5, pp. 1163–1228.

Kohler-Hausmann, Issa and Robin Dembroff (2022). "Supreme Confusion About Causality at the Supreme Court". In: *City University of New York Law Review* 25.1, pp. 57–92.

Landgrave, Michelangelo and Nicholas Weller (2022). "Do Name-Based Treatments Violate Information Equivalence? Evidence from a Correspondence Audit Experiment". In: *Political Analysis* 30.1, pp. 142–148.

Lax, Jeffrey R., Justin H. Phillips, and Adam Zelizer (2019). "The Party or the Purse? Unequal Representation in the US Senate". In: *The American Political Science Review* 113.4, pp. 917–940.

Leavitt, Thomas and Viviana Rivera-Burgos (2024). "Audit Experiments of Racial Discrimination and the Importance of Symmetry in Exposure to Cues". In: *Political Analysis* 32.4, pp. 445–462.

— (2026). "Navigating the Mismeasurement of Intermediary Variables in Message-Based Experiments". In: *Political Science Research and Methods*. First View, pp. 1–14.

Lippert-Rasmussen, Kasper (2011). "'We are all Different': Statistical Discrimination and the Right to be Treated as an Individual". In: *The Journal of Ethics* 15.1/2, pp. 47–59.

Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart (2021). "What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory". In: *American Sociological Review* 86.3, pp. 532–565.

Ma, Debbie S., Joshua Correll, and Bernd Wittenbrink (2015). "The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data". In: *Behavior Research Methods* 47.4, pp. 1122–1135.

Maddox, Keith B. (2004). "Perspectives on Racial Phenotypicality Bias". In: *Personality and Social Psychology Review* 8.4, pp. 383–401.

Mansbridge, Jane (1999). "Should Blacks Represent Blacks and Women Represent Women? A Contingent "Yes"". In: *The Journal of Politics* 61.3, pp. 628–657.

Masuoka, Natalie and Jane Junn (2013). *The Politics of Belonging: Race, Public Opinion, and Immigration*. Chicago: University of Chicago Press.

Mcdermott, Monika L. (1998). "Race and Gender Cues in Low-Information Elections". In: *Political Research Quarterly* 51.4, pp. 895–918.

McGuire, Thomas G. et al. (2006). "Implementing the Institute of Medicine Definition of Disparities: An Application to Mental Health Care". In: *Health Services Research* 41.5, pp. 1979–2005.

Michener, Jamila (2018). *Fragmented Democracy: Medicaid, Federalism, and Unequal Politics*. New York, NY: Cambridge University Press.

— (2020). "Race, Politics, and the Affordable Care Act". In: *Journal of Health Politics, Policy and Law* 45.4, pp. 547–566.

Miller, Warren E. and Donald E. Stokes (1963). "Constituency Influence in Congress". In: *The American Political Science Review* 57.1, pp. 45–56.

Mills, Charles W. (1998). "'But What Are You Really?': The Metaphysics of Race". In: *Blackness Visible: Essays on Philosophy and Rac*. Ed. by Charles W. Mills. Ithaca, NY: Cornell University Press. Chap. 3, pp. 41–66.

Monk, Jr., Ellis P. (2022). "Inequality without Groups: Contemporary Theories of Categories, Intersectional Typicality, and the Disaggregation of Difference". In: *Sociological Theory* 40.1, pp. 3–27.

Ocampo-Roland, Angie N. (2025). "Group Prototypicality and Boundary Definition: Comparing White and Black Perceptions of Whether Latinos Are American". In: *American Political Science Review* 119.4, pp. 1581–1598.

Olvera, Javier G., Candis Watts Smith, and Nadia D. von Lockette (2023). "The Effect of the Affordable Care Act and Racial Dynamics on Federal Medicaid Transfers". In: *Journal of Public Policy* 43.3, pp. 533–555.

Pager, Devah (2003). "The Mark of a Criminal Record". In: *The American Journal of Sociology* 108.5, pp. 937–975.

Pashley, Nicole E. (2022). "Note on the Delta Method for Finite Population Inference with Applications to Causal Inference". In: *Statistics & Probability Letters* 188, p. 109540.

Pearl, Judea (2001). "Direct and Indirect Effects". In: *UAI'01: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Ed. by Jack Breese and Daphne Koller. Association for Computing Machinery. San Francisco, CA: Morgan Kaufmann Publishers, pp. 411–420.

Phelps, Edmund S. (1972). "The Statistical Theory of Racism and Sexism". In: *The American Economic Review* 62.4, pp. 659–661.

Pierson, Emma et al. (2020). "A Large-Scale Analysis of Racial Disparities in Police Stops across the United States". In: *Nature Human Behaviour* 4, pp. 736–745.

Piper, Adrian (1992). "Passing for White, Passing for Black". In: *Transition* 58, pp. 4–32.

Pitkin, Hanna Fenichel (1967). *The Concept of Representation*. Berkeley, CA: University of California Press.

Purnell, Thomas, William Idsardi, and John Baugh (1999). "Perceptual and Phonetic Experiments on American English Dialect Identification". In: *Journal of Language and Social Psychology* 18.1, pp. 10–30.

Quillian, Lincoln, Devah Pager, Ole Hexel, and Arnfinn H. Midtbøen (2017). "Meta-analysis of field experiments shows no change in racial discrimination in hiring over time". In: *Proceedings of the National Academy of Sciences of the United States of America* 114.1, pp. 364–377.

Rivera-Burgos, Viviana (2025). "Do Legislators Represent Racial Minority Opinions? Evidence on Disparities at the Congressional District Level". In: *Journal of Race, Ethnicity, and Politics*. FirstView.

Robins, James M. and Sander Greenland (1992). "Identifiability and Exchangeability for Direct and Indirect Effects". In: *Epidemiology* 3.2, pp. 143–155.

Robins, James M. and Thomas S. Richardson (2010). "Alternative Graphical Causal Models and the Identification of Direct Effects". In: *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*. Ed. by Patrick E. Shrout, Katherine M. Keyes, and Katherine Ornstein. Oxford: Oxford University Press, pp. 103–158.

Rosa, Jonathan (2018). *Looking Like a Language, Sounding Like a Race: Raciolinguistic Ideologies and the Learning of Latinidad*. New York, NY: Oxford University Press.

Rosch, Eleanor H. (1973). "Natural Categories". In: *Cognitive Psychology* 4.3, pp. 328–350.

— (1975). "Cognitive Representations of Semantic Categories". In: *Journal of Experimental Psychology: General* 104.3, pp. 192–233.

Rosch, Eleanor H. and Carolyn B. Mervis (1975). "Family resemblances: Studies in the internal structure of categories". In: *Cognitive Psychology* 7.4, pp. 573–605.

Rosenbaum, Paul R. and Donald B. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects". In: *Biometrika* 70.1, pp. 41–55.

Rubin, Donald B. (2008). "For Objective Causal Inference, Design Trumps Analysis". In: *The Annals of Applied Statistics* 2.3, pp. 808–840.

Sanchez, Diana T., George Chavez, Jessica J. Good, and Leigh S. Wilton (2012). "The Language of Acceptance: Spanish Proficiency and Perceived Intragroup Rejection Among Latinos". In: *Journal of Cross-Cultural Psychology* 43.6, pp. 1019–1033.

Sanchez, Gabriel R. (2008). "Latino Group Consciousness and Perceptions of Commonality with African Americans". In: *Social Science Quarterly* 89.2, pp. 428–444.

Sanchez, Gabriel R. and Natalie Masuoka (2010). "Brown-Utility Heuristic? The Presence and Contributing Factors of Latino Linked Fate". In: *Hispanic Journal of Behavioral Sciences* 32.4, pp. 519–531.

Sen, Maya and Omar Wasow (2016). "Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics". In: *Annual Review of Political Science* 19.1, pp. 499–522.

Spencer, Quayshawn (2014). "A Radical Solution to the Race Problem". In: *Philosophy of Science* 81.5, pp. 1025–1038.

Taylor, Paul C. (2003). *Race: A Philosophical Introduction*. Cambridge, UK: Polity Press.

Terkildsen, Nayda (1993). "When White Voters Evaluate Black Candidates: The Processing Implications of Candidate Skin Color, Prejudice, and Self-Monitoring". In: *American Journal of Political Science* 37.4, pp. 1032–1053.

VanderWeele, Tyler J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York, NY: Oxford University Press.

VanderWeele, Tyler J. and Whitney R. Robinson (2014). "On the Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables". In: *Epidemiology* 25.4, pp. 473–484.

Weber, Max (1978). "Bureaucracy". In: *Economy and Society: An Outline of Interpretive Sociology*. Ed. by Guenther Roth and Claus Wittich. Vol. 2. Berkeley, CA: University of California Press. Chap. XI, pp. 956–1005.

Whitehead, Margaret (1992). "The Concepts and Principles of Equity in Health". In: *International Journal of Health Services* 22.3, pp. 429–445.

Wittgenstein, Ludwig (1953). *Philosophical Investigations*. Oxford, UK: Blackwell Publishers.

Zárate, Marques G., Enrique Quezada-Llanes, and Angel D. Armenta (2024). "Se Habla Español: Spanish-Language Appeals and Candidate Evaluations in the United States". In: *The American Political Science Review* 118.1, pp. 363–379.

Zou, Linda X. and Sapna Cheryan (2017). "Two Axes of Subordination: A New Model of Racial Position". In: *Journal of Personality and Social Psychology* 112.5, pp. 696–717.

Zuberi, Tukufu (2001). *Thicker Than Blood: How Racial Statistics Lie*. Minneapolis, MN: University of Minnesota Press.

Supplementary Material for "Which Effect of Race? Causal Inference without Holding All Else Equal"

Thomas Leavitt

Contents

S1 Formal Results

S1.1 From decision-makers and decisions to units

Section 3 of the main text indexes units directly and points here for the underlying structure. There is a set of decision-makers indexed by $j \in \{1, \ldots, J\}$, and each decision-maker j faces K_j decisions indexed by $k \in \{1, \ldots, K_j\}$, giving $N \coloneqq \sum_j K_j$ decision-maker–decision pairs. Under the exposure-based regime, decision-maker j in decision k encounters a configuration $\boldsymbol{t}_{jk} \in \mathcal{T}$. Let $\boldsymbol{t} \coloneqq (\boldsymbol{t}_{11}, \ldots, \boldsymbol{t}_{JK_J}) \in \mathcal{T}^N$ stack these configurations into a *configuration assignment*. In principle, the potential outcome of the pair (j, k) may depend on the entire assignment.

Assumption S1 (SUTVA)**.** *For all decision-makers j, decisions k, and configuration assignments $\boldsymbol{t}, \boldsymbol{t}' \in \mathcal{T}^N$, $y_{jk}(\boldsymbol{t}) = y_{jk}(\boldsymbol{t}')$ whenever $\boldsymbol{t}_{jk} = \boldsymbol{t}'_{jk}$.*

Assumption S1 rules out interference between units and carryover across a decision-maker's own decisions, encompassing the "stability and no carryover effects" assumption of Hain-

mueller et al. (2014, p. 8). This assumption licenses reindexing the pairs (j,k) as units $i \in \{1,\ldots,N\}$ whose potential outcomes depend only on their own configurations, defining the function $y_i : \mathcal{T} \to \mathbb{R}$ used throughout the main text.

Under the perception-based regime, only the cue is assigned. Let $\boldsymbol{w} \coloneqq (w_{11},\ldots,w_{JK_J}) \in \{0,1\}^N$ stack the cues into a *cue assignment*. The analogue of Assumption S1 applies to the response, the racial perception, and the nonracial perceptions at once.

Assumption S2 (SUTVA for the cue)**.** *For all decision-makers j, decisions k, and cue assignments $\boldsymbol{w}, \boldsymbol{w}' \in \{0,1\}^N$,*

$$y_{jk}(\boldsymbol{w}) = y_{jk}(\boldsymbol{w}'), \quad z_{jk}(\boldsymbol{w}) = z_{jk}(\boldsymbol{w}'), \quad \textit{and} \quad \boldsymbol{v}_{jk}(\boldsymbol{w}) = \boldsymbol{v}_{jk}(\boldsymbol{w}')$$

whenever $w_{jk} = w'_{jk}$.

Reindexing as before yields the unit-level potential outcomes $z_i(w)$, $\boldsymbol{v}_i(w)$, and $y_i(w)$ of Section 3.3.

S1.2 The gap between two estimands

A single identity, stated as (S1) in Proposition S1 below, gives the gap between any two members of the estimand family. The identity is purely algebraic: Assumption S1 makes the estimands well defined, but the identity itself invokes no assumption beyond the estimands' definitions.

Proposition S1 (Gap between two members of the estimand family)**.** *Write*

$$\theta(g_0, g_1) \coloneqq \frac{1}{N} \sum_{i=1}^{N} \sum_{\boldsymbol{v} \in \mathcal{V}} \left[y_i(1,\boldsymbol{v})\, g_1(\boldsymbol{v}) - y_i(0,\boldsymbol{v})\, g_0(\boldsymbol{v}) \right]$$

for the member of the estimand family that weights nonracial profiles by g_0 under racial condition $z = 0$ and by g_1 under racial condition $z = 1$ — the estimand targeted by the plug-in estimator of the estimand family, (9) *of the main text — and $\bar{y}(z,\boldsymbol{v}) \coloneqq (1/N)\sum_{i=1}^{N} y_i(z,\boldsymbol{v})$. For any two members of the estimand family,*

$$\theta(g_0, g_1) - \theta(g'_0, g'_1) = \sum_{\boldsymbol{v} \in \mathcal{V}} \left\{ \bar{y}(1,\boldsymbol{v}) \left[g_1(\boldsymbol{v}) - g'_1(\boldsymbol{v}) \right] - \bar{y}(0,\boldsymbol{v}) \left[g_0(\boldsymbol{v}) - g'_0(\boldsymbol{v}) \right] \right\}. \tag{S1}$$

Proof. The proof subtracts the two estimands term by term and then exchanges the order of the two finite sums, so that each difference of weights multiplies an average potential outcome.

Applying the definition of θ to each pair of weights,

$$\begin{aligned}\theta(g_0, g_1) - \theta(g_0', g_1') &= \frac{1}{N}\sum_{i=1}^{N}\sum_{\boldsymbol{v}\in\mathcal{V}}\left[y_i(1,\boldsymbol{v})\, g_1(\boldsymbol{v}) - y_i(0,\boldsymbol{v})\, g_0(\boldsymbol{v})\right] \\ &\quad - \frac{1}{N}\sum_{i=1}^{N}\sum_{\boldsymbol{v}\in\mathcal{V}}\left[y_i(1,\boldsymbol{v})\, g_1'(\boldsymbol{v}) - y_i(0,\boldsymbol{v})\, g_0'(\boldsymbol{v})\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{\boldsymbol{v}\in\mathcal{V}}\left\{y_i(1,\boldsymbol{v})\left[g_1(\boldsymbol{v}) - g_1'(\boldsymbol{v})\right] - y_i(0,\boldsymbol{v})\left[g_0(\boldsymbol{v}) - g_0'(\boldsymbol{v})\right]\right\}\end{aligned}$$

where the second equality combines the two sums term by term and, within each $(i, \boldsymbol{v})$ term, factors $y_i(1, \boldsymbol{v})$ and $y_i(0, \boldsymbol{v})$ out of their respective differences of weights. Both index sets are finite, so the sums over i and $\boldsymbol{v}$ may be exchanged, and the average $(1/N)\sum_i$ distributes over the potential outcomes because the differences of weights do not depend on i:

$$\begin{aligned}\theta(g_0, g_1) - \theta(g_0', g_1') = \sum_{\boldsymbol{v}\in\mathcal{V}}\Bigg\{ &\left[\frac{1}{N}\sum_{i=1}^{N} y_i(1,\boldsymbol{v})\right]\left[g_1(\boldsymbol{v}) - g_1'(\boldsymbol{v})\right] \\ &- \left[\frac{1}{N}\sum_{i=1}^{N} y_i(0,\boldsymbol{v})\right]\left[g_0(\boldsymbol{v}) - g_0'(\boldsymbol{v})\right]\Bigg\}.\end{aligned}$$

Substituting $\bar{y}(z, \boldsymbol{v}) = (1/N)\sum_i y_i(z, \boldsymbol{v})$ then yields the identity (S1). □

The difference between the AEWR and the AEE is the instance of the identity (S1) with $(g_0, g_1) = (q_0, q_1)$ and $(g_0', g_1') = (\rho, \rho)$. A mixed contrast enters the identity (S1) through the mixed weights $g_z(\boldsymbol{v}) = \rho(\boldsymbol{v}_{\text{fix}})\, q_z(\boldsymbol{v}_{\text{free}} \,|\, \boldsymbol{v}_{\text{fix}})$, stated as (S5) in Section S1.5. Substituting the mixed weights on one side of (S1) gives the gap between a mixed effect and the AEE or the AEWR, and substituting the mixed weights of two different partitions on the two sides gives the gap between two mixed effects.

Two members of the estimand family coincide when the right-hand side of the identity (S1) vanishes. The right-hand side vanishes when the weights agree, $g_z = g_z'$ for $z \in \{0, 1\}$,

and also when $\bar{y}(z, \boldsymbol{v})$ does not depend on $\boldsymbol{v}$. Writing $\bar{y}(z)$ for the common value and factoring the common value out of the sum,

$$\sum_{\boldsymbol{v}\in\mathcal{V}} \bar{y}(z,\boldsymbol{v})\,[g_z(\boldsymbol{v}) - g'_z(\boldsymbol{v})] = \bar{y}(z)\left[\sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v}) - \sum_{\boldsymbol{v}\in\mathcal{V}} g'_z(\boldsymbol{v})\right] = \bar{y}(z)\,(1-1) = 0,$$

because g_z and g'_z are probability distributions on $\mathcal{V}$ and therefore each normalizes to one: $\sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v}) = \sum_{\boldsymbol{v}\in\mathcal{V}} g'_z(\boldsymbol{v}) = 1$.

The AEWR-versus-AEE pairing gives the cleanest reading of the gap as a failure of two operations to commute, differencing across racial conditions and averaging over nonracial profiles: The AEE differences at each shared profile and then averages over ρ, whereas the AEWR averages each racial condition over its own conditional q_z and then differences the two race-specific averages. The two estimands agree exactly when the two operations commute.

S1.3 The double duty of a profile-blind assignment

Section 6.1 of the main text states Proposition 1 and points here for the proof of part *(ii)*; part *(i)* is the unbiasedness of the plug-in estimator for the targeted member of the estimand family, proved in Section S1.8.

Proof of part (ii) of Proposition 1. The weighting read off the assignment is $g_z(\boldsymbol{v}) \coloneqq \Pr(\boldsymbol{V} = \boldsymbol{v} \mid Z = z)$ for each racial condition $z \in \{0, 1\}$, and profile-blindness (Definition 1) is the condition that $\Pr(Z = z \mid \boldsymbol{V} = \boldsymbol{v})$ does not vary with $\boldsymbol{v}$. The weightings are well defined because of the proposition's hypothesis that $p(z, \boldsymbol{v}) > 0$ on every profile with positive weight under the target. The target's weights are probability distributions, so each racial condition contains at least one profile with positive weight, and therefore $\Pr(Z = z) = \sum_{\boldsymbol{v}} p(z, \boldsymbol{v}) > 0$ for each racial condition z. Conditioning on $\boldsymbol{V} = \boldsymbol{v}$ is understood at profiles with $\Pr(\boldsymbol{V} = \boldsymbol{v}) > 0$, and at profiles with $\Pr(\boldsymbol{V} = \boldsymbol{v}) = 0$ both weightings place zero weight, $g_1(\boldsymbol{v}) = g_0(\boldsymbol{v}) = 0$.

Part *(ii)* asserts an equivalence, and the proof establishes the equivalence one implica-

tion at a time. The first implication is that profile-blindness forces the weightings read off the two racial conditions to coincide; the second is that coinciding weightings force profile-blindness. Both implications rest on two elementary facts. The first fact is that a joint probability factors through either of its conditionals,

$$\Pr(A, B) = \Pr(A \mid B)\Pr(B) = \Pr(B \mid A)\Pr(A),$$

which is the definition of conditional probability rearranged; the proof below refers to the factorization as the multiplication rule. The second fact is the law of total probability. Each implication also establishes, along the way, the same intermediate fact — derived within each argument, not presupposed — that a weighting common to the two racial conditions can only be the marginal profile distribution $\Pr(\boldsymbol{V} = \boldsymbol{v})$.

Suppose first that the assignment is profile-blind, so that $\Pr(Z = z \mid \boldsymbol{V} = \boldsymbol{v})$ equals a constant c_z at every profile with $\Pr(\boldsymbol{V} = \boldsymbol{v}) > 0$. By the law of total probability, the constant equals the marginal probability of the racial condition,

$$\Pr(Z = z) = \sum_{\boldsymbol{v}} \Pr(Z = z \mid \boldsymbol{V} = \boldsymbol{v})\Pr(\boldsymbol{V} = \boldsymbol{v}) = c_z \sum_{\boldsymbol{v}} \Pr(\boldsymbol{V} = \boldsymbol{v}) = c_z.$$

Then, at every profile with $\Pr(\boldsymbol{V} = \boldsymbol{v}) > 0$, writing the joint probability by the multiplication rule in each of its two orders,

$$\begin{aligned} g_z(\boldsymbol{v}) &= \frac{\Pr(Z = z, \boldsymbol{V} = \boldsymbol{v})}{\Pr(Z = z)} = \frac{\Pr(Z = z \mid \boldsymbol{V} = \boldsymbol{v})\Pr(\boldsymbol{V} = \boldsymbol{v})}{\Pr(Z = z)} \\ &= \frac{c_z \Pr(\boldsymbol{V} = \boldsymbol{v})}{c_z} = \Pr(\boldsymbol{V} = \boldsymbol{v}), \end{aligned}$$

where the constant c_z cancels between the numerator and the denominator. The right-hand side does not depend on z, so $g_1(\boldsymbol{v}) = g_0(\boldsymbol{v}) = \Pr(\boldsymbol{V} = \boldsymbol{v})$ at every profile. The weightings read off the two racial conditions therefore coincide, and the common weighting equals the marginal profile distribution, which is the intermediate fact.

Suppose conversely that the weighting read off the assignment is common across the two

racial conditions, $g_1 = g_0 =: \rho$. The law of total probability over the two racial conditions shows that the common weighting ρ is the marginal profile distribution,

$$\Pr(\boldsymbol{V} = \boldsymbol{v}) = \sum_{z \in \{0,1\}} \Pr(\boldsymbol{V} = \boldsymbol{v} \mid Z = z) \Pr(Z = z) = \rho(\boldsymbol{v}) \sum_{z \in \{0,1\}} \Pr(Z = z) = \rho(\boldsymbol{v}),$$

where the middle equality factors $\rho(\boldsymbol{v})$ out of the sum because the common weighting does not depend on the racial condition z. Then, at every profile with $\Pr(\boldsymbol{V} = \boldsymbol{v}) > 0$, again writing the joint probability by the multiplication rule in each of its two orders,

$$\begin{aligned} \Pr(Z = z \mid \boldsymbol{V} = \boldsymbol{v}) &= \frac{\Pr(Z = z, \boldsymbol{V} = \boldsymbol{v})}{\Pr(\boldsymbol{V} = \boldsymbol{v})} = \frac{g_z(\boldsymbol{v}) \Pr(Z = z)}{\Pr(\boldsymbol{V} = \boldsymbol{v})} \\ &= \frac{\rho(\boldsymbol{v}) \Pr(Z = z)}{\rho(\boldsymbol{v})} = \Pr(Z = z), \end{aligned}$$

where the common weight $\rho(\boldsymbol{v})$ cancels between the numerator and the denominator. The right-hand side does not vary with $\boldsymbol{v}$, so the assignment is profile-blind. □

For a researcher who reads the estimand's weights off the assignment, as part *(ii)* supposes, the intermediate fact of the proof has a design consequence. Each weighting $g_z(\boldsymbol{v}) = \Pr(\boldsymbol{V} = \boldsymbol{v} \,|\, Z = z)$ is the distribution of nonracial profiles within racial condition z, and the marginal profile distribution mixes the two weightings in proportion to the probabilities of the racial conditions,

$$\Pr(\boldsymbol{V} = \boldsymbol{v}) = g_0(\boldsymbol{v}) \Pr(Z = 0) + g_1(\boldsymbol{v}) \Pr(Z = 1),$$

by the law of total probability, because each configuration falls under exactly one of the two racial conditions. When $g_1 = g_0 = \rho$, the mixture of the two equal weightings is

$$\rho(\boldsymbol{v}) \left[\Pr(Z = 0) + \Pr(Z = 1)\right] = \rho(\boldsymbol{v}),$$

so the common weighting coincides with the marginal profile distribution — the distribution written as ρ in Figure 3 of the main text. For the researcher who reads the estimand's

weights off the assignment, selecting the common weighting ρ and selecting the assignment's marginal distribution over nonracial profiles are therefore one choice, not two.

S1.4 Identification of the NREC in the perception-based regime

Section 6.2 of the main text states Proposition 2 and Proposition 3 and points here for the proofs and for the two standard conditions on how the cue would move each unit's perceived race, that is, on the potential outcomes $z_i(0)$ and $z_i(1)$ of the cue.

The first condition rules out defiers, the units that would invert the cue. With Black and White as the two racial conditions of the running example, a defier would perceive the Black cue as White and the White cue as Black.

Assumption S3 (No defiers)**.** *For all units* $i \in \{1, \ldots, N\}$*,* $(z_i(0), z_i(1)) \neq (1, 0)$*.*

To illustrate what Assumption S3 permits and what the assumption rules out, consider the name-based racial cue of the running example, with DeShawn coded to read as Black and Connor as White (Section 3.3 of the main text). A decision-maker might read "Connor" as Black rather than White, but would then plausibly read "DeShawn" as Black as well; a decision-maker who read "DeShawn" as White would likewise plausibly read "Connor" as White. Either decision-maker would fail to register the cue's intended racial contrast, a failure that Assumption S3 permits; the assumption rules out only the decision-maker who would invert the contrast. Once defiers are ruled out, the denominator of (7) of the main text — the cue's average effect on perceived race — equals the share of compliers, with no defiers to offset them.

The second condition is a positive complier share. The cue must move at least one unit's perceived race in the direction the cue intends — in the running example, toward White under the White cue and toward Black under the Black cue — and movement alone does not suffice, because a defier's perceived race also moves, in the opposite direction.

Assumption S4 (Existence of compliers)**.** *At least one of the* $i \in \{1, \ldots, N\}$ *units has* $z_i(0) = 0$ *and* $z_i(1) = 1$*, i.e.,* $\mathcal{C} \neq \emptyset$*.*

Without such a complier the complier share is zero; the denominator of (7) then vanishes and the NREC, an average over $\mathcal{C}$, is undefined.

Under Assumption S3 and Assumption S4, together with the hypotheses that Proposition 2 states in the main text — the cue-level SUTVA (Assumption S2), perception sufficiency, and Assumption 2 — the proposition follows.

Proof of Proposition 2. The proof shows that non-compliers contribute zero to both the numerator and the denominator of the ratio in (7), so that the numerator and the denominator each reduce to a sum over compliers, and the ratio of those two sums is the NREC.

By Assumption S3, each unit belongs to one of three types: a complier ($z_i(0) = 0, z_i(1) = 1$), an always-taker ($z_i(0) = z_i(1) = 1$), or a never-taker ($z_i(0) = z_i(1) = 0$). For non-compliers, $z_i(0) = z_i(1)$, and by Assumption 2, $\boldsymbol{v}_i(0) = \boldsymbol{v}_i(1)$ as well, so $y_i\left(z_i(1), \boldsymbol{v}_i(1)\right) - y_i\left(z_i(0), \boldsymbol{v}_i(0)\right) = 0$ and $z_i(1) - z_i(0) = 0$ for every non-complier. The numerator and denominator of (7) therefore reduce to sums over compliers:

$$\frac{1}{N}\sum_{i=1}^{N}\left[y_i\left(z_i(1), \boldsymbol{v}_i(1)\right) - y_i\left(z_i(0), \boldsymbol{v}_i(0)\right)\right] = \frac{1}{N}\sum_{i\in\mathcal{C}}\left[y_i\left(1, \boldsymbol{v}_i(1)\right) - y_i\left(0, \boldsymbol{v}_i(0)\right)\right],$$
$$\frac{1}{N}\sum_{i=1}^{N}\left[z_i(1) - z_i(0)\right] = \frac{|\mathcal{C}|}{N}.$$

By Assumption S4, $|\mathcal{C}| \geq 1$, so the denominator $|\mathcal{C}|/N$ is positive and the ratio is well defined. Taking the ratio of the two sums over compliers,

$$\frac{\frac{1}{N}\sum_{i\in\mathcal{C}}\left[y_i\left(1, \boldsymbol{v}_i(1)\right) - y_i\left(0, \boldsymbol{v}_i(0)\right)\right]}{|\mathcal{C}|/N} = \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\left[y_i\left(1, \boldsymbol{v}_i(1)\right) - y_i\left(0, \boldsymbol{v}_i(0)\right)\right],$$

and applying the definition of the NREC in (6) of the main text yields the result. □

The numerator of (7) — the cue's average effect on the outcome over all N units — is a lower bound in magnitude on the NREC. In practice, many cue-based studies cannot compute the ratio: The denominator — the cue's average effect on perceived race — requires perceived race to be measured on the decision-makers who supply the outcome,

and most studies measure perceived race, if at all, in a separate sample (Elder and Hayes, 2023; Gaddis, 2017; Landgrave and Weller, 2022). Such studies can estimate the numerator alone, and the numerator alone is still informative.

Corollary S1 (The reduced-form effect bounds the NREC)**.** *Under the conditions of Proposition 2, the numerator of* (7) *equals* $(|\mathcal{C}|/N)$ *NREC, where the complier share* $|\mathcal{C}|/N$ *lies in* $(0, 1]$*.*

Corollary S1 follows from a single equation. Writing Δ_Y for the numerator of the ratio in (7), the equation is $\Delta_Y = (|\mathcal{C}|/N)\,\text{NREC}$. In words, the numerator Δ_Y averages the cue's effects on the outcome over all N units, and every non-complier contributes zero to that average, so the average dilutes the compliers' average effect by the factor $|\mathcal{C}|/N$, the complier share.

The proof below derives the equation $\Delta_Y = (|\mathcal{C}|/N)\,\text{NREC}$ and records three of the equation's consequences. First, because the complier share is at most one, the dilution of the compliers' average effect cannot enlarge that effect, so $|\Delta_Y| \leq |\text{NREC}|$. Second, the two magnitudes $|\Delta_Y|$ and $|\text{NREC}|$ are equal exactly when the dilution changes nothing, either because every unit is a complier ($|\mathcal{C}| = N$) or because there is no effect to dilute ($\text{NREC} = 0$). Third, because the complier share is positive, dividing Δ_Y by the complier share undoes the dilution and recovers the NREC, $\text{NREC} = (N/|\mathcal{C}|)\,\Delta_Y$, with the factor $N/|\mathcal{C}|$ confined to the finite set $\{N/k : k = 1, \ldots, N\}$.

Proof of Corollary S1. The proof shows that the denominator of the ratio in (7) equals the complier share and combines that value with Proposition 2 to obtain $\Delta_Y = (|\mathcal{C}|/N)\,\text{NREC}$; the three consequences then follow from $0 < |\mathcal{C}|/N \leq 1$.

Write $\Delta_Y \coloneqq \frac{1}{N}\sum_{i=1}^{N}\left[y_i\left(z_i(1), \boldsymbol{v}_i(1)\right) - y_i\left(z_i(0), \boldsymbol{v}_i(0)\right)\right]$ for the numerator of (7) and $\Delta_Z \coloneqq \frac{1}{N}\sum_{i=1}^{N}\left[z_i(1) - z_i(0)\right]$ for its denominator. As established in the proof of Proposition 2, each unit is a complier, always-taker, or never-taker (Assumption S3), contributing $z_i(1) - z_i(0)$ equal to 1, 0, and 0 respectively, so $\Delta_Z = |\mathcal{C}|/N$. Proposition 2 gives $\Delta_Y/\Delta_Z = \text{NREC}$, hence $\Delta_Y = \Delta_Z\,\text{NREC} = (|\mathcal{C}|/N)\,\text{NREC}$.

By Assumption S4, $|\mathcal{C}| \geq 1$, and $|\mathcal{C}| \leq N$ since $\mathcal{C} \subseteq \{1, \ldots, N\}$; therefore $|\mathcal{C}|/N \in (0, 1]$.

Taking absolute values of both sides of $\Delta_Y = (|\mathcal{C}|/N)\,\text{NREC}$,

$$|\Delta_Y| = \frac{|\mathcal{C}|}{N}\,|\text{NREC}| \leq |\text{NREC}|,$$

because $|\mathcal{C}|/N \leq 1$. The inequality holds with equality exactly when $|\mathcal{C}|/N = 1$ — equivalently, $|\mathcal{C}| = N$ — or when $\text{NREC} = 0$. Finally, because $|\mathcal{C}|/N > 0$, dividing both sides of $\Delta_Y = (|\mathcal{C}|/N)\,\text{NREC}$ by the complier share recovers the NREC from the numerator alone, $\text{NREC} = (N/|\mathcal{C}|)\,\Delta_Y$; as $|\mathcal{C}|$ ranges over the integers $\{1, \ldots, N\}$, the factor $N/|\mathcal{C}|$ ranges over $\{N/k : k = 1, \ldots, N\}$. □

An estimator that uses the numerator alone therefore shares the NREC's sign but understates the NREC's magnitude: The estimator is conservative, guaranteed never to overstate. Kaufman et al. (2026) suggest that the understatement may be substantial, and a researcher could treat the unknown complier share as a sensitivity parameter, reassessing the implied NREC across the share's possible values, analogous to the sensitivity analyses in Leavitt and Rivera-Burgos (2024, 2026).

Finally, Proposition 3 of the main text decomposes cue stability into two conjuncts and assigns a role to each conjunct of the decomposition.

Proof of Proposition 3. The proof proceeds in three parts: The partition of units into compliers and non-compliers splits cue stability into the two conjuncts; conjunct (i) is then shown to be necessary and sufficient for the ratio to equal the NREC; and, given conjunct (i), conjunct (ii) is shown to be necessary and sufficient for the NREC to equal the AEE among compliers. Each sufficiency claim holds for all potential outcomes, and each necessity claim is established by constructing potential outcomes under which the conjunct fails and the corresponding equality fails with it.

By Assumption S3, the units partition into compliers ($z_i(0) = 0$, $z_i(1) = 1$) and non-compliers ($z_i(0) = z_i(1)$). Cue stability requires $\boldsymbol{v}_i(0) = \boldsymbol{v}_i(1)$ for every unit, so cue stability holds if and only if the equality $\boldsymbol{v}_i(0) = \boldsymbol{v}_i(1)$ holds on each cell of the partition: on non-compliers, which is Assumption 2 of the main text, conjunct (i), and on compliers, which

is conjunct (ii).

Sufficiency of conjunct (i) for the ratio in (7) to equal the NREC is Proposition 2. Proposition 2 concludes that the ratio equals the NREC for all potential outcomes satisfying the proposition's assumptions, so, to show that conjunct (i) is necessary for that conclusion, it suffices to construct one set of potential outcomes that satisfies every assumption except conjunct (i) and under which the ratio differs from the NREC. The argument below constructs such a set of potential outcomes.

Suppose that conjunct (i) fails for some non-complier j, so that $\boldsymbol{v}_j(0) \neq \boldsymbol{v}_j(1)$. Consider potential outcomes with $y_j\left(z_j(0), \boldsymbol{v}_j(1)\right) - y_j\left(z_j(0), \boldsymbol{v}_j(0)\right) = \kappa \neq 0$, with every complier's contrast equal to zero, and with the corresponding term $y_i\left(z_i(0), \boldsymbol{v}_i(1)\right) - y_i\left(z_i(0), \boldsymbol{v}_i(0)\right)$ equal to zero for every non-complier $i \neq j$. The denominator of (7) still equals $|\mathcal{C}|/N > 0$, while the numerator equals $\kappa/N \neq 0$ because unit j's term is the only nonzero term, so the ratio equals $\kappa/|\mathcal{C}| \neq 0 = \text{NREC}$.

The third part of the proof takes conjunct (i) as given and shows that conjunct (ii) is necessary and sufficient for the NREC to equal the AEE among compliers. Given conjunct (i), the ratio in (7) equals the NREC by the sufficiency just established, so recovery is settled, and the remaining question is whether the NREC equals the AEE among compliers — the average over $\mathcal{C}$ of the all-else-equal contrast evaluated, for each complier, at the nonracial perception the complier would hold under cue value 0, $\frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\left[y_i\left(1, \boldsymbol{v}_i(0)\right) - y_i\left(0, \boldsymbol{v}_i(0)\right)\right]$. Subtracting the AEE among compliers from the NREC in (6) term by term,

$$
\begin{aligned}
&\frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\left[y_i\left(1, \boldsymbol{v}_i(1)\right) - y_i\left(0, \boldsymbol{v}_i(0)\right)\right] - \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\left[y_i\left(1, \boldsymbol{v}_i(0)\right) - y_i\left(0, \boldsymbol{v}_i(0)\right)\right] \\
&= \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\left\{\left[y_i\left(1, \boldsymbol{v}_i(1)\right) - y_i\left(0, \boldsymbol{v}_i(0)\right)\right] - \left[y_i\left(1, \boldsymbol{v}_i(0)\right) - y_i\left(0, \boldsymbol{v}_i(0)\right)\right]\right\} \\
&= \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\left[y_i\left(1, \boldsymbol{v}_i(1)\right) - y_i\left(1, \boldsymbol{v}_i(0)\right)\right],
\end{aligned}
$$

where the first equality combines the two sums term by term and the second cancels the terms $y_i\left(0, \boldsymbol{v}_i(0)\right)$ within each summand. The difference between the NREC and the AEE

among compliers therefore equals the final sum, and both directions of the third part read off that sum.

To establish sufficiency, suppose conjunct (ii) holds. Every term of the final sum vanishes because $\boldsymbol{v}_i(1) = \boldsymbol{v}_i(0)$ for each complier $i \in \mathcal{C}$, so the NREC equals the AEE among compliers for all potential outcomes.

To establish necessity, suppose conjunct (ii) fails for some complier j, so that $\boldsymbol{v}_j(1) \neq \boldsymbol{v}_j(0)$; a counterexample analogous to the one for conjunct (i) then breaks the equality. Consider potential outcomes with $y_j\left(1, \boldsymbol{v}_j(1)\right) - y_j\left(1, \boldsymbol{v}_j(0)\right) \neq 0$ and with the corresponding terms of all other compliers equal to zero. The final sum is then nonzero, so the NREC differs from the AEE among compliers, and the equality cannot hold for all potential outcomes without conjunct (ii). □

Why conjunct (ii) has no classical counterpart. The identification of the NREC in Proposition 2 has the same structure as the identification of local average treatment effects (Angrist et al., 1996; Gerber and Green, 2012; Imbens and Angrist, 1994): The cue w is the instrument, and the perceived configuration $(z_i(w), \boldsymbol{v}_i(w))$ — racial and nonracial components together — is the treatment. Yet the classical treatment of instrumental variables demands no condition like conjunct *(ii)*, and the reason is the dimension of the treatment. In the classical setting the treatment is a single variable, the analogue of perceived race $z_i(w)$ alone, so compliance fully specifies a complier's treatment at both values of the instrument, $z_i(0) = 0$ and $z_i(1) = 1$, and nothing about what compliers would receive is left to assume. Under the perception-based regime, compliance is defined by the racial component alone, so the same specification, $z_i(0) = 0$ and $z_i(1) = 1$, leaves the nonracial component $\boldsymbol{v}_i(w)$ of what compliers would perceive unspecified. Conjunct *(ii)* supplies the specification, $\boldsymbol{v}_i(0) = \boldsymbol{v}_i(1)$ for each complier $i \in \mathcal{C}$.

A condition that specifies which treatment the compliers would receive selects the estimand rather than recovers it. Under conjunct *(ii)*, each complier's induced contrast $y_i\left(1, \boldsymbol{v}_i(1)\right) - y_i\left(0, \boldsymbol{v}_i(0)\right)$ becomes $y_i\left(1, \boldsymbol{v}_i(0)\right) - y_i\left(0, \boldsymbol{v}_i(0)\right)$, a contrast that differs in per-

ceived race alone, so the NREC collapses into the AEE among compliers. Recovery needs no such condition, because, given conjunct *(i)* — $\boldsymbol{v}_i(0) = \boldsymbol{v}_i(1)$ for every non-complier — the ratio in (7) equals the NREC with or without conjunct *(ii)* (Proposition 3). The distinction between selecting and recovering the estimand is the one that Section 6.2 of the main text draws from the decomposition of cue stability, and the classical setting could leave the distinction implicit because a one-dimensional treatment leaves conjunct *(ii)* nothing to specify.

S1.5 The mixed contrast

Section 4.1 of the main text introduces the mixed contrast as one that holds some nonracial features fixed across racial conditions and lets the rest vary by race. This subsection defines the partition of the nonracial features into a held-fixed set and a free set. From the joint PMF $q(z, \boldsymbol{v})$ and the common weighting ρ of Section 4 of the main text, the subsection then derives the two distributions a mixed contrast needs, namely the conditional distribution of the free features given the held-fixed features within each racial condition and the marginal distribution of the held-fixed features.

Partition the indices of the nonracial features, $\{1, \ldots, L\} = \mathcal{I}_{\text{fix}} \cup \mathcal{I}_{\text{free}}$ with $\mathcal{I}_{\text{fix}} \cap \mathcal{I}_{\text{free}} = \emptyset$. Write $\boldsymbol{v} = (\boldsymbol{v}_{\text{fix}}, \boldsymbol{v}_{\text{free}})$, with $\boldsymbol{v}_{\text{fix}} \in \mathcal{V}_{\text{fix}} \coloneqq \prod_{\ell \in \mathcal{I}_{\text{fix}}} \mathcal{V}_\ell$ collecting the features held fixed and $\boldsymbol{v}_{\text{free}} \in \mathcal{V}_{\text{free}} \coloneqq \prod_{\ell \in \mathcal{I}_{\text{free}}} \mathcal{V}_\ell$ collecting the free features. The AEE is the case $\mathcal{I}_{\text{free}} = \emptyset$ (every feature held fixed) and the AEWR the case $\mathcal{I}_{\text{fix}} = \emptyset$ (no feature held fixed).

When $\mathcal{I}_{\text{fix}}$ and $\mathcal{I}_{\text{free}}$ are both nonempty, the weighting follows the partition, as Section 4.1 of the main text describes. A distribution common to the two racial conditions weights the held-fixed features, as the common weighting ρ weights all of $\boldsymbol{v}$ for the AEE, and race-specific distributions weight the free features given the held-fixed value, as the q_z weight all of $\boldsymbol{v}$ for the AEWR. The construction below gives the two distributions and then assembles the mixed effect in a single expression.

Within racial condition z at held-fixed value $\boldsymbol{v}_{\text{fix}}$, the free features follow the conditional

distribution

$$q_z(\boldsymbol{v}_{\text{free}} \mid \boldsymbol{v}_{\text{fix}}) \coloneqq \frac{q(z, \boldsymbol{v}_{\text{fix}}, \boldsymbol{v}_{\text{free}})}{\sum_{\boldsymbol{v}'_{\text{free}} \in \mathcal{V}_{\text{free}}} q(z, \boldsymbol{v}_{\text{fix}}, \boldsymbol{v}'_{\text{free}})}, \tag{S2}$$

which the joint PMF $q(z, \boldsymbol{v})$ of Section 4 of the main text induces, and which is defined at each pair of a racial condition and a held-fixed value for which the denominator of (S2) is positive. The held-fixed features play the role that the whole profile plays in the AEE, so their weights must be common to the two racial conditions. The aggregation over the held-fixed values uses the marginal distribution of the held-fixed features under the common weighting ρ of Section 4 of the main text,

$$\rho(\boldsymbol{v}_{\text{fix}}) \coloneqq \sum_{\boldsymbol{v}_{\text{free}} \in \mathcal{V}_{\text{free}}} \rho(\boldsymbol{v}_{\text{fix}}, \boldsymbol{v}_{\text{free}}). \tag{S3}$$

For unit i, the race effect under the mixed contrast combines the two distributions in one expression,

$$\begin{aligned} \sum_{\boldsymbol{v}_{\text{fix}} \in \mathcal{V}_{\text{fix}}} \Bigg[& \sum_{\boldsymbol{v}_{\text{free}} \in \mathcal{V}_{\text{free}}} y_i(1, \boldsymbol{v}_{\text{fix}}, \boldsymbol{v}_{\text{free}})\, q_1(\boldsymbol{v}_{\text{free}} \mid \boldsymbol{v}_{\text{fix}}) \\ & - \sum_{\boldsymbol{v}_{\text{free}} \in \mathcal{V}_{\text{free}}} y_i(0, \boldsymbol{v}_{\text{fix}}, \boldsymbol{v}_{\text{free}})\, q_0(\boldsymbol{v}_{\text{free}} \mid \boldsymbol{v}_{\text{fix}}) \Bigg] \rho(\boldsymbol{v}_{\text{fix}}). \end{aligned} \tag{S4}$$

Read the expression from the inside out. Each inner sum averages unit i's potential outcomes over the free features within one racial condition at the held-fixed value $\boldsymbol{v}_{\text{fix}}$ with the weights in (S2). The difference between the two inner sums contrasts the racial conditions at that common held-fixed value, so the contrast is all-else-equal on the held-fixed features and within-race on the free features. The outer sum aggregates the contrasts over the held-fixed values with the common weights in (S3).

Averaging the race effect (S4) over the N units gives the mixed effect, and collecting the weight that multiplies each potential outcome $y_i(z, \boldsymbol{v}_{\text{fix}}, \boldsymbol{v}_{\text{free}})$ shows that the mixed effect

is the member of the estimand family with the mixed weights

$$g_z(\boldsymbol{v}) = \rho(\boldsymbol{v}_{\text{fix}})\, q_z(\boldsymbol{v}_{\text{free}} \mid \boldsymbol{v}_{\text{fix}}), \tag{S5}$$

which are common on the held-fixed features and race-specific on the free features given the held-fixed value. Section S1.2 substitutes the mixed weights in (S5) into the gap identity (S1) to obtain the gap between a mixed effect and any other member of the estimand family — the AEE, the AEWR, or a mixed effect built on a different partition of the nonracial features.

S1.6 The structural tradeoff

Section 5 of the main text argues that, when racial classifications index different regions of the attribute space, no common all-else-equal weighting can both restrict attention to profiles typical of both groups and represent the typical members of each group. This subsection makes that claim precise and proves it.

Fix $\varepsilon \in (0,1)$. For each race $z \in \{0,1\}$, call a set $\mathcal{R}_z \subseteq \mathcal{V}$ a *typical region for race z* if

$$q_z(\mathcal{R}_z) \coloneqq \sum_{\boldsymbol{v} \in \mathcal{R}_z} q_z(\boldsymbol{v}) \geq 1 - \varepsilon,$$

so that the profiles of at least a $1-\varepsilon$ share of individuals with race z lie in $\mathcal{R}_z$. On the thick constructivist conception of Section 2 of the main text, a racial category indexes a distribution of nonracial attributes; the typical region for race z is the region of the attribute space on which the indexed distribution q_z concentrates. The intersection $\mathcal{R}_0 \cap \mathcal{R}_1$ is the *both-typical region*.

Proposition S2 (Structural tradeoff)**.** *Fix $\varepsilon \in (0,1)$ and typical regions $\mathcal{R}_0, \mathcal{R}_1 \subseteq \mathcal{V}$. Suppose that for at least one race $z \in \{0,1\}$,*

$$q_z(\mathcal{R}_0 \cap \mathcal{R}_1) < 1 - \varepsilon. \tag{S6}$$

Then no weighting distribution ρ on $\mathcal{V}$, common across the two racial conditions, satisfies

both of the following constraints:

(i) $\text{supp}(\rho) \subseteq \mathcal{R}_0 \cap \mathcal{R}_1$*, so that* ρ *places all of its weight on both-typical profiles; and*

(ii) $q_z(\text{supp}(\rho)) \geq 1 - \varepsilon$ *for each race* $z \in \{0, 1\}$*, so that the support of* ρ *retains at least a* $1 - \varepsilon$ *share of each group.*

Proof. The proof supposes constraint (i) and derives the negation of constraint (ii) from the proposition's hypothesis, the inequality (S6).

Let $z^\star \in \{0, 1\}$ be a race for which (S6) holds, and suppose a common weighting distribution ρ satisfies constraint (i), so that $\text{supp}(\rho) \subseteq \mathcal{R}_0 \cap \mathcal{R}_1$. For any set of profiles $\mathcal{A} \subseteq \mathcal{V}$, the quantity $q_{z^\star}(\mathcal{A}) = \sum_{\boldsymbol{v} \in \mathcal{A}} q_{z^\star}(\boldsymbol{v})$ is the share of individuals with race $z^\star$ whose profiles lie in $\mathcal{A}$. Each term of the sum is nonnegative, so enlarging $\mathcal{A}$ to any set $\mathcal{B}$ with $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ adds terms and removes none; hence $q_{z^\star}(\mathcal{A}) \leq q_{z^\star}(\mathcal{B})$ (monotonicity with respect to set inclusion).

Take $\mathcal{A} = \text{supp}(\rho)$ and $\mathcal{B} = \mathcal{R}_0 \cap \mathcal{R}_1$; the inclusion $\mathcal{A} \subseteq \mathcal{B}$ is then exactly constraint (i). Monotonicity with respect to set inclusion at these two sets, $q_{z^\star}(\text{supp}(\rho)) \leq q_{z^\star}(\mathcal{R}_0 \cap \mathcal{R}_1)$, gives the weak inequality, and the hypothesis (S6), $q_{z^\star}(\mathcal{R}_0 \cap \mathcal{R}_1) < 1 - \varepsilon$, gives the strict inequality, in the chain of inequalities

$$q_{z^\star}(\text{supp}(\rho)) \;\leq\; q_{z^\star}(\mathcal{R}_0 \cap \mathcal{R}_1) \;<\; 1 - \varepsilon. \tag{S7}$$

In the chain (S7), a weak inequality is followed by a strict inequality, and the two inequalities compose by transitivity into the strict inequality between the ends, $q_{z^\star}(\text{supp}(\rho)) < 1 - \varepsilon$. The strict inequality between the ends states, in words, that the support of ρ retains less than a $1 - \varepsilon$ share of the individuals with race $z^\star$, and the statement is the negation of constraint (ii) at $z = z^\star$. Hence no common weighting distribution ρ satisfies constraints (i) and (ii) together. □

Proposition S2 admits two logically equivalent statements — the proposition as stated and the proposition's contrapositive — and each statement shows what imposing one of the two constraints forces the researcher to give up. As stated, the proposition says that

if the common weighting ρ places all of its weight on both-typical profiles, as constraint (i) requires, then the support of ρ retains less than a $1-\varepsilon$ share of at least one group, so constraint (ii) fails. In the contrapositive, the proposition says that if the common weighting ρ retains at least a $1-\varepsilon$ share of each group, as constraint (ii) requires, then ρ must place positive weight on profiles outside the typical region for at least one race, so constraint (i) fails.

Worked example. A stylized version of the running example shows, in numbers, what each constraint of Proposition S2 forces the researcher to give up. The example follows the shape that Figure 2 of the main text depicts: two distributions of neighborhoods and prior records that overlap only slightly (Section 5 of the main text). Code the two nonracial features of the running example as binary — $v_1 = 1$ for a disadvantaged neighborhood and $v_2 = 1$ for an extensive prior record — so that the profiles $\boldsymbol{v} = (v_1, v_2)$ take four possible values. Suppose that, among arrestees, Black, relative to White, indexes a greater likelihood of arrest in a disadvantaged neighborhood and a greater likelihood of an extensive prior record. The distribution q_1 of profiles among Black arrestees then places its largest share on the profile with both features, smaller shares on the two profiles with one feature, and its smallest share on the profile with neither:

$$q_1(1,1) = .60, \quad q_1(1,0) = .20, \quad q_1(0,1) = .15, \quad q_1(0,0) = .05.$$

The distribution q_0 of profiles among White arrestees concentrates on the profile with neither feature and places the same shares as q_1 on the single-feature profiles: $q_0(0,0) = .60$, $q_0(1,0) = .20$, $q_0(0,1) = .15$, and $q_0(1,1) = .05$. For both groups, the profile with an extensive prior record but no disadvantaged neighborhood is the rarer of the two single-feature profiles, $q_z(0,1) = .15 < .20 = q_z(1,0)$, because an extensive prior record arises partly from the greater policing of disadvantaged neighborhoods.

Fix $\varepsilon = .10$, so that a typical region must hold at least a .90 share of its group. The set $\mathcal{R}_1 = \{(1,1),(1,0),(0,1)\}$ is a typical region for race 1, with $q_1(\mathcal{R}_1) = .95 \geq .90$, and

the set $\mathcal{R}_0 = \{(0,0),(0,1),(1,0)\}$ is a typical region for race 0, with $q_0(\mathcal{R}_0) = .95$. The both-typical region is the pair of single-feature profiles, $\mathcal{R}_0 \cap \mathcal{R}_1 = \{(1,0),(0,1)\}$. The both-typical region holds only a .35 share of each group, $q_z(\mathcal{R}_0 \cap \mathcal{R}_1) = .35 < .90$, so the hypothesis (S6) holds for both races.

A common weighting ρ that satisfies constraint (i) places all of its weight on the two single-feature profiles, so the support of ρ retains at most a .35 share of each group — far below the .90 share that constraint (ii) demands. A common weighting that satisfies constraint (ii) must retain at least a .90 share of group 1, and the support of any such weighting includes the profile $(1,1)$, because the three profiles other than $(1,1)$ together hold only a .40 share of group 1. The profile $(1,1)$ carries a .05 share of group 0, so the weighting places weight on a profile rare under group 0's distribution. The same argument with the groups exchanged forces weight on $(0,0)$, a profile rare under group 1's distribution.

Whenever the hypothesis (S6) holds, so that the both-typical region contains less than a $1-\varepsilon$ share of at least one group, constraints (i) and (ii) of Proposition S2 cannot hold together, and the researcher must give up either the restriction to both-typical profiles or the representation of each group's typical members.

S1.7 Distribution specifications

Section 5.4 of the main text motivates specifications of q_z and ρ; this subsection develops the specifications in full. The three parts below take up the restriction of the support to exclude implausible profiles, the information that calibrating the weighting distributions to a reference population requires, and the translation of graded typicality into a distribution that concentrates around each group's prototype.

S1.7.1 Restricting the support

Restrictions on implausible attribute combinations operate directly on the joint PMF $q(z, \boldsymbol{v}_{\text{fix}}, \boldsymbol{v}_{\text{free}})$ of Section 4 of the main text, defined on the space $\mathcal{T}$ of configurations. Hainmueller et al. (2014, pp. 13–16) build randomization restrictions for implausible com-

binations into the canonical conjoint framework and note that causal effects are then defined only over the combinations the design retains (p. 20).

The researcher judges some configurations in $\mathcal{T}$ implausible and retains the rest in the comparison. Let $\mathcal{T}^* \subseteq \mathcal{T}$ denote the set of retained configurations. Excluding the implausible configurations from the comparison is a choice about the estimand's weighting, not a claim that the excluded configurations cannot occur: The new joint PMF q^* places zero weight on every configuration outside $\mathcal{T}^*$ and rescales the weights that q places on the configurations in $\mathcal{T}^*$ so that the rescaled weights sum to one. Formally, define

$$q^*(z, \boldsymbol{v}_{\text{fix}}, \boldsymbol{v}_{\text{free}}) \coloneqq \frac{q(z, \boldsymbol{v}_{\text{fix}}, \boldsymbol{v}_{\text{free}}) \mathbb{1}\{(z, \boldsymbol{v}_{\text{fix}}, \boldsymbol{v}_{\text{free}}) \in \mathcal{T}^*\}}{\sum_{(z', \boldsymbol{v}'_{\text{fix}}, \boldsymbol{v}'_{\text{free}}) \in \mathcal{T}^*} q(z', \boldsymbol{v}'_{\text{fix}}, \boldsymbol{v}'_{\text{free}})}, \tag{S8}$$

whenever the denominator is positive.

For a configuration retained in $\mathcal{T}^*$, the indicator in the numerator of (S8) equals one, so q^* returns the weight $q(z, \boldsymbol{v}_{\text{fix}}, \boldsymbol{v}_{\text{free}})$ divided by the total weight that q places on $\mathcal{T}^*$. For a configuration outside $\mathcal{T}^*$, the indicator equals zero, so q^* returns zero. The values of q^* are therefore nonnegative and sum to one over $\mathcal{T}^*$, so q^* is a PMF on $\mathcal{T}^*$, proportional to q on the configurations retained in $\mathcal{T}^*$. All subsequent quantities — including the conditional distribution $q_z(\boldsymbol{v}_{\text{free}} \mid \boldsymbol{v}_{\text{fix}})$ in (S2) and the marginal distribution $\rho(\boldsymbol{v}_{\text{fix}})$ in (S3) — are then computed from q^* rather than from q.

Replacing q with q^* leaves the formal structure of the mixed effect in (S4) unchanged. The replacement can nonetheless leave parts of the mixed effect with nothing to average or to compare, because each inner average in (S4) draws its terms from the configurations retained in $\mathcal{T}^*$ alone. To state when each part of the mixed effect exists, collect, for racial condition z at held-fixed value $\boldsymbol{v}_{\text{fix}}$, the free-feature values that the restriction to $\mathcal{T}^*$ leaves available,

$$\mathcal{V}^*_{\text{free}}(z, \boldsymbol{v}_{\text{fix}}) \coloneqq \{\boldsymbol{v}_{\text{free}} \in \mathcal{V}_{\text{free}} : (z, \boldsymbol{v}_{\text{fix}}, \boldsymbol{v}_{\text{free}}) \in \mathcal{T}^*\}. \tag{S9}$$

Two requirements for the existence of the mixed effect follow. First, under q^*, the inner

average over the free features within racial condition z at held-fixed value $\boldsymbol{v}_{\text{fix}}$ becomes

$$\sum_{\boldsymbol{v}_{\text{free}} \in \mathcal{V}^*_{\text{free}}(z, \boldsymbol{v}_{\text{fix}})} y_i(z, \boldsymbol{v}_{\text{fix}}, \boldsymbol{v}_{\text{free}})\, q^*_z(\boldsymbol{v}_{\text{free}} \mid \boldsymbol{v}_{\text{fix}}),$$

because $q^*_z(\boldsymbol{v}_{\text{free}} \mid \boldsymbol{v}_{\text{fix}})$ is positive only at the free-feature values in $\mathcal{V}^*_{\text{free}}(z, \boldsymbol{v}_{\text{fix}})$. The inner average therefore exists only when $\mathcal{V}^*_{\text{free}}(z, \boldsymbol{v}_{\text{fix}}) \neq \emptyset$. Second, the contrast at held-fixed value $\boldsymbol{v}_{\text{fix}}$ subtracts the inner average for racial condition 0 from the inner average for racial condition 1. The contrast therefore exists only when both sets of free-feature values are nonempty, $\mathcal{V}^*_{\text{free}}(0, \boldsymbol{v}_{\text{fix}}) \neq \emptyset$ and $\mathcal{V}^*_{\text{free}}(1, \boldsymbol{v}_{\text{fix}}) \neq \emptyset$.

If $\mathcal{V}^*_{\text{free}}(z, \boldsymbol{v}_{\text{fix}})$ is empty for exactly one of the two racial conditions — as when a combination of features is judged plausible for one group alone — the contrast at that held-fixed value is undefined. The outer summation in (S4) weights each held-fixed value by the restricted marginal distribution $\rho^*(\boldsymbol{v}_{\text{fix}})$, which places positive weight only on the held-fixed values that appear in at least one configuration retained in $\mathcal{T}^*$. The mixed effect under q^* therefore exists when the second requirement holds at every held-fixed value on which ρ^* places positive weight.

The existence condition is membership in the estimand family, whose weights are probability distributions (Section 4 of the main text). For each racial condition z, the mixed weights under the restriction to $\mathcal{T}^*$ sum to

$$\sum_{\boldsymbol{v}_{\text{fix}} \in \mathcal{V}_{\text{fix}}} \rho^*(\boldsymbol{v}_{\text{fix}}) \sum_{\boldsymbol{v}_{\text{free}} \in \mathcal{V}_{\text{free}}} q^*_z(\boldsymbol{v}_{\text{free}} \mid \boldsymbol{v}_{\text{fix}}).$$

The inner sum equals one at each held-fixed value where the conditional distribution $q^*_z(\boldsymbol{v}_{\text{free}} \,|\, \boldsymbol{v}_{\text{fix}})$ exists, so the mixed weights are well defined and sum to one exactly when the conditional distribution exists at every held-fixed value on which ρ^* places positive weight.

S1.7.2 Calibration to a reference population

The construction of the mixed contrast in Section S1.5 partitions the nonracial features into a held-fixed set and a free set. Under the partition, calibrating the joint PMF $q(z, \boldsymbol{v}_{\text{fix}}, \boldsymbol{v}_{\text{free}})$

to match a reference population requires more information than the analogous calibration in de la Cuesta et al. (2022). The reason lies in which pieces of the joint PMF the mixed effect uses. Written with the mixed weights of (S5), the race effect under the mixed contrast for unit i in (S4) is

$$\sum_{\boldsymbol{v} \in \mathcal{V}} \left[y_i(1, \boldsymbol{v})\, \rho(\boldsymbol{v}_{\text{fix}})\, q_1(\boldsymbol{v}_{\text{free}} \mid \boldsymbol{v}_{\text{fix}}) - y_i(0, \boldsymbol{v})\, \rho(\boldsymbol{v}_{\text{fix}})\, q_0(\boldsymbol{v}_{\text{free}} \mid \boldsymbol{v}_{\text{fix}}) \right],$$

so the mixed effect uses the joint PMF only through the marginal distribution $\rho(\boldsymbol{v}_{\text{fix}})$ of the held-fixed features and the conditional distributions $q_z(\boldsymbol{v}_{\text{free}} \mid \boldsymbol{v}_{\text{fix}})$ of the free features within each racial condition. Calibrating the mixed effect means supplying the marginal distribution and the conditional distributions from data on the reference population.

The lighter requirement in de la Cuesta et al. (2022, p. 21) — each attribute's marginal distribution alone — rests on the assumption that the attributes have no three-way or higher-order interactions, and supplying partial joint distributions relaxes the assumption (de la Cuesta et al., 2022, pp. 21, 31). The assumption restricts the outcomes, and under that restriction the estimand of de la Cuesta et al. (2022) needs no more from the profile distribution than the marginal distributions provide. No restriction on the outcomes does the same for the mixed effect, because the weights of the mixed effect are themselves built from the joint PMF. A researcher could instead restrict the reference population — supposing, for example, that the attributes are independent within each racial condition — and calibrate from marginal distributions alone. The supposition would then fix the weights, and with the weights the estimand, by assumption rather than by data.

A reference population is often summarized by *single-attribute margins* — for each attribute separately, the share of the population at each value of that attribute, with no record of how the attributes combine. Many joint distributions share the same single-attribute margins while disagreeing about the dependence across attributes. Both distributions that the calibration needs are properties of the dependence: The marginal distribution $\rho(\boldsymbol{v}_{\text{fix}})$ assigns probability to whole bundles of held-fixed values, and the conditional distribution

$q_z(\boldsymbol{v}_{\text{free}} \mid \boldsymbol{v}_{\text{fix}})$ records how the free features vary with the racial condition and the held-fixed value.

Worked example. The example exhibits two populations that share every single-attribute margin yet differ in the conditional distribution the calibration needs. Take the two nonracial features of the running example of the main text — the location of the arrest and the extent of the prior record — coded as binary: $v_1 = 1$ for an arrest in a disadvantaged neighborhood and $v_1 = 0$ otherwise, and $v_2 = 1$ for an extensive prior record and $v_2 = 0$ otherwise. Consider a stylized reference population of 100 arrestees that splits evenly on each feature, $\Pr(v_1 = 1) = \Pr(v_2 = 1) = 1/2$, and two ways the features can combine within the stylized population (Table S1). In the left population of the table, the two features are independent, and each of the four combinations of (v_1, v_2) holds 25 arrestees. In the right population, the two features coincide, with the matching combinations $(1, 1)$ and $(0, 0)$ holding 50 arrestees each and the combinations $(1, 0)$ and $(0, 1)$ holding none. The conditional distribution of prior record given neighborhood is even in the left population, $\Pr(v_2 = 1 \mid v_1) = 1/2$, and degenerate in the right population, $\Pr(v_2 = v_1 \mid v_1) = 1$. The shading in Table S1 marks what the two populations share: The shaded row and column totals — the single-attribute margins — are identical, while the unshaded interior cells differ.

Independent features

	v_2		
v_1	1	0	Total
1	25	25	50
0	25	25	50
Total	50	50	100

Coinciding features

	v_2		
v_1	1	0	Total
1	50	0	50
0	0	50	50
Total	50	50	100

Table S1: Two stylized populations of 100 arrestees, one per table. Cells count arrestees by neighborhood v_1 (rows) and prior record v_2 (columns); because each population holds 100 arrestees, dividing any cell by 100 gives the corresponding population proportion. The shaded cells — the row and column totals, which are the single-attribute margins — are identical across the two populations; the unshaded interior cells differ.

The conditional distributions $q_z(\boldsymbol{v}_{\text{free}} \mid \boldsymbol{v}_{\text{fix}})$ are underdetermined in the same way as

the conditional distribution of the worked example, and more severely. The conditioning includes the racial condition z, so the calibration also needs the dependence between the nonracial features and race. Identifying the marginal distribution $\rho(\boldsymbol{v}_{\text{fix}})$ and the conditional distributions $q_z(\boldsymbol{v}_{\text{free}} \,|\, \boldsymbol{v}_{\text{fix}})$ therefore requires joint information: cross-tabulations of the free features against the held-fixed features and race in the reference population, or individual-level data on the members of the reference population, from which any such cross-tabulation can be computed.

S1.7.3 Typicality

The construction sketched in Section 5.4 of the main text translates graded typicality into the conditional distribution $q_z(\boldsymbol{v}_{\text{free}})$. Three elements of the construction come from the theory of conceptual spaces (Gärdenfors, 2000, 2014): the representation of objects as points in a space of *quality dimensions* — scales on which the objects can be compared (Gärdenfors, 2000, p. 2) — with similarity decreasing in distance (Gärdenfors, 2000, p. 44); the representation of a category by a prototype, the category's most typical point (Gärdenfors, 2000, pp. 87–91); and a distance that weights each dimension by the dimension's salience (Gärdenfors, 2000, p. 104). The exponential decay of similarity in distance follows Nosofsky (1986) and Shepard (1987), whom Gärdenfors (2000, p. 20) also follows. The assembly is this paper's: group-specific prototypes, salience weights, and concentration parameters, applied to the finite set $\mathcal{V}_{\text{free}}$ and normalized into a weighting distribution for the estimand family, together with the conditional-prototype variant below.

The construction needs three ingredients.

- An embedding $\phi : \mathcal{V}_{\text{free}} \to \mathbb{R}^D$ — a function that represents the elements of one set as points of a space — maps each free-feature profile to a point of a D-dimensional conceptual space, with each coordinate one quality dimension. The number of dimensions D may be smaller than the number of free features, as when several features collapse into a single scale, or larger, as when one feature occupies several scales — neighborhood disadvantage in the running example, for instance, may occupy one coordinate

for income and one for intensity of policing.

- A racial prototype $\boldsymbol{p}_z \in \mathbb{R}^D$ locates the most typical bundle of free features for group z in the conceptual space.
- Salience weights $\boldsymbol{a}_z \in [0,1]^D$ with $\sum_{k=1}^{D} a_{z,k} = 1$ record how diagnostic each dimension of the conceptual space is for membership in group z. A difference along a highly salient dimension reduces typicality more than the same difference along a less salient dimension.

Typicality is similarity to the prototype, and similarity in a conceptual space decreases with distance, so grading typicality requires a distance from the prototype in which differences along more salient dimensions matter more. The embedding, the prototype, and the salience weights combine into such a distance, the weighted Euclidean distance from the prototype,

$$d_z(\boldsymbol{v}_{\text{free}}, \boldsymbol{p}_z) \coloneqq \sqrt{(\phi(\boldsymbol{v}_{\text{free}}) - \boldsymbol{p}_z)^\top \operatorname{diag}(\boldsymbol{a}_z)\,(\phi(\boldsymbol{v}_{\text{free}}) - \boldsymbol{p}_z)}, \tag{S10}$$

where $\operatorname{diag}(\boldsymbol{a}_z)$ denotes the $D \times D$ diagonal matrix with the salience weights $a_{z,1}, \ldots, a_{z,D}$ on its diagonal. The distance is zero exactly when $\phi(\boldsymbol{v}_{\text{free}}) = \boldsymbol{p}_z$ and positive otherwise.

Exponentiate the negative squared distance, scaled by a concentration parameter $\lambda_z \geq 0$ that sets how fast a profile's weight declines with distance from the prototype, and normalize over $\mathcal{V}_{\text{free}}$. The result is a conditional distribution that concentrates around the prototype (Nosofsky, 1986; Shepard, 1987):

$$q_z(\boldsymbol{v}_{\text{free}}; \lambda_z) \coloneqq \frac{\exp\left(-\lambda_z\, d_z(\boldsymbol{v}_{\text{free}}, \boldsymbol{p}_z)^2\right)}{\sum_{\boldsymbol{v}'_{\text{free}} \in \mathcal{V}_{\text{free}}} \exp\left(-\lambda_z\, d_z(\boldsymbol{v}'_{\text{free}}, \boldsymbol{p}_z)^2\right)}. \tag{S11}$$

In words, a profile's weight under (S11) declines exponentially in the profile's weighted squared distance from the prototype, so profiles near the prototype carry most of the weight and profiles far from the prototype contribute little weight.

The two extremes of the concentration parameter show its role: At $\lambda_z = 0$, every profile

in $\mathcal{V}_{\text{free}}$ receives equal weight; as $\lambda_z \to \infty$, the distribution collapses to a point mass on the profile closest to the prototype $\boldsymbol{p}_z$. The two groups may differ in their concentration parameters λ_0 and λ_1 — the formal counterpart of the example in Section 5.4 of the main text, where DeShawn may register as more atypically White than Greg does atypically Black because typicality declines at different rates for the two groups.

Worked example. The application of Section 8.1 of the main text supplies a deliberately simple instance — two profiles and one dimension — in which the typicality construction reproduces the application's weights exactly and shows what each parameter contributes. Among the candidates who campaigned in Spanish in the content analysis of Zárate et al. (2024), the within-race accent distributions over $\mathcal{V}_{\text{free}} = \{\text{Non-Native Spanish}, \text{Native Spanish}\}$ are $q_0 = (.96, .04)$ for Anglo and $q_1 = (.17, .83)$ for Hispanic candidates (Table 1 of the main text). Take a single quality dimension recording the nativeness of the accent, $D = 1$, with the embedding $\phi(\text{Non-Native Spanish}) = 0$ and $\phi(\text{Native Spanish}) = 1$; with one dimension, the salience weights are trivial, $a_{z,1} = 1$ for both groups, and salience plays no role until a second dimension enters. Set each group's prototype at the accent typical of the group's own campaigners, $\boldsymbol{p}_0 = 0$ (non-native) and $\boldsymbol{p}_1 = 1$ (native), so the distance (S10) between a profile and a prototype equals 0 or 1.

With the distances so arranged, the distribution (S11) places weight $1/(1 + e^{-\lambda_z})$ on the accent at group z's prototype and weight $e^{-\lambda_z}/(1 + e^{-\lambda_z})$ on the other accent, so the concentration parameter equals the logarithm of the group's odds of its typical accent. The exact odds come from the counts of the content analysis: 47 non-native and 2 native Spanish speakers among the Anglo campaigners, and 13 non-native and 64 native among the Hispanic campaigners. Setting $\lambda_1 = \ln(64/13) \approx 1.59$ recovers q_1 exactly, and setting $\lambda_0 = \ln(47/2) \approx 3.16$ recovers q_0 exactly. The larger λ_0 says that Anglo campaigners concentrate more tightly on their typical accent than Hispanic campaigners do on theirs. The equality between the typicality weighting and the campaign distributions is the case that Section 8.1 of the main text describes, in which typicality is defined by the target population itself.

Conditional-prototype variant. The most typical bundle of free features for a group may differ across values of the features that the mixed contrast holds fixed. For each pair of a group z and a value $\boldsymbol{v}_{\text{fix}}$ of the held-fixed features, a conditional prototype $\boldsymbol{p}_{z,\boldsymbol{v}_{\text{fix}}} \in \mathbb{R}^D$ locates the most typical bundle of free features for group z among individuals whose held-fixed features equal $\boldsymbol{v}_{\text{fix}}$. Replacing the prototype $\boldsymbol{p}_z$ with the conditional prototype $\boldsymbol{p}_{z,\boldsymbol{v}_{\text{fix}}}$ in the distance (S10) and the distribution (S11) yields

$$q_z(\boldsymbol{v}_{\text{free}} \mid \boldsymbol{v}_{\text{fix}}; \lambda_z) \coloneqq \frac{\exp\left(-\lambda_z \, d_z(\boldsymbol{v}_{\text{free}}, \boldsymbol{p}_{z,\boldsymbol{v}_{\text{fix}}})^2\right)}{\sum_{\boldsymbol{v}'_{\text{free}} \in \mathcal{V}_{\text{free}}} \exp\left(-\lambda_z \, d_z(\boldsymbol{v}'_{\text{free}}, \boldsymbol{p}_{z,\boldsymbol{v}_{\text{fix}}})^2\right)}. \tag{S12}$$

For example, if $\boldsymbol{v}_{\text{fix}}$ contains the neighborhood of the running example and $\boldsymbol{v}_{\text{free}}$ contains the prior record, the most typical prior record for group z among arrestees from a disadvantaged neighborhood may differ from the most typical prior record for group z overall, so the conditional prototype shifts with the neighborhood.

S1.8 Randomization-based properties of the plug-in estimator

This subsection establishes the randomization-based properties of the plug-in estimator of the estimand family, $\widehat{\theta}$ of (9) of the main text, restated with its ingredients in (S16) below. The estimator targets the finite-population estimand

$$\theta \coloneqq \sum_{\boldsymbol{v} \in \mathcal{V}} g_1(\boldsymbol{v}) \, \mu(1, \boldsymbol{v}) - \sum_{\boldsymbol{v} \in \mathcal{V}} g_0(\boldsymbol{v}) \, \mu(0, \boldsymbol{v}), \tag{S13}$$

where $\mu(z, \boldsymbol{v}) \coloneqq \frac{1}{N} \sum_{i=1}^{N} y_i(z, \boldsymbol{v})$ denotes the mean of the potential outcomes at profile $(z, \boldsymbol{v})$ over all N units.

The mean $\mu(z, \boldsymbol{v})$ equals the average potential outcome $\bar{y}(z, \boldsymbol{v})$ of Proposition S1; this subsection writes $\mu(z, \boldsymbol{v})$ so that each population mean pairs with its estimator, the cell mean $\hat{\mu}(z, \boldsymbol{v})$ of (8) of the main text. Exchanging the finite sums over i and $\boldsymbol{v}$ shows that θ coincides with $\theta(g_0, g_1)$ of Proposition S1.

The AEE ($g_0 = g_1 = \rho$), the AEWR ($g_z = q_z$), and the mixed contrasts (g_z as in (S5))

are the special cases obtained by fixing the weights g_0 and g_1, so the results below cover the whole estimand family.

Let $n_{z,\boldsymbol{v}}$ be fixed positive integers for all $(z, \boldsymbol{v}) \in \mathcal{T}$ satisfying $\sum_{\boldsymbol{v}\in\mathcal{V}} n_{0,\boldsymbol{v}} + \sum_{\boldsymbol{v}\in\mathcal{V}} n_{1,\boldsymbol{v}} = N$. For each $(z, \boldsymbol{v}) \in \mathcal{T}$, refer to the set of units assigned to profile $(z, \boldsymbol{v})$ as the *cell* corresponding to profile $(z, \boldsymbol{v})$; the integer $n_{z,\boldsymbol{v}}$ fixes the size of that cell.

Define the set of configuration assignments consistent with these fixed cell sizes as

$$\Omega \coloneqq \bigcap_{(z,\boldsymbol{v})\in\mathcal{T}} \left\{ \boldsymbol{t} \in \mathcal{T}^N : \sum_{i=1}^{N} \mathbb{1}\{\boldsymbol{t}_i = (z, \boldsymbol{v})\} = n_{z,\boldsymbol{v}} \right\}. \tag{S14}$$

Thus, Ω consists of all configuration assignments $\boldsymbol{t} \in \mathcal{T}^N$ such that, for every profile $(z, \boldsymbol{v}) \in \mathcal{T}$, exactly $n_{z,\boldsymbol{v}}$ components of $\boldsymbol{t}$ equal $(z, \boldsymbol{v})$.

Complete random assignment (Assumption 3) makes the random configuration assignment $\boldsymbol{T}$ uniformly distributed on Ω, so $\Pr(\boldsymbol{T} = \boldsymbol{t}) = 1/|\Omega|$ for all $\boldsymbol{t} \in \Omega$, where $|\Omega|$, the number of assignments in Ω, equals $N!/\prod_{(z,\boldsymbol{v})\in\mathcal{T}} n_{z,\boldsymbol{v}}!$.

To construct the plug-in estimator $\widehat{\theta}$, define the observed outcome for unit i as $Y_i \coloneqq y_i(\boldsymbol{T})$, a random quantity through $\boldsymbol{T}$. For each $(z, \boldsymbol{v}) \in \mathcal{T}$, define the mean of the observed outcomes within the cell corresponding to profile $(z, \boldsymbol{v})$ as

$$\hat{\mu}(z, \boldsymbol{v}) \coloneqq \frac{1}{n_{z,\boldsymbol{v}}} \sum_{i=1}^{N} Y_i \, \mathbb{1}\left\{\boldsymbol{T}_i = (z, \boldsymbol{v})\right\}, \tag{S15}$$

which is a function of the random configuration assignment $\boldsymbol{T}$. The plug-in estimator $\widehat{\theta}$ is then

$$\widehat{\theta} \coloneqq \sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})\hat{\mu}(1, \boldsymbol{v}) - \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})\hat{\mu}(0, \boldsymbol{v}). \tag{S16}$$

All randomness in $\widehat{\theta}$ arises from the random assignment $\boldsymbol{T}$, which enters through the cell means $\hat{\mu}(z, \boldsymbol{v})$.

Lemma S1 (Randomization moments)**.** *Under complete random assignment (Assump-*

tion 3), for every unit $i \in \{1, \ldots, N\}$ and every profile $(z, \boldsymbol{v}) \in \mathcal{T}$,

$$\mathrm{E}\left[\mathbb{1}\{\boldsymbol{T}_i = (z, \boldsymbol{v})\}\right] = \frac{n_{z,\boldsymbol{v}}}{N} \tag{S17}$$

$$\mathrm{Var}\left[\mathbb{1}\{\boldsymbol{T}_i = (z, \boldsymbol{v})\}\right] = \frac{n_{z,\boldsymbol{v}}(N - n_{z,\boldsymbol{v}})}{N^2} \tag{S18}$$

and, for any units i and j and any profiles $(z, \boldsymbol{v}), (z', \boldsymbol{v}') \in \mathcal{T}$,

$$\mathrm{Cov}\left[\mathbb{1}\left\{\boldsymbol{T}_i = (z, \boldsymbol{v})\right\}, \mathbb{1}\left\{\boldsymbol{T}_j = (z, \boldsymbol{v})\right\}\right] = -\frac{n_{z,\boldsymbol{v}}(N - n_{z,\boldsymbol{v}})}{N^2(N-1)} \text{ for } i \neq j \tag{S19}$$

$$\mathrm{Cov}\left[\mathbb{1}\left\{\boldsymbol{T}_i = (z, \boldsymbol{v})\right\}, \mathbb{1}\left\{\boldsymbol{T}_j = (z', \boldsymbol{v}')\right\}\right] = \frac{n_{z,\boldsymbol{v}} n_{z',\boldsymbol{v}'}}{N^2(N-1)} \text{ for } i \neq j \text{ and } (z, \boldsymbol{v}) \neq (z', \boldsymbol{v}') \tag{S20}$$

$$\mathrm{Cov}\left[\mathbb{1}\{\boldsymbol{T}_i = (z, \boldsymbol{v})\}, \mathbb{1}\{\boldsymbol{T}_i = (z', \boldsymbol{v}')\}\right] = -\frac{n_{z,\boldsymbol{v}} n_{z',\boldsymbol{v}'}}{N^2} \text{ for } (z, \boldsymbol{v}) \neq (z', \boldsymbol{v}'). \tag{S21}$$

Proof. The proof computes each moment by counting assignments. The counting is justified by the design: Complete random assignment places equal probability $1/|\Omega|$ on every assignment in Ω, so the probability of any event equals the number of assignments in the event divided by $|\Omega|$. Every moment in the lemma reduces to such a probability because the expectation of a product of indicators is the probability that all of the indicated events occur. Each count then follows from fixing the coordinates of the units that the event names and distributing the remaining units over the remaining places in each cell.

Fix a unit $i \in \{1, \ldots, N\}$ and a profile $(z, \boldsymbol{v}) \in \mathcal{T}$. Then

$$\begin{aligned}\mathrm{E}\left[\mathbb{1}\{\boldsymbol{T}_i = (z, \boldsymbol{v})\}\right] &= \Pr\left(\boldsymbol{T}_i = (z, \boldsymbol{v})\right) \\ &= \sum_{\boldsymbol{t} \in \Omega} \mathbb{1}\{\boldsymbol{t}_i = (z, \boldsymbol{v})\} \Pr(\boldsymbol{T} = \boldsymbol{t}).\end{aligned}$$

Under complete random assignment (Assumption 3), $\Pr(\boldsymbol{T} = \boldsymbol{t}) = 1/|\Omega|$ for all $\boldsymbol{t} \in \Omega$. Therefore,

$$\Pr\left(\boldsymbol{T}_i = (z, \boldsymbol{v})\right) = \frac{1}{|\Omega|} \sum_{\boldsymbol{t} \in \Omega} \mathbb{1}\{\boldsymbol{t}_i = (z, \boldsymbol{v})\}.$$

To evaluate the number of assignments in Ω for which $\boldsymbol{t}_i = (z, \boldsymbol{v})$, note that fixing $\boldsymbol{t}_i = (z, \boldsymbol{v})$ reduces the remaining counts to $n_{z,\boldsymbol{v}} - 1$ in cell $(z, \boldsymbol{v})$ and $n_{z',\boldsymbol{v}'}$ in every other cell

$(z', \boldsymbol{v}') \neq (z, \boldsymbol{v})$. Thus,

$$\sum_{\boldsymbol{t} \in \Omega} \mathbb{1}\{\boldsymbol{t}_i = (z, \boldsymbol{v})\} = \frac{(N-1)!}{(n_{z,\boldsymbol{v}} - 1)! \prod_{(z',\boldsymbol{v}') \neq (z,\boldsymbol{v})} n_{z',\boldsymbol{v}'}!}.$$

Writing the size of Ω, stated in the introduction to this subsection, with the factor for cell $(z, \boldsymbol{v})$ separated from the product of factorials,

$$|\Omega| = \frac{N!}{n_{z,\boldsymbol{v}}! \prod_{(z',\boldsymbol{v}') \neq (z,\boldsymbol{v})} n_{z',\boldsymbol{v}'}!},$$

which implies that

$$\begin{aligned}
\Pr\left(\boldsymbol{T}_i = (z, \boldsymbol{v})\right) &= \frac{1}{|\Omega|} \sum_{\boldsymbol{t} \in \Omega} \mathbb{1}\{\boldsymbol{t}_i = (z, \boldsymbol{v})\} \\
&= \frac{(N-1)!}{(n_{z,\boldsymbol{v}} - 1)! \prod_{(z',\boldsymbol{v}') \neq (z,\boldsymbol{v})} n_{z',\boldsymbol{v}'}!} \frac{n_{z,\boldsymbol{v}}! \prod_{(z',\boldsymbol{v}') \neq (z,\boldsymbol{v})} n_{z',\boldsymbol{v}'}!}{N!} \\
&= \frac{n_{z,\boldsymbol{v}}}{N},
\end{aligned}$$

which establishes (S17).

Because an indicator equals its own square, the definition of variance then gives

$$\begin{aligned}
\operatorname{Var}\left[\mathbb{1}\{\boldsymbol{T}_i = (z, \boldsymbol{v})\}\right] &= \mathrm{E}\left[\mathbb{1}\{\boldsymbol{T}_i = (z, \boldsymbol{v})\}^2\right] - \mathrm{E}\left[\mathbb{1}\{\boldsymbol{T}_i = (z, \boldsymbol{v})\}\right]^2 \\
&= \mathrm{E}\left[\mathbb{1}\{\boldsymbol{T}_i = (z, \boldsymbol{v})\}\right] - \mathrm{E}\left[\mathbb{1}\{\boldsymbol{T}_i = (z, \boldsymbol{v})\}\right]^2 \\
&= \frac{n_{z,\boldsymbol{v}}}{N} - \left(\frac{n_{z,\boldsymbol{v}}}{N}\right)^2 \\
&= \frac{n_{z,\boldsymbol{v}}(N - n_{z,\boldsymbol{v}})}{N^2},
\end{aligned}$$

which establishes (S18).

Next, fix distinct units $i \neq j$. For the same profile $(z, \boldsymbol{v})$, the product $\mathbb{1}\{\boldsymbol{T}_i = (z, \boldsymbol{v})\}\, \mathbb{1}\{\boldsymbol{T}_j = (z, \boldsymbol{v})\}$ equals 1 exactly when both units fall in cell $(z, \boldsymbol{v})$, so its expectation is the joint probability $\Pr\left(\boldsymbol{T}_i = (z, \boldsymbol{v}), \boldsymbol{T}_j = (z, \boldsymbol{v})\right)$. The count parallels the derivation of (S17): Fixing $\boldsymbol{t}_i = \boldsymbol{t}_j = (z, \boldsymbol{v})$ reduces the remaining counts to $n_{z,\boldsymbol{v}} - 2$ in cell $(z, \boldsymbol{v})$ and

$n_{z',\boldsymbol{v}'}$ in every other cell $(z',\boldsymbol{v}') \neq (z,\boldsymbol{v})$, so

$$\begin{aligned}\Pr\left(\boldsymbol{T}_i = (z,\boldsymbol{v}), \boldsymbol{T}_j = (z,\boldsymbol{v})\right) &= \frac{(N-2)!}{(n_{z,\boldsymbol{v}}-2)!\prod_{(z',\boldsymbol{v}')\neq(z,\boldsymbol{v})} n_{z',\boldsymbol{v}'}!} \frac{n_{z,\boldsymbol{v}}!\prod_{(z',\boldsymbol{v}')\neq(z,\boldsymbol{v})} n_{z',\boldsymbol{v}'}!}{N!} \\ &= \frac{(N-2)!}{N!}\frac{n_{z,\boldsymbol{v}}!}{(n_{z,\boldsymbol{v}}-2)!} \\ &= \frac{n_{z,\boldsymbol{v}}(n_{z,\boldsymbol{v}}-1)}{N(N-1)}.\end{aligned}$$

The count assumes $n_{z,\boldsymbol{v}} \geq 2$; when $n_{z,\boldsymbol{v}} = 1$, no assignment places two distinct units in the cell, so the joint probability equals 0, which the final expression also gives. Therefore,

$$\begin{aligned}\operatorname{Cov}\left[\mathbb{1}\{\boldsymbol{T}_i = (z,\boldsymbol{v})\}, \mathbb{1}\{\boldsymbol{T}_j = (z,\boldsymbol{v})\}\right] &= \mathrm{E}\left[\mathbb{1}\{\boldsymbol{T}_i = (z,\boldsymbol{v})\}\mathbb{1}\{\boldsymbol{T}_j = (z,\boldsymbol{v})\}\right] \\ &\quad - \mathrm{E}\left[\mathbb{1}\{\boldsymbol{T}_i = (z,\boldsymbol{v})\}\right]\mathrm{E}\left[\mathbb{1}\{\boldsymbol{T}_j = (z,\boldsymbol{v})\}\right] \\ &= \frac{n_{z,\boldsymbol{v}}(n_{z,\boldsymbol{v}}-1)}{N(N-1)} - \frac{n_{z,\boldsymbol{v}}^2}{N^2} \\ &= -\frac{n_{z,\boldsymbol{v}}(N-n_{z,\boldsymbol{v}})}{N^2(N-1)},\end{aligned}$$

which establishes (S19).

For $(z,\boldsymbol{v}) \neq (z',\boldsymbol{v}')$, still with $i \neq j$, the count is analogous: Fixing $\boldsymbol{t}_i = (z,\boldsymbol{v})$ and $\boldsymbol{t}_j = (z',\boldsymbol{v}')$ reduces the remaining counts to $n_{z,\boldsymbol{v}} - 1$ and $n_{z',\boldsymbol{v}'} - 1$ in those two cells, so

$$\mathrm{E}\left[\mathbb{1}\{\boldsymbol{T}_i = (z,\boldsymbol{v})\}\,\mathbb{1}\{\boldsymbol{T}_j = (z',\boldsymbol{v}')\}\right] = \Pr\left(\boldsymbol{T}_i = (z,\boldsymbol{v}), \boldsymbol{T}_j = (z',\boldsymbol{v}')\right) = \frac{n_{z,\boldsymbol{v}}}{N}\frac{n_{z',\boldsymbol{v}'}}{N-1}.$$

Therefore,

$$\begin{aligned}\operatorname{Cov}\left[\mathbb{1}\{\boldsymbol{T}_i = (z,\boldsymbol{v})\}, \mathbb{1}\{\boldsymbol{T}_j = (z',\boldsymbol{v}')\}\right] &= \frac{n_{z,\boldsymbol{v}}}{N}\frac{n_{z',\boldsymbol{v}'}}{N-1} - \frac{n_{z,\boldsymbol{v}}}{N}\frac{n_{z',\boldsymbol{v}'}}{N} \\ &= \frac{n_{z,\boldsymbol{v}}n_{z',\boldsymbol{v}'}}{N^2(N-1)},\end{aligned}$$

which establishes (S20).

Finally, fix a single unit i and two distinct profiles $(z,\boldsymbol{v}) \neq (z',\boldsymbol{v}')$. A unit receives exactly one profile, so the two indicator events are disjoint and the product of the two

indicators equals zero at every assignment in Ω:

$$\mathbb{1}\{\boldsymbol{T}_i = (z, \boldsymbol{v})\}\mathbb{1}\{\boldsymbol{T}_i = (z', \boldsymbol{v}')\} = 0.$$

Therefore,

$$\begin{aligned}\operatorname{Cov}\left[\mathbb{1}\{\boldsymbol{T}_i = (z, \boldsymbol{v})\}, \mathbb{1}\{\boldsymbol{T}_i = (z', \boldsymbol{v}')\}\right] &= 0 - \frac{n_{z,\boldsymbol{v}}}{N}\frac{n_{z',\boldsymbol{v}'}}{N} \\ &= -\frac{n_{z,\boldsymbol{v}} n_{z',\boldsymbol{v}'}}{N^2},\end{aligned}$$

which establishes (S21). □

The first proposition of this subsection establishes part *(i)* of Proposition 1 of the main text — the plug-in estimator $\widehat{\theta}$ is unbiased for the estimand θ — the part whose proof Section S1.3 defers to this subsection.

Proposition S3 (Unbiasedness of the plug-in estimator)**.** *Under Assumption S1, together with Assumption 3 of the main text,* $\widehat{\theta}$ *in* (S16) *is unbiased for* θ *in* (S13)*:*

$$\mathrm{E}\left[\widehat{\theta}\right] = \theta, \tag{S22}$$

where the expectation E *is taken with respect to the randomization distribution of the configuration assignment* $\boldsymbol{T}$*, which Assumption 3 fixes as the uniform distribution on* Ω*.*

Proof. The proof writes each cell mean as a sum of fixed potential outcomes multiplied by random assignment indicators and applies the expectation of each indicator from Lemma S1.

Under SUTVA (Assumption S1), $Y_i = y_i(\boldsymbol{T}) = y_i(\boldsymbol{T}_i)$, so

$$\hat{\mu}(z, \boldsymbol{v}) = \frac{1}{n_{z,\boldsymbol{v}}} \sum_{i=1}^{N} y_i(\boldsymbol{T}_i)\mathbb{1}\{\boldsymbol{T}_i = (z, \boldsymbol{v})\}.$$

Each summand is nonzero only when $\boldsymbol{T}_i = (z, \boldsymbol{v})$, so replacing the random argument $\boldsymbol{T}_i$ of

y_i with the fixed profile $(z, \boldsymbol{v})$ leaves every summand unchanged:

$$\hat{\mu}(z, \boldsymbol{v}) = \frac{1}{n_{z,\boldsymbol{v}}} \sum_{i=1}^{N} y_i(z, \boldsymbol{v}) \mathbb{1}\{\boldsymbol{T}_i = (z, \boldsymbol{v})\}. \tag{S23}$$

Since $g_1(\boldsymbol{v})$ and $g_0(\boldsymbol{v})$ are fixed constants for each attribute profile $\boldsymbol{v} \in \mathcal{V}$, the linearity of expectation gives

$$\mathrm{E}\left[\widehat{\theta}\right] = \sum_{\boldsymbol{v} \in \mathcal{V}} g_1(\boldsymbol{v}) \, \mathrm{E}\left[\hat{\mu}(1, \boldsymbol{v})\right] - \sum_{\boldsymbol{v} \in \mathcal{V}} g_0(\boldsymbol{v}) \, \mathrm{E}\left[\hat{\mu}(0, \boldsymbol{v})\right]. \tag{S24}$$

Under complete random assignment (Assumption 3), the cell sizes $n_{z,\boldsymbol{v}}$ are fixed across all $\boldsymbol{t} \in \Omega$, and the potential outcomes $y_i(z, \boldsymbol{v})$ are fixed features of the finite population. Substituting (S23) into (S24) and again applying the linearity of expectation therefore yields

$$\begin{aligned} \mathrm{E}\left[\widehat{\theta}\right] = & \sum_{\boldsymbol{v} \in \mathcal{V}} g_1(\boldsymbol{v}) \frac{1}{n_{1,\boldsymbol{v}}} \sum_{i=1}^{N} y_i(1, \boldsymbol{v}) \, \mathrm{E}\left[\mathbb{1}\{\boldsymbol{T}_i = (1, \boldsymbol{v})\}\right] \\ & - \sum_{\boldsymbol{v} \in \mathcal{V}} g_0(\boldsymbol{v}) \frac{1}{n_{0,\boldsymbol{v}}} \sum_{i=1}^{N} y_i(0, \boldsymbol{v}) \, \mathrm{E}\left[\mathbb{1}\{\boldsymbol{T}_i = (0, \boldsymbol{v})\}\right]. \end{aligned}$$

Then, under complete random assignment (Assumption 3), (S17) of Lemma S1 implies that

$$\mathrm{E}\left[\widehat{\theta}\right] = \sum_{\boldsymbol{v} \in \mathcal{V}} g_1(\boldsymbol{v}) \frac{1}{n_{1,\boldsymbol{v}}} \sum_{i=1}^{N} y_i(1, \boldsymbol{v}) \left(n_{1,\boldsymbol{v}}/N\right) - \sum_{\boldsymbol{v} \in \mathcal{V}} g_0(\boldsymbol{v}) \frac{1}{n_{0,\boldsymbol{v}}} \sum_{i=1}^{N} y_i(0, \boldsymbol{v}) \left(n_{0,\boldsymbol{v}}/N\right).$$

Since $\dfrac{1}{n_{z,\boldsymbol{v}}} \left(n_{z,\boldsymbol{v}}/N\right) = \dfrac{1}{N}$, each inner sum reduces to $\dfrac{1}{N} \sum_{i=1}^{N} y_i(z, \boldsymbol{v}) = \mu(z, \boldsymbol{v})$, so

$$\mathrm{E}\left[\widehat{\theta}\right] = \sum_{\boldsymbol{v} \in \mathcal{V}} g_1(\boldsymbol{v}) \mu(1, \boldsymbol{v}) - \sum_{\boldsymbol{v} \in \mathcal{V}} g_0(\boldsymbol{v}) \mu(0, \boldsymbol{v}),$$

which equals θ in (S13), thereby completing the proof. □

The second proposition of this subsection derives the exact variance of the plug-in estimator $\widehat{\theta}$. The variance involves two population quantities. For each profile $(z, \boldsymbol{v}) \in \mathcal{T}$,

define the finite-population variance of the potential outcomes at $(z, \boldsymbol{v})$ as

$$S^2(z, \boldsymbol{v}) \coloneqq \frac{1}{N-1} \sum_{i=1}^{N} \left[y_i(z, \boldsymbol{v}) - \mu(z, \boldsymbol{v}) \right]^2, \tag{S25}$$

and, for any two profiles $(z, \boldsymbol{v}), (z', \boldsymbol{v}') \in \mathcal{T}$, define the finite-population covariance of the potential outcomes at $(z, \boldsymbol{v})$ and $(z', \boldsymbol{v}')$ as

$$S(z, \boldsymbol{v}; z', \boldsymbol{v}') \coloneqq \frac{1}{N-1} \sum_{i=1}^{N} \left[y_i(z, \boldsymbol{v}) - \mu(z, \boldsymbol{v}) \right] \left[y_i(z', \boldsymbol{v}') - \mu(z', \boldsymbol{v}') \right]. \tag{S26}$$

Proposition S4 (Exact variance of the plug-in estimator)**.** *Under Assumption S1, together with Assumption 3 of the main text,*

$$\begin{aligned} \operatorname{Var}\left[\widehat{\theta}\right] = & \sum_{\boldsymbol{v} \in \mathcal{V}} g_1(\boldsymbol{v})^2 \frac{N - n_{1,\boldsymbol{v}}}{N} \frac{S^2(1, \boldsymbol{v})}{n_{1,\boldsymbol{v}}} + \sum_{\boldsymbol{v} \in \mathcal{V}} g_0(\boldsymbol{v})^2 \frac{N - n_{0,\boldsymbol{v}}}{N} \frac{S^2(0, \boldsymbol{v})}{n_{0,\boldsymbol{v}}} \\ & - \frac{1}{N} \sum_{\boldsymbol{v} \neq \boldsymbol{v}'} g_1(\boldsymbol{v}) g_1(\boldsymbol{v}') S(1, \boldsymbol{v}; 1, \boldsymbol{v}') - \frac{1}{N} \sum_{\boldsymbol{v} \neq \boldsymbol{v}'} g_0(\boldsymbol{v}) g_0(\boldsymbol{v}') S(0, \boldsymbol{v}; 0, \boldsymbol{v}') \\ & + \frac{2}{N} \sum_{\boldsymbol{v} \in \mathcal{V}} \sum_{\boldsymbol{v}' \in \mathcal{V}} g_1(\boldsymbol{v}) g_0(\boldsymbol{v}') S(1, \boldsymbol{v}; 0, \boldsymbol{v}'), \end{aligned}$$

where the variance Var *is taken with respect to the same randomization distribution of* $\boldsymbol{T}$ *as in Proposition S3.*

Proof. The proof decomposes the variance of $\widehat{\theta}$ into three components — the variance of the weighted sum under each racial condition and the covariance between the two weighted sums — and evaluates each component with the moments of the indicators from Lemma S1.

The estimator $\widehat{\theta}$ in (S16) is a linear combination of the cell means $\hat{\mu}(z, \boldsymbol{v})$ with fixed weights $g_1(\boldsymbol{v})$ and $-g_0(\boldsymbol{v})$. For fixed constants α_k and β_l and random variables X_k and Y_l, covariance distributes over weighted sums: $\operatorname{Cov}\left[\sum_k \alpha_k X_k, \sum_l \beta_l Y_l\right] = \sum_k \sum_l \alpha_k \beta_l \operatorname{Cov}\left[X_k, Y_l\right]$ (the bilinearity of covariance). The variance of a random variable is the covariance of the random variable with itself. Writing $\operatorname{Var}[\widehat{\theta}]$ as the covariance of $\widehat{\theta}$ with itself and applying the bilinearity of covariance to the difference of the two weighted

sums in (S16),

$$\begin{aligned}\operatorname{Var}\left[\widehat{\theta}\right] = \operatorname{Var}\left[\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})\hat{\mu}(1,\boldsymbol{v})\right] + \operatorname{Var}\left[\sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})\hat{\mu}(0,\boldsymbol{v})\right] \\ - 2\operatorname{Cov}\left[\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})\hat{\mu}(1,\boldsymbol{v}),\ \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})\hat{\mu}(0,\boldsymbol{v})\right].\end{aligned} \tag{S27}$$

Take the three components of (S27) one at a time. For the first two, again by the bilinearity of covariance,

$$\operatorname{Var}\left[\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})\hat{\mu}(1,\boldsymbol{v})\right] = \sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})^2 \operatorname{Var}\left[\hat{\mu}(1,\boldsymbol{v})\right] + \sum_{\boldsymbol{v}\neq\boldsymbol{v}'} g_1(\boldsymbol{v})g_1(\boldsymbol{v}')\operatorname{Cov}\left[\hat{\mu}(1,\boldsymbol{v}),\hat{\mu}(1,\boldsymbol{v}')\right] \tag{S28}$$

$$\operatorname{Var}\left[\sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})\hat{\mu}(0,\boldsymbol{v})\right] = \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})^2 \operatorname{Var}\left[\hat{\mu}(0,\boldsymbol{v})\right] + \sum_{\boldsymbol{v}\neq\boldsymbol{v}'} g_0(\boldsymbol{v})g_0(\boldsymbol{v}')\operatorname{Cov}\left[\hat{\mu}(0,\boldsymbol{v}),\hat{\mu}(0,\boldsymbol{v}')\right]. \tag{S29}$$

Under SUTVA (Assumption S1), the cell mean $\hat{\mu}(z,\boldsymbol{v})$ takes the form (S23), so

$$\operatorname{Var}\left[\hat{\mu}(z,\boldsymbol{v})\right] = \operatorname{Var}\left[\frac{1}{n_{z,\boldsymbol{v}}}\sum_{i=1}^{N} y_i(z,\boldsymbol{v})\mathbb{1}\{\boldsymbol{T}_i = (z,\boldsymbol{v})\}\right],$$

which, by the bilinearity of covariance, is

$$\begin{aligned}\operatorname{Var}\left[\hat{\mu}(z,\boldsymbol{v})\right] = \frac{1}{n_{z,\boldsymbol{v}}^2}\sum_{i=1}^{N} y_i(z,\boldsymbol{v})^2 \operatorname{Var}\left[\mathbb{1}\{\boldsymbol{T}_i = (z,\boldsymbol{v})\}\right] \\ + \frac{1}{n_{z,\boldsymbol{v}}^2}\sum_{i\neq j} y_i(z,\boldsymbol{v})y_j(z,\boldsymbol{v})\operatorname{Cov}\left[\mathbb{1}\{\boldsymbol{T}_i = (z,\boldsymbol{v})\},\mathbb{1}\{\boldsymbol{T}_j = (z,\boldsymbol{v})\}\right].\end{aligned}$$

Then, under complete random assignment (Assumption 3), (S18) and (S19) of Lemma S1 imply that

$$\begin{aligned}\operatorname{Var}\left[\hat{\mu}(z,\boldsymbol{v})\right] = \frac{1}{n_{z,\boldsymbol{v}}^2}\sum_{i=1}^{N} y_i(z,\boldsymbol{v})^2\left(\frac{n_{z,\boldsymbol{v}}(N-n_{z,\boldsymbol{v}})}{N^2}\right) \\ + \frac{1}{n_{z,\boldsymbol{v}}^2}\sum_{i\neq j} y_i(z,\boldsymbol{v})y_j(z,\boldsymbol{v})\left(-\frac{n_{z,\boldsymbol{v}}(N-n_{z,\boldsymbol{v}})}{N^2(N-1)}\right).\end{aligned}$$

Factoring $\dfrac{n_{z,\boldsymbol{v}}(N-n_{z,\boldsymbol{v}})}{n_{z,\boldsymbol{v}}^2 N^2} = \dfrac{N-n_{z,\boldsymbol{v}}}{n_{z,\boldsymbol{v}}N^2}$ out of both terms of the display above, so that the

two sums collect into a single factor in brackets,

$$\operatorname{Var}\left[\hat{\mu}(z,\boldsymbol{v})\right] = \frac{N - n_{z,\boldsymbol{v}}}{n_{z,\boldsymbol{v}} N^2}\left[\sum_{i=1}^{N} y_i(z,\boldsymbol{v})^2 - \frac{1}{N-1}\sum_{i\neq j} y_i(z,\boldsymbol{v}) y_j(z,\boldsymbol{v})\right]. \tag{S30}$$

The remaining steps rewrite the factor in brackets in (S30) as a multiple of a finite-population variance. Splitting the square of the sum $\left(\sum_{i=1}^{N} y_i(z,\boldsymbol{v})\right)^2 = \sum_{i=1}^{N} y_i(z,\boldsymbol{v})^2 + \sum_{i\neq j} y_i(z,\boldsymbol{v}) y_j(z,\boldsymbol{v})$ into its $i = j$ and $i \neq j$ terms and using $\sum_{i=1}^{N} y_i(z,\boldsymbol{v}) = N\mu(z,\boldsymbol{v})$,

$$\sum_{i\neq j} y_i(z,\boldsymbol{v}) y_j(z,\boldsymbol{v}) = N^2\mu(z,\boldsymbol{v})^2 - \sum_{i=1}^{N} y_i(z,\boldsymbol{v})^2.$$

Substituting this expression for $\sum_{i\neq j} y_i(z,\boldsymbol{v}) y_j(z,\boldsymbol{v})$ into the factor in brackets in (S30) and combining the two terms of that factor over the common denominator $N - 1$,

$$\sum_{i=1}^{N} y_i(z,\boldsymbol{v})^2 - \frac{N^2\mu(z,\boldsymbol{v})^2 - \sum_{i=1}^{N} y_i(z,\boldsymbol{v})^2}{N-1} = \frac{N\left[\sum_{i=1}^{N} y_i(z,\boldsymbol{v})^2 - N\mu(z,\boldsymbol{v})^2\right]}{N-1}.$$

Expanding the square in $\sum_{i=1}^{N}\left[y_i(z,\boldsymbol{v}) - \mu(z,\boldsymbol{v})\right]^2$ and again using $\sum_{i=1}^{N} y_i(z,\boldsymbol{v}) = N\mu(z,\boldsymbol{v})$ shows that $\sum_{i=1}^{N}\left[y_i(z,\boldsymbol{v}) - \mu(z,\boldsymbol{v})\right]^2 = \sum_{i=1}^{N} y_i(z,\boldsymbol{v})^2 - N\mu(z,\boldsymbol{v})^2$. The factor in brackets in (S30) therefore equals $\frac{N}{N-1}\sum_{i=1}^{N}\left[y_i(z,\boldsymbol{v}) - \mu(z,\boldsymbol{v})\right]^2$, which is $NS^2(z,\boldsymbol{v})$ by the definition of the finite-population variance in (S25), so that

$$\operatorname{Var}\left[\hat{\mu}(z,\boldsymbol{v})\right] = \frac{N - n_{z,\boldsymbol{v}}}{n_{z,\boldsymbol{v}} N^2}\cdot\frac{N}{N-1}\sum_{i=1}^{N}\left[y_i(z,\boldsymbol{v}) - \mu(z,\boldsymbol{v})\right]^2 = \frac{N - n_{z,\boldsymbol{v}}}{N}\frac{S^2(z,\boldsymbol{v})}{n_{z,\boldsymbol{v}}}. \tag{S31}$$

For the covariance terms in (S28) and (S29), fix z and $\boldsymbol{v} \neq \boldsymbol{v}'$. SUTVA (Assumption S1) and the bilinearity of covariance imply that

$$\begin{aligned}\operatorname{Cov}\left[\hat{\mu}(z,\boldsymbol{v}), \hat{\mu}(z,\boldsymbol{v}')\right] &= \operatorname{Cov}\left[\frac{1}{n_{z,\boldsymbol{v}}}\sum_{i=1}^{N} y_i(z,\boldsymbol{v})\mathbb{1}\{\boldsymbol{T}_i = (z,\boldsymbol{v})\},\ \frac{1}{n_{z,\boldsymbol{v}'}}\sum_{j=1}^{N} y_j(z,\boldsymbol{v}')\mathbb{1}\{\boldsymbol{T}_j = (z,\boldsymbol{v}')\}\right] \\ &= \frac{1}{n_{z,\boldsymbol{v}} n_{z,\boldsymbol{v}'}}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i(z,\boldsymbol{v}) y_j(z,\boldsymbol{v}')\operatorname{Cov}\left[\mathbb{1}\{\boldsymbol{T}_i = (z,\boldsymbol{v})\}, \mathbb{1}\{\boldsymbol{T}_j = (z,\boldsymbol{v}')\}\right].\end{aligned}$$

Then (S20) and (S21) of Lemma S1 imply that

$$\begin{aligned}\operatorname{Cov}\left[\hat{\mu}(z,\boldsymbol{v}),\hat{\mu}(z,\boldsymbol{v}')\right] &= \frac{1}{n_{z,\boldsymbol{v}}n_{z,\boldsymbol{v}'}}\left[\sum_{i=1}^{N} y_i(z,\boldsymbol{v})y_i(z,\boldsymbol{v}')\left(-\frac{n_{z,\boldsymbol{v}}n_{z,\boldsymbol{v}'}}{N^2}\right)\right.\\ &\qquad\left.+\sum_{i\neq j} y_i(z,\boldsymbol{v})y_j(z,\boldsymbol{v}')\frac{n_{z,\boldsymbol{v}}n_{z,\boldsymbol{v}'}}{N^2(N-1)}\right]\\ &= \frac{1}{N^2}\left[-\sum_{i=1}^{N} y_i(z,\boldsymbol{v})y_i(z,\boldsymbol{v}')+\frac{1}{N-1}\sum_{i\neq j} y_i(z,\boldsymbol{v})y_j(z,\boldsymbol{v}')\right]. \qquad \text{(S32)}\end{aligned}$$

The remaining steps rewrite the factor in brackets in (S32) as a multiple of a finite-population covariance. Splitting the product of the two sums into $i = j$ and $i \neq j$ terms and using $\sum_{i=1}^{N} y_i(z,\boldsymbol{v}) = N\mu(z,\boldsymbol{v})$,

$$\begin{aligned}\sum_{i\neq j} y_i(z,\boldsymbol{v})y_j(z,\boldsymbol{v}') &= \left(\sum_{i=1}^{N} y_i(z,\boldsymbol{v})\right)\left(\sum_{j=1}^{N} y_j(z,\boldsymbol{v}')\right)-\sum_{i=1}^{N} y_i(z,\boldsymbol{v})y_i(z,\boldsymbol{v}')\\ &= N^2\mu(z,\boldsymbol{v})\,\mu(z,\boldsymbol{v}')-\sum_{i=1}^{N} y_i(z,\boldsymbol{v})y_i(z,\boldsymbol{v}').\end{aligned}$$

Substituting this expression for $\sum_{i\neq j} y_i(z,\boldsymbol{v})y_j(z,\boldsymbol{v}')$ into the factor in brackets in (S32) and combining the two terms of that factor over the common denominator $N-1$, which gives the product term $\sum_{i=1}^{N} y_i(z,\boldsymbol{v})y_i(z,\boldsymbol{v}')$ the coefficient $-(N-1)-1=-N$,

$$\begin{aligned}&-\sum_{i=1}^{N} y_i(z,\boldsymbol{v})y_i(z,\boldsymbol{v}')+\frac{N^2\mu(z,\boldsymbol{v})\,\mu(z,\boldsymbol{v}')-\sum_{i=1}^{N} y_i(z,\boldsymbol{v})y_i(z,\boldsymbol{v}')}{N-1}\\ &=-\frac{N}{N-1}\left[\sum_{i=1}^{N} y_i(z,\boldsymbol{v})y_i(z,\boldsymbol{v}')-N\mu(z,\boldsymbol{v})\,\mu(z,\boldsymbol{v}')\right].\end{aligned}$$

Expanding the product of the two deviations from the cell means $\mu(z,\boldsymbol{v})$ and $\mu(z,\boldsymbol{v}')$ in each summand below and using the identities $\sum_{i=1}^{N} y_i(z,\boldsymbol{v}) = N\mu(z,\boldsymbol{v})$ and $\sum_{i=1}^{N} y_i(z,\boldsymbol{v}') = N\mu(z,\boldsymbol{v}')$ shows that

$$\sum_{i=1}^{N}\left[y_i(z,\boldsymbol{v})-\mu(z,\boldsymbol{v})\right]\left[y_i(z,\boldsymbol{v}')-\mu(z,\boldsymbol{v}')\right]=\sum_{i=1}^{N} y_i(z,\boldsymbol{v})y_i(z,\boldsymbol{v}')-N\mu(z,\boldsymbol{v})\,\mu(z,\boldsymbol{v}').$$

By the definition of the finite-population covariance in (S26), the left-hand side of the

display above equals $(N-1)\,S\,(z,\boldsymbol{v};z,\boldsymbol{v}')$, so the factor in brackets in (S32) equals $-\dfrac{N}{N-1}\cdot (N-1)\,S\,(z,\boldsymbol{v};z,\boldsymbol{v}') = -NS\,(z,\boldsymbol{v};z,\boldsymbol{v}')$. Restoring the leading factor $1/N^2$ that multiplies the brackets in (S32),

$$\operatorname{Cov}\left[\hat{\mu}(z,\boldsymbol{v}),\hat{\mu}(z,\boldsymbol{v}')\right] = -\frac{1}{N}S\,(z,\boldsymbol{v};z,\boldsymbol{v}')\,.$$

Consequently, the first two components of $\operatorname{Var}[\widehat{\theta}]$ in (S28) and (S29), respectively, are

$$\operatorname{Var}\left[\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})\hat{\mu}(1,\boldsymbol{v})\right] = \sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})^2\frac{N-n_{1,\boldsymbol{v}}}{N}\frac{S^2(1,\boldsymbol{v})}{n_{1,\boldsymbol{v}}} - \sum_{\boldsymbol{v}\neq\boldsymbol{v}'} g_1(\boldsymbol{v})g_1(\boldsymbol{v}')\frac{1}{N}S\,(1,\boldsymbol{v};1,\boldsymbol{v}') \tag{S33}$$

$$\operatorname{Var}\left[\sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})\hat{\mu}(0,\boldsymbol{v})\right] = \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})^2\frac{N-n_{0,\boldsymbol{v}}}{N}\frac{S^2(0,\boldsymbol{v})}{n_{0,\boldsymbol{v}}} - \sum_{\boldsymbol{v}\neq\boldsymbol{v}'} g_0(\boldsymbol{v})g_0(\boldsymbol{v}')\frac{1}{N}S\,(0,\boldsymbol{v};0,\boldsymbol{v}')\,. \tag{S34}$$

For the third component of $\operatorname{Var}[\widehat{\theta}]$ in (S27),

$$-2\operatorname{Cov}\left[\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})\hat{\mu}(1,\boldsymbol{v}),\ \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})\hat{\mu}(0,\boldsymbol{v})\right], \tag{S35}$$

the bilinearity of covariance implies that

$$\operatorname{Cov}\left[\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})\hat{\mu}(1,\boldsymbol{v}),\ \sum_{\boldsymbol{v}'\in\mathcal{V}} g_0(\boldsymbol{v}')\hat{\mu}(0,\boldsymbol{v}')\right] = \sum_{\boldsymbol{v}\in\mathcal{V}}\sum_{\boldsymbol{v}'\in\mathcal{V}} g_1(\boldsymbol{v})g_0(\boldsymbol{v}')\operatorname{Cov}\left[\hat{\mu}(1,\boldsymbol{v}),\hat{\mu}(0,\boldsymbol{v}')\right]. \tag{S36}$$

Then SUTVA (Assumption S1), together with (S20) and (S21) of Lemma S1, yields

$$\begin{aligned}
\operatorname{Cov}\left[\hat{\mu}(1,\boldsymbol{v}),\hat{\mu}(0,\boldsymbol{v}')\right] &= \frac{1}{n_{1,\boldsymbol{v}}n_{0,\boldsymbol{v}'}}\left[\sum_{i=1}^{N} y_i(1,\boldsymbol{v})y_i(0,\boldsymbol{v}')\left(-\frac{n_{1,\boldsymbol{v}}n_{0,\boldsymbol{v}'}}{N^2}\right)\right.\\
&\qquad \left. + \sum_{i\neq j} y_i(1,\boldsymbol{v})y_j(0,\boldsymbol{v}')\frac{n_{1,\boldsymbol{v}}n_{0,\boldsymbol{v}'}}{N^2(N-1)}\right]\\
&= \frac{1}{N^2}\left[-\sum_{i=1}^{N} y_i(1,\boldsymbol{v})y_i(0,\boldsymbol{v}') + \frac{1}{N-1}\sum_{i\neq j} y_i(1,\boldsymbol{v})y_j(0,\boldsymbol{v}')\right],
\end{aligned}$$

which reduces by the same three steps as the covariance $\operatorname{Cov}\left[\hat{\mu}(z,\boldsymbol{v}),\hat{\mu}(z,\boldsymbol{v}')\right]$ above — splitting the product of the two sums, combining over the common denominator $N-1$, and recognizing the finite-population covariance (S26) — now with $y_i(1,\boldsymbol{v})$ and $y_i(0,\boldsymbol{v}')$

in place of $y_i(z, \boldsymbol{v})$ and $y_i(z, \boldsymbol{v}')$. Restoring the leading factor $1/N^2$ that multiplies the brackets in the display above then yields

$$\operatorname{Cov}\left[\hat{\mu}(1, \boldsymbol{v}), \hat{\mu}(0, \boldsymbol{v}')\right] = -\frac{1}{N} S\left(1, \boldsymbol{v}; 0, \boldsymbol{v}'\right). \tag{S37}$$

Substituting (S37) into (S36) then yields

$$\operatorname{Cov}\left[\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})\hat{\mu}(1, \boldsymbol{v}),\ \sum_{\boldsymbol{v}'\in\mathcal{V}} g_0(\boldsymbol{v}')\hat{\mu}(0, \boldsymbol{v}')\right] = -\frac{1}{N}\sum_{\boldsymbol{v}\in\mathcal{V}}\sum_{\boldsymbol{v}'\in\mathcal{V}} g_1(\boldsymbol{v})g_0(\boldsymbol{v}')S(1, \boldsymbol{v}; 0, \boldsymbol{v}').$$

Multiplying the covariance by -2 shows that the third component (S35) of the variance decomposition in (S27) equals

$$\frac{2}{N}\sum_{\boldsymbol{v}\in\mathcal{V}}\sum_{\boldsymbol{v}'\in\mathcal{V}} g_1(\boldsymbol{v})g_0(\boldsymbol{v}')S(1, \boldsymbol{v}; 0, \boldsymbol{v}'). \tag{S38}$$

Finally, putting (S33), (S34), and (S38) together yields

$$\begin{aligned}\operatorname{Var}\left[\widehat{\theta}\right] = &\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})^2 \frac{N - n_{1,\boldsymbol{v}}}{N}\frac{S^2(1, \boldsymbol{v})}{n_{1,\boldsymbol{v}}} + \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})^2 \frac{N - n_{0,\boldsymbol{v}}}{N}\frac{S^2(0, \boldsymbol{v})}{n_{0,\boldsymbol{v}}} \\ &- \frac{1}{N}\sum_{\boldsymbol{v}\neq\boldsymbol{v}'} g_1(\boldsymbol{v})g_1(\boldsymbol{v}')S(1, \boldsymbol{v}; 1, \boldsymbol{v}') - \frac{1}{N}\sum_{\boldsymbol{v}\neq\boldsymbol{v}'} g_0(\boldsymbol{v})g_0(\boldsymbol{v}')S(0, \boldsymbol{v}; 0, \boldsymbol{v}') \\ &+ \frac{2}{N}\sum_{\boldsymbol{v}\in\mathcal{V}}\sum_{\boldsymbol{v}'\in\mathcal{V}} g_1(\boldsymbol{v})g_0(\boldsymbol{v}')S(1, \boldsymbol{v}; 0, \boldsymbol{v}'),\end{aligned} \tag{S39}$$

which completes the proof. □

The exact variance in Proposition S4 cannot be estimated. Each unit reveals the potential outcome of at most one cell — the fundamental problem of causal inference (Holland, 1986) — so no assignment identifies the cross-cell covariances $S(z, \boldsymbol{v}; z', \boldsymbol{v}')$, which pair potential outcomes that no unit exhibits together. Corollary S2 therefore replaces the exact variance with an upper bound whose every ingredient, including the correction term Γ, is a within-cell variance $S^2(z, \boldsymbol{v})$ or its square root, each identified by the units assigned to the corresponding cell. Confidence intervals computed from the upper bound are conservative

rather than exact.

Corollary S2. *Under Assumption S1, together with Assumption 3 of the main text,*

$$\operatorname{Var}\left[\widehat{\theta}\right] \leq \overline{\operatorname{Var}}\left[\widehat{\theta}\right] = \sum_{z\in\{0,1\}}\sum_{\boldsymbol{v}\in\mathcal{V}} \frac{g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v})}{n_{z,\boldsymbol{v}}} - \frac{1}{N}\max\{0,\,\Gamma\},$$

where $S(z,\boldsymbol{v}) \coloneqq \sqrt{S^2(z,\boldsymbol{v})}$ *and the correction term* Γ *is*

$$\Gamma \coloneqq 2\sum_{z\in\{0,1\}}\sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v}) - \left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v}) S\left(1,\boldsymbol{v}\right) + \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v}) S\left(0,\boldsymbol{v}\right)\right)^2.$$

Proof. From the expression for the variance in Proposition S4, $\operatorname{Var}[\widehat{\theta}]$ is

$$\operatorname{Var}\left[\widehat{\theta}\right] = \sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})^2 \frac{N-n_{1,\boldsymbol{v}}}{N}\frac{S^2(1,\boldsymbol{v})}{n_{1,\boldsymbol{v}}} - \frac{1}{N}\sum_{\boldsymbol{v}\neq\boldsymbol{v}'} g_1(\boldsymbol{v})g_1(\boldsymbol{v}')S(1,\boldsymbol{v};1,\boldsymbol{v}') \tag{S40}$$

$$+ \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})^2 \frac{N-n_{0,\boldsymbol{v}}}{N}\frac{S^2(0,\boldsymbol{v})}{n_{0,\boldsymbol{v}}} - \frac{1}{N}\sum_{\boldsymbol{v}\neq\boldsymbol{v}'} g_0(\boldsymbol{v})g_0(\boldsymbol{v}')S(0,\boldsymbol{v};0,\boldsymbol{v}') \tag{S41}$$

$$+ \frac{2}{N}\sum_{\boldsymbol{v}\in\mathcal{V}}\sum_{\boldsymbol{v}'\in\mathcal{V}} g_1(\boldsymbol{v})g_0(\boldsymbol{v}')S(1,\boldsymbol{v};0,\boldsymbol{v}'). \tag{S42}$$

The proof proceeds in three steps. Step 1 rewrites the exact variance so that every unidentified covariance is collected into a single nonnegative quantity, the variance of the unit-level contrasts; dropping that quantity gives the first upper bound. Step 2 instead bounds each unidentified covariance by Cauchy-Schwarz, which gives the second upper bound and the correction term Γ. Step 3 takes the smaller of the two upper bounds, which is the bound the corollary states.

Step 1: an exact rewriting of the variance. For each unit i, define the unit-level contrast

$$\psi_i \coloneqq \sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})\, y_i(1,\boldsymbol{v}) - \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})\, y_i(0,\boldsymbol{v}).$$

Exchanging the finite sums over i and $\boldsymbol{v}$ shows that the mean $\bar{\psi} \coloneqq (1/N)\sum_{i=1}^N \psi_i$ equals θ in (S13), so the finite-population variance $S^2(\psi) \coloneqq \frac{1}{N-1}\sum_{i=1}^N(\psi_i-\bar{\psi})^2$ measures the spread of the unit-level contrasts around the estimand. By the bilinearity of the finite-population covariance, substituting the definition of ψ_i into $S^2(\psi)$ and multiplying out the squared

difference of the two weighted sums yields one term per pair of cells, with diagonal terms $S(z, \boldsymbol{v}; z, \boldsymbol{v}) = S^2(z, \boldsymbol{v})$:

$$\begin{aligned} S^2(\psi) = \sum_{z\in\{0,1\}} \sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})^2 S^2(z, \boldsymbol{v}) + \sum_{z\in\{0,1\}} \sum_{\boldsymbol{v}\neq\boldsymbol{v}'} g_z(\boldsymbol{v}) g_z(\boldsymbol{v}') S(z, \boldsymbol{v}; z, \boldsymbol{v}') \\ - 2 \sum_{\boldsymbol{v}\in\mathcal{V}} \sum_{\boldsymbol{v}'\in\mathcal{V}} g_1(\boldsymbol{v}) g_0(\boldsymbol{v}') S(1, \boldsymbol{v}; 0, \boldsymbol{v}'). \end{aligned} \tag{S43}$$

Since $(N - n_{z,\boldsymbol{v}})/N n_{z,\boldsymbol{v}} = 1/n_{z,\boldsymbol{v}} - 1/N$, the first term of (S40) and of (S41) decomposes as

$$\sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})^2 \frac{N - n_{z,\boldsymbol{v}}}{N} \frac{S^2(z, \boldsymbol{v})}{n_{z,\boldsymbol{v}}} = \sum_{\boldsymbol{v}\in\mathcal{V}} \frac{g_z(\boldsymbol{v})^2 S^2(z, \boldsymbol{v})}{n_{z,\boldsymbol{v}}} - \frac{1}{N} \sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})^2 S^2(z, \boldsymbol{v}), \tag{S44}$$

so (S40)–(S42) regroup as the cell-variance term $\sum_{z\in\{0,1\}} \sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})^2 S^2(z, \boldsymbol{v})/n_{z,\boldsymbol{v}}$ minus $1/N$ times the right-hand side of (S43). The regrouping matches term by term: The decomposition (S44) supplies the $-1/N$ multiples of the diagonal terms $g_z(\boldsymbol{v})^2 S^2(z, \boldsymbol{v})$; the second terms of (S40) and (S41) are the $-1/N$ multiples of the within-condition covariances; and (S42), with its sign flipped by the leading minus, supplies the $+2/N$ multiples of the cross-condition covariances. The regrouping yields the identity

$$\operatorname{Var}\left[\widehat{\theta}\right] = \sum_{z\in\{0,1\}} \sum_{\boldsymbol{v}\in\mathcal{V}} \frac{g_z(\boldsymbol{v})^2 S^2(z, \boldsymbol{v})}{n_{z,\boldsymbol{v}}} - \frac{S^2(\psi)}{N}, \tag{S45}$$

the multi-cell analogue of the classical decomposition of the variance of a difference in means, in which $S^2(\psi)$ plays the role of the variance of the unit-level effects. Every unidentified covariance now enters the variance only through $S^2(\psi)$. As a sum of squared deviations divided by $N-1$, $S^2(\psi) \geq 0$, so dropping the term $-S^2(\psi)/N$ from the identity (S45) can only increase the right-hand side. Dropping the term gives the first upper bound,

$$\operatorname{Var}\left[\widehat{\theta}\right] \leq \sum_{z\in\{0,1\}} \sum_{\boldsymbol{v}\in\mathcal{V}} \frac{g_z(\boldsymbol{v})^2 S^2(z, \boldsymbol{v})}{n_{z,\boldsymbol{v}}}. \tag{S46}$$

Step 2: the Cauchy-Schwarz bound. The second route bounds the unidentified covariances in (S40)–(S42) one pair of cells at a time, replacing each covariance by a product of

identified standard deviations. Fix $z \in \{0,1\}$. By Cauchy-Schwarz,

$$|S(z,\boldsymbol{v};z,\boldsymbol{v}')| \leq \sqrt{S^2(z,\boldsymbol{v})S^2(z,\boldsymbol{v}')},$$

and in particular

$$S(z,\boldsymbol{v};z,\boldsymbol{v}') \geq -\sqrt{S^2(z,\boldsymbol{v})S^2(z,\boldsymbol{v}')}.$$

The lower bound is the direction the proof needs because the within-condition covariances enter (S40) and (S41) with a minus sign. Multiplying the lower bound by the nonnegative product of weights $g_z(\boldsymbol{v})\, g_z(\boldsymbol{v}')$ preserves the bound's direction, negating turns the lower bound into an upper bound, and summing over the pairs $\boldsymbol{v} \neq \boldsymbol{v}'$ preserves the upper bound, so

$$\begin{aligned} -\frac{1}{N}\sum_{\boldsymbol{v}\neq\boldsymbol{v}'} g_z(\boldsymbol{v})g_z(\boldsymbol{v}')S(z,\boldsymbol{v};z,\boldsymbol{v}') &\leq \frac{1}{N}\sum_{\boldsymbol{v}\neq\boldsymbol{v}'} g_z(\boldsymbol{v})g_z(\boldsymbol{v}')\sqrt{S^2(z,\boldsymbol{v})S^2(z,\boldsymbol{v}')} \\ &= \frac{1}{N}\sum_{\boldsymbol{v}\neq\boldsymbol{v}'} g_z(\boldsymbol{v})g_z(\boldsymbol{v}')S(z,\boldsymbol{v})S(z,\boldsymbol{v}'), \end{aligned} \tag{S47}$$

where the equality uses the definition $S(z,\boldsymbol{v}) = \sqrt{S^2(z,\boldsymbol{v})}$ from the statement of the corollary.

Then, by the decomposition (S44) of the first term of (S40) and (S41), together with the identity

$$\sum_{\boldsymbol{v}\neq\boldsymbol{v}'} g_z(\boldsymbol{v})S(z,\boldsymbol{v})g_z(\boldsymbol{v}')S(z,\boldsymbol{v}') = \left(\sum_{\boldsymbol{v}} g_z(\boldsymbol{v})S(z,\boldsymbol{v})\right)^2 - \sum_{\boldsymbol{v}} g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v}), \tag{S48}$$

it follows that

$$\frac{1}{N}\sum_{\boldsymbol{v}\neq\boldsymbol{v}'} g_z(\boldsymbol{v})g_z(\boldsymbol{v}')S(z,\boldsymbol{v})S(z,\boldsymbol{v}') = \frac{1}{N}\left[\left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})S(z,\boldsymbol{v})\right)^2 - \sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v})\right].$$

Combining, for fixed z, the decomposition (S44), the covariance bound (S47), and the

identity (S48) bounds each of (S40) and (S41) by identified quantities alone:

$$\begin{aligned}
&\sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})^2 \frac{N-n_{z,\boldsymbol{v}}}{N}\frac{S^2(z,\boldsymbol{v})}{n_{z,\boldsymbol{v}}} - \frac{1}{N}\sum_{\boldsymbol{v}\neq\boldsymbol{v}'} g_z(\boldsymbol{v})g_z(\boldsymbol{v}')S(z,\boldsymbol{v};z,\boldsymbol{v}') \\
&\leq \sum_{\boldsymbol{v}\in\mathcal{V}} \frac{g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v})}{n_{z,\boldsymbol{v}}} - \frac{1}{N}\sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v}) + \frac{1}{N}\left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})S(z,\boldsymbol{v})\right)^2 - \frac{1}{N}\sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v}) \\
&= \sum_{\boldsymbol{v}\in\mathcal{V}} \frac{g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v})}{n_{z,\boldsymbol{v}}} - \frac{2}{N}\sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v}) + \frac{1}{N}\left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})S(z,\boldsymbol{v})\right)^2,
\end{aligned}$$

where the equality collects the two $-\frac{1}{N}\sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v})$ terms. Therefore, at $z=1$ and $z=0$ respectively, (S40) and (S41) are bounded above by

$$\sum_{\boldsymbol{v}\in\mathcal{V}} \frac{g_1(\boldsymbol{v})^2 S^2(1,\boldsymbol{v})}{n_{1,\boldsymbol{v}}} - \frac{2}{N}\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})^2 S^2(1,\boldsymbol{v}) + \frac{1}{N}\left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})S(1,\boldsymbol{v})\right)^2 \tag{S49}$$

and

$$\sum_{\boldsymbol{v}\in\mathcal{V}} \frac{g_0(\boldsymbol{v})^2 S^2(0,\boldsymbol{v})}{n_{0,\boldsymbol{v}}} - \frac{2}{N}\sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})^2 S^2(0,\boldsymbol{v}) + \frac{1}{N}\left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})S(0,\boldsymbol{v})\right)^2, \tag{S50}$$

respectively.

For (S42), Cauchy-Schwarz implies that

$$|S(1,\boldsymbol{v};0,\boldsymbol{v}')| \leq \sqrt{S^2(1,\boldsymbol{v})S^2(0,\boldsymbol{v}')} = S(1,\boldsymbol{v})S(0,\boldsymbol{v}'),$$

and, since $g_1(\boldsymbol{v}), g_0(\boldsymbol{v}') \geq 0$,

$$\begin{aligned}
\frac{2}{N}\sum_{\boldsymbol{v},\boldsymbol{v}'} g_1(\boldsymbol{v})g_0(\boldsymbol{v}')S(1,\boldsymbol{v};0,\boldsymbol{v}') &\leq \frac{2}{N}\sum_{\boldsymbol{v},\boldsymbol{v}'} g_1(\boldsymbol{v})g_0(\boldsymbol{v}')S(1,\boldsymbol{v})S(0,\boldsymbol{v}') \\
&= \frac{2}{N}\left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})S(1,\boldsymbol{v})\right)\left(\sum_{\boldsymbol{v}'\in\mathcal{V}} g_0(\boldsymbol{v}')S(0,\boldsymbol{v}')\right).
\end{aligned} \tag{S51}$$

Summing the three upper bounds — (S49) for (S40), (S50) for (S41), and (S51) for

(S42) — bounds the whole variance:

$$\begin{aligned}\operatorname{Var}\left[\widehat{\theta}\right] \leq \sum_{z\in\{0,1\}}\sum_{\boldsymbol{v}\in\mathcal{V}} \frac{g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v})}{n_{z,\boldsymbol{v}}} - \frac{2}{N}\sum_{z\in\{0,1\}}\sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v}) \\ + \frac{1}{N}\left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})S(1,\boldsymbol{v})\right)^2 + \frac{1}{N}\left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})S(0,\boldsymbol{v})\right)^2 \\ + \frac{2}{N}\left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})S(1,\boldsymbol{v})\right)\left(\sum_{\boldsymbol{v}'\in\mathcal{V}} g_0(\boldsymbol{v}')S(0,\boldsymbol{v}')\right).\end{aligned}$$

Finally, the three square terms in the bound above collect into a single square,

$$\begin{aligned}&\frac{1}{N}\left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})S(1,\boldsymbol{v})\right)^2 + \frac{1}{N}\left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})S(0,\boldsymbol{v})\right)^2 \\ &+ \frac{2}{N}\left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})S(1,\boldsymbol{v})\right)\left(\sum_{\boldsymbol{v}'\in\mathcal{V}} g_0(\boldsymbol{v}')S(0,\boldsymbol{v}')\right) \\ &= \frac{1}{N}\left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})S(1,\boldsymbol{v}) + \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})S(0,\boldsymbol{v})\right)^2.\end{aligned}$$

By the definition of Γ in the statement of the corollary, the subtracted diagonal term $-\frac{2}{N}\sum_z\sum_{\boldsymbol{v}} g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v})$ and the collected square together equal $-\Gamma/N$. The substitution expresses the bound through Γ,

$$\operatorname{Var}\left[\widehat{\theta}\right] \leq \sum_{z\in\{0,1\}}\sum_{\boldsymbol{v}\in\mathcal{V}} \frac{g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v})}{n_{z,\boldsymbol{v}}} - \frac{\Gamma}{N}. \tag{S52}$$

By the identity (S45), the Cauchy-Schwarz bound (S52) is equivalent to the lower bound $S^2(\psi) \geq \Gamma$. The two steps thus bound the same unidentified quantity $S^2(\psi)$ from below, once by 0 and once by Γ, and neither lower bound is always the larger of the two because Γ may take either sign.

Step 3: combining the two bounds. Steps 1 and 2 each establish a valid upper bound on $\operatorname{Var}[\widehat{\theta}]$ — (S46) and (S52) — so the variance is at most the smaller of the two upper bounds. Subtracting $\max\{0,\Gamma\}/N$ from the cell-variance term subtracts the larger of the

two correction terms, 0 or Γ, and therefore selects the smaller upper bound:

$$\operatorname{Var}\left[\widehat{\theta}\right] \leq \sum_{z\in\{0,1\}} \sum_{\boldsymbol{v}\in\mathcal{V}} \frac{g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v})}{n_{z,\boldsymbol{v}}} - \frac{1}{N} \max\left\{0,\, \Gamma\right\},$$

which completes the proof. □

The identity (S45) is the multi-cell form of a decomposition that goes back to Neyman (1923) and appears, for two arms, as Theorem 6.2 of Imbens and Rubin (2015, p. 89) and, for a general contrast of the means of the arms of a multi-arm experiment, as Theorem 3 of Li and Ding (2017). Discarding the variance of the unit-level contrasts outright, as in (S46), is the standard route to a conservative variance in randomization-based inference: For two arms, the resulting bound is what the Neyman variance estimator estimates (Imbens and Rubin, 2015, Theorem 6.3 and pp. 92–93), and for the general contrast, the bound and its plug-in estimator appear in Li and Ding (2017, Proposition 3). Bounding unidentified covariances by Cauchy-Schwarz likewise goes back to Neyman (1923): Given only the identified second moments, the Cauchy-Schwarz bounds on an unidentified covariance are the sharpest available, and sharper bounds require bringing the marginal distributions of the potential outcomes to bear (Aronow et al., 2014, pp. 852–854).

How the two branches relate depends on the number of cells. With a single profile in $\mathcal{V}$ — the classical two-arm experiment — the correction term reduces to $\Gamma = \left[g_1(\boldsymbol{v})S(1,\boldsymbol{v}) - g_0(\boldsymbol{v})S(0,\boldsymbol{v})\right]^2 \geq 0$. The Cauchy-Schwarz branch is therefore never looser than the branch that discards $S^2(\psi)$, the two branches coincide exactly when $g_1(\boldsymbol{v})S(1,\boldsymbol{v}) = g_0(\boldsymbol{v})S(0,\boldsymbol{v})$, and the maximum always selects the Cauchy-Schwarz branch, which is the classical Neyman bound for a two-arm experiment (Aronow et al., 2014, p. 853). The two classical routes to a conservative variance do coincide in the two-arm case once Cauchy-Schwarz is completed, as in Neyman (1923), with the inequality of arithmetic and geometric means: The completion bounds $2g_1(\boldsymbol{v})g_0(\boldsymbol{v})S(1,\boldsymbol{v})S(0,\boldsymbol{v})$ by $g_1(\boldsymbol{v})^2S^2(1,\boldsymbol{v}) + g_0(\boldsymbol{v})^2S^2(0,\boldsymbol{v})$, and the resulting bound is exactly (S46), the bound that discarding $S^2(\psi)$ produces.

With two or more profiles, no ordering between the branches holds: Γ may take either sign, so either branch can be the tighter one, and the maximum genuinely chooses. Example S1 shows the two branches taking different values on one population, with the true variance below both.

Example S1 (The two branches diverging in a four-prosecutor population)**.** Simplify the charging example of Figure 3 in the main text to four prosecutors, indexed $i \in \{1, 2, 3, 4\}$. Each case file pairs a race, White ($z = 0$) or Black ($z = 1$), with a single nonracial feature, neighborhood disadvantage $\boldsymbol{v} \in \{0, 1\}$, coded $\boldsymbol{v} = 1$ for a disadvantaged neighborhood as in the worked examples above. The pairing gives four configurations. The design assigns each of the four configurations to exactly one of the four prosecutors, uniformly at random, so every cell count is $n_{z,\boldsymbol{v}} = 1$.

Cells of size one are permissible here because the example evaluates the exact variance and the two bounds, which are population quantities defined for any positive cell counts. The requirement $n_{z,\boldsymbol{v}} \geq 2$ applies only to the plug-in estimator of the variance bound, stated in (S53) below, which the example does not compute.

The two racial conditions share the uniform weighting of panel (a) of Figure 3, $g_1(\boldsymbol{v}) = g_0(\boldsymbol{v}) = 1/2$ at both values of $\boldsymbol{v}$, so the target is the AEE. The potential outcomes record whether each prosecutor would charge (1) or decline (0) each file: Prosecutor 1 would charge every file, prosecutors 2 and 3 would charge exactly the Black-marked file from the disadvantaged neighborhood, the configuration $(1, 1)$, and prosecutor 4 would charge no file.

	White ($z = 0$)		Black ($z = 1$)	
$y_i(z, \boldsymbol{v})$	$\boldsymbol{v} = 0$	$\boldsymbol{v} = 1$	$\boldsymbol{v} = 0$	$\boldsymbol{v} = 1$
Prosecutor 1	1	1	1	1
Prosecutor 2	0	0	0	1
Prosecutor 3	0	0	0	1
Prosecutor 4	0	0	0	0
Cell mean $\mu(z, \boldsymbol{v})$	1/4	1/4	1/4	3/4
Cell variance $S^2(z, \boldsymbol{v})$	1/4	1/4	1/4	1/4

Each column contains either one or three ones, so every cell mean is 1/4 or 3/4, every cell variance is $S^2(z, \boldsymbol{v}) = 1/4$, and $S(z, \boldsymbol{v}) = 1/2$ for every cell.

The two branches of the variance bound in Corollary S2 now separate. The branch that discards $S^2(\psi)$ gives the upper bound on $\mathrm{Var}[\widehat{\theta}]$

$$\sum_{z \in \{0,1\}} \sum_{\boldsymbol{v}} \frac{g_z(\boldsymbol{v})^2 S^2(z, \boldsymbol{v})}{n_{z,\boldsymbol{v}}} = 4 \cdot \left(\frac{1}{2}\right)^2 \cdot \frac{1/4}{1} = \frac{1}{4}.$$

The correction term is

$$\Gamma = 2 \sum_{z\in\{0,1\}} \sum_{\boldsymbol{v}} g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v}) - \left(\sum_{z\in\{0,1\}} \sum_{\boldsymbol{v}} g_z(\boldsymbol{v}) S(z,\boldsymbol{v}) \right)^2 = 2 \cdot \frac{1}{4} - \left(4 \cdot \frac{1}{2} \cdot \frac{1}{2}\right)^2 = -\frac{1}{2},$$

so the Cauchy-Schwarz branch gives the upper bound $\frac{1}{4} - \left(-\frac{1}{2}\right)/4 = \frac{1}{4} + \frac{1}{8} = \frac{3}{8}$ on $\mathrm{Var}[\widehat{\theta}]$.

For the exact variance, evaluate the unit-level contrasts $\psi_i = \frac{1}{2}[y_i(1,0) + y_i(1,1)] - \frac{1}{2}[y_i(0,0) + y_i(0,1)]$. The four values are $(0, \frac{1}{2}, \frac{1}{2}, 0)$, with mean $1/4$; the four deviations from the mean are $\pm 1/4$, so the variance of the unit-level contrasts is $S^2(\psi) = \frac{1}{3} \cdot 4 \cdot \frac{1}{16} = \frac{1}{12}$. The identity (S45) then gives

$$\mathrm{Var}\left[\widehat{\theta}\right] = \frac{1}{4} - \frac{1/12}{4} = \frac{11}{48}.$$

The same value follows from direct calculation over the set of assignments. The direct calculation uses only the definition of the variance over equally likely assignments, so agreement with the value above checks the identity (S45) and the expression for the variance in Proposition S4 on this population. Under Assumption 3 of the main text, the set of assignments Ω contains the $4! = 24$ ways to assign the four configurations to the four prosecutors, each with probability $1/24$.

Prosecutors 2 and 3 have identical potential outcomes, so $\widehat{\theta}$ depends on the assignment only through the configurations that prosecutors 1 and 4 read. Prosecutor 1 reads any one of the four configurations, and prosecutor 4 then reads any one of the remaining three, so the pair of configurations takes $4 \times 3 = 12$ equally likely values, each value collecting two of the 24 assignments. Each cell mean is the potential outcome of the lone prosecutor assigned to that configuration: The mean $\hat{\mu}(1,1)$ equals 1 unless prosecutor 4 reads the file $(1,1)$, and each of the other three cell means equals 1 exactly when prosecutor 1 reads the corresponding file. Therefore

$$\widehat{\theta} = \frac{1}{2}\Big[\mathbb{1}\{\text{prosecutor 4 does not read } (1,1)\} + \mathbb{1}\{\text{prosecutor 1 reads } (1,0)\} \\ - \mathbb{1}\{\text{prosecutor 1 reads } (0,1)\} - \mathbb{1}\{\text{prosecutor 1 reads } (0,0)\}\Big],$$

and evaluating the four indicators at each of the twelve values of the pair of configurations gives the value of $\widehat{\theta}$ at every assignment:

	Configuration read by prosecutor 4			
Configuration read by prosecutor 1	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$(0,0)$	—	0	0	$-1/2$
$(0,1)$	0	—	0	$-1/2$
$(1,0)$	1	1	—	$1/2$
$(1,1)$	$1/2$	$1/2$	$1/2$	—

The twelve equally likely entries give the distribution of $\widehat{\theta}$: the value 1 with probability 2/12, the value 1/2 with probability 4/12, the value 0 with probability 4/12, and the value $-1/2$ with probability 2/12. The mean over the set of assignments is

$$\mathrm{E}\left[\widehat{\theta}\right] = \frac{2 \cdot 1 + 4 \cdot \frac{1}{2} + 4 \cdot 0 + 2 \cdot \left(-\frac{1}{2}\right)}{12} = \frac{3}{12} = \frac{1}{4},$$

which equals the estimand $\theta = \frac{1}{2}\left[\mu(1,0) + \mu(1,1)\right] - \frac{1}{2}\left[\mu(0,0) + \mu(0,1)\right] = \frac{1}{4}$, as the unbiasedness of $\widehat{\theta}$ for θ in Proposition S3 guarantees. The second moment over the set of assignments is

$$\mathrm{E}\left[\widehat{\theta}^2\right] = \frac{2 \cdot 1 + 4 \cdot \frac{1}{4} + 4 \cdot 0 + 2 \cdot \frac{1}{4}}{12} = \frac{7/2}{12} = \frac{7}{24},$$

so the variance calculated directly over the set of assignments is

$$\mathrm{Var}\left[\widehat{\theta}\right] = \frac{7}{24} - \left(\frac{1}{4}\right)^2 = \frac{14}{48} - \frac{3}{48} = \frac{11}{48},$$

in exact agreement with the value that (S45) produced above from the expression for the variance in Proposition S4. The three quantities are ordered as

$$\mathrm{Var}\left[\widehat{\theta}\right] = \frac{11}{48} \;<\; \frac{12}{48} = \frac{1}{4} \text{ (discarding } S^2(\psi)) \;<\; \frac{18}{48} = \frac{3}{8} \text{ (Cauchy-Schwarz).}$$

Both branches bound the true variance, and the two upper bounds are not equal. Because $\Gamma < 0$, the maximum in Corollary S2 equals $\max\{0, \Gamma\} = 0$, so the corollary's bound is the smaller of the two upper bounds — here 12/48, from the branch that discards $S^2(\psi)$.

The example reverses the ordering of the two-arm case, in which the Cauchy-Schwarz branch is never the looser branch. The reversal has a plain source. Each branch replaces a quantity that the data cannot identify with the value that makes the resulting bound largest. The Cauchy-Schwarz branch replaces every cross-cell covariance at once: each within-condition covariance with its most negative possible value and each cross-condition covariance with its most positive possible value. Yet no single population of potential outcomes places all of the covariances at those extremes simultaneously when four cells have equal spread, so the Cauchy-Schwarz bound exceeds the exact variance of every such population. The branch that discards $S^2(\psi)$ replaces a single quantity, the variance of the unit-level contrasts, with zero, and a population attains zero whenever the unit-level contrasts ψ_i do not vary.

The Cauchy-Schwarz branch becomes the tighter one under a small change to the potential outcomes. Change the potential outcomes of prosecutor 1 so that prosecutor 1 would charge only the file $(1, 1)$; the other three prosecutors keep their potential outcomes.

No prosecutor would charge any file other than $(1,1)$, and the column of the configuration $(1,1)$ is unchanged. Every cell except $(1,1)$ is then constant, with $S^2(z,\boldsymbol{v}) = 0$, while $S^2(1,1) = 1/4$ as before. The branch that discards $S^2(\psi)$ gives the upper bound $(1/2)^2 \cdot \frac{1/4}{1} = \frac{1}{16}$. The correction term is now positive, $\Gamma = 2 \cdot \frac{1}{16} - \left(\frac{1}{2} \cdot \frac{1}{2}\right)^2 = \frac{1}{16}$, so the Cauchy-Schwarz branch gives the upper bound $\frac{1}{16} - \frac{1/16}{4} = \frac{3}{64}$, and the corollary's bound is the Cauchy-Schwarz branch. The unit-level contrasts are $\psi_i = \frac{1}{2}\, y_i(1,1)$, with variance $S^2(\psi) = \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{16} = \Gamma$, so the Cauchy-Schwarz bound equals the exact variance, 3/64. The mechanism from the four-varying-cell case runs in reverse: With a single varying cell, every cross-cell covariance is zero, so the Cauchy-Schwarz replacements introduce no slack at all, while the branch that discards $S^2(\psi)$ still replaces the positive variance $S^2(\psi) = 1/16$ with zero.

Now define a plug-in estimator of the variance bound, requiring $n_{z,\boldsymbol{v}} \geq 2$ for every $(z,\boldsymbol{v}) \in \mathcal{T}$ so that the sample variance defined below exists for every cell:

$$\widehat{\overline{\mathrm{Var}}}\left[\widehat{\theta}\right] = \sum_{z\in\{0,1\}} \sum_{\boldsymbol{v}\in\mathcal{V}} \frac{g_z(\boldsymbol{v})^2 \hat{S}^2(z,\boldsymbol{v})}{n_{z,\boldsymbol{v}}} - \frac{1}{N} \max\left\{0,\, \widehat{\Gamma}\right\}, \tag{S53}$$

$$\widehat{\Gamma} \coloneqq 2 \sum_{z\in\{0,1\}} \sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})^2 \hat{S}^2(z,\boldsymbol{v}) - \left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v}) \hat{S}\left(1,\boldsymbol{v}\right) + \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v}) \hat{S}\left(0,\boldsymbol{v}\right)\right)^2, \tag{S54}$$

where, for each profile $(z,\boldsymbol{v})$,

$$\hat{S}^2\left(z,\boldsymbol{v}\right) \coloneqq \frac{1}{n_{z,\boldsymbol{v}} - 1} \sum_{i=1}^{N} \mathbb{1}\{\boldsymbol{T}_i = (z,\boldsymbol{v})\} \left[Y_i - \hat{\mu}(z,\boldsymbol{v})\right]^2,$$

and $\hat{S}(z,\boldsymbol{v}) \coloneqq \sqrt{\hat{S}^2(z,\boldsymbol{v})}$.

Within each cell, the units assigned to the cell form a simple random sample without replacement from the N units, so the sample variance $\hat{S}^2(z,\boldsymbol{v})$ is unbiased for $S^2(z,\boldsymbol{v})$ by the classical result for sampling without replacement (Cochran, 1977, Theorem 2.4). The plug-in estimator of the variance bound as a whole is nonetheless not unbiased for $\overline{\mathrm{Var}}[\widehat{\theta}]$ in finite samples: The square root and the maximum are nonlinear, so Jensen's inequality introduces bias through each nonlinearity. The property established below for the plug-in estimator of the variance bound is therefore consistency, not unbiasedness.

The remaining results establish, over a sequence of finite populations of increasing size,

the consistency of the plug-in estimator $\widehat{\theta}$ for the limiting parameter, the consistency of the plug-in estimator of the variance bound for the scaled limiting bound, and a central limit theorem for $\widehat{\theta}$. The following regularity conditions govern the sequence.

Assumption S5 (Asymptotic regularity conditions)**.** *Consider a sequence of finite populations indexed by N, with potential outcomes $\{y_i(z, \boldsymbol{v}) : i = 1, \ldots, N;\ z \in \{0, 1\}, \boldsymbol{v} \in \mathcal{V}\}$ and treatment profiles assigned under complete random assignment with fixed cell counts $n_{z,\boldsymbol{v}}$ for $z \in \{0, 1\}$ and $\boldsymbol{v} \in \mathcal{V}$, as in Assumption 3. The potential outcomes, the cell counts, the moments below, and the estimator $\widehat{\theta}$ all depend on N. The notation suppresses the index except where an argument compares quantities across population sizes, as with the estimand θ_N, the limiting parameter θ_∞, and the scaled variance $\sigma^2_{\theta,N}$ below. Throughout the asymptotic results, a plain arrow $\to$ denotes the convergence of a nonrandom sequence of real numbers as $N \to \infty$. Suppose:*

(a) The set of attribute profiles $\mathcal{V}$ is finite, and for each profile $(z, \boldsymbol{v})$ there exists $\pi_{z,\boldsymbol{v},\infty} \in (0, 1)$ such that

$$\frac{n_{z,\boldsymbol{v}}}{N} \to \pi_{z,\boldsymbol{v},\infty} \quad \text{as } N \to \infty.$$

(b) For each profile $(z, \boldsymbol{v})$, the finite-population second moments are uniformly bounded:

$$\frac{1}{N}\sum_{i=1}^{N} y_i(z, \boldsymbol{v})^2 \le C_2$$

for some constant $C_2 < \infty$ that does not depend on N. The finite-population means $\mu(z, \boldsymbol{v})$, the finite-population variances $S^2(z, \boldsymbol{v})$ of (S25), and the finite-population covariances $S(z, \boldsymbol{v}; z', \boldsymbol{v}')$ of (S26) converge to finite limits, denoted

$$\mu_{z,\boldsymbol{v},\infty} \in \mathbb{R}, \qquad S^2_{z,\boldsymbol{v},\infty} \in [0, \infty), \qquad S_{z,\boldsymbol{v};z',\boldsymbol{v}',\infty} \in \mathbb{R},$$

and we write $S_{z,\boldsymbol{v},\infty} \coloneqq \sqrt{S^2_{z,\boldsymbol{v},\infty}}$. The Lindeberg-type condition

$$\max_{1 \le i \le N} \frac{[y_i(z, \boldsymbol{v}) - \mu(z, \boldsymbol{v})]^2}{N} \to 0$$

also holds for each profile $(z, \boldsymbol{v})$.

(c) The weights $g_z(\boldsymbol{v})$, for $z \in \{0, 1\}$ and $\boldsymbol{v} \in \mathcal{V}$, do not vary with N.

(d) (Nondegeneracy of the limiting variance.) Under (a)–(c), the scaled variance $N \operatorname{Var}\left[\widehat{\theta}\right]$ converges to a finite limit. By the expression for the variance of the

plug-in estimator in (S45)*,*

$$N \operatorname{Var}\left[\widehat{\theta}\right] = \sum_{z\in\{0,1\}} \sum_{\boldsymbol{v}\in\mathcal{V}} \frac{N}{n_{z,\boldsymbol{v}}} g_z(\boldsymbol{v})^2 S^2(z,\boldsymbol{v}) - S^2(\psi),$$

and each term on the right-hand side converges under (a)–(c): the ratios $N/n_{z,\boldsymbol{v}}$ *to* $1/\pi_{z,\boldsymbol{v},\infty}$ *under (a), the variances* $S^2(z,\boldsymbol{v})$ *to* $S^2_{z,\boldsymbol{v},\infty}$ *under (b), and* $S^2(\psi)$ *— a combination of the variances and covariances with weights fixed under (c), as* (S43) *shows — to the same combination of the limits in (b). Condition (d) itself is the assumption that the limit is strictly positive,*

$$\sigma^2_{\theta,\infty} \coloneqq \lim_{N\to\infty} N \operatorname{Var}\left[\widehat{\theta}\right] \in (0,\infty).$$

The convergence statements for random quantities in the results below — consistency and the central limit theorem — are with respect to the randomization distribution that complete random assignment induces, conditional on the potential outcomes.

Define the scaled variance and the scaled limiting bound,

$$\sigma^2_{\theta,N} \coloneqq N \operatorname{Var}\left[\widehat{\theta}\right], \qquad \overline{\sigma}^2_{\theta,\infty} \coloneqq \lim_{N\to\infty} N \overline{\operatorname{Var}}\left[\widehat{\theta}\right],$$

where the limit defining $\overline{\sigma}^2_{\theta,\infty}$ exists and is finite by Assumption S5(a)–(c) and Corollary S2. Explicitly,

$$\overline{\sigma}^2_{\theta,\infty} = \sum_{z\in\{0,1\}} \sum_{\boldsymbol{v}\in\mathcal{V}} \frac{g_z(\boldsymbol{v})^2 S^2_{z,\boldsymbol{v},\infty}}{\pi_{z,\boldsymbol{v},\infty}} - \max\{0,\, \Gamma_\infty\}, \tag{S55}$$

$$\Gamma_\infty \coloneqq 2 \sum_{z\in\{0,1\}} \sum_{\boldsymbol{v}\in\mathcal{V}} g_z(\boldsymbol{v})^2 S^2_{z,\boldsymbol{v},\infty} - \left(\sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v}) S_{1,\boldsymbol{v},\infty} + \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v}) S_{0,\boldsymbol{v},\infty}\right)^2, \tag{S56}$$

which follows from Assumption S5(a)–(c) by taking the limit of each term of the scaled bound $N\overline{\operatorname{Var}}[\widehat{\theta}]$: The finite sums converge term by term, with $N/n_{z,\boldsymbol{v}} \to 1/\pi_{z,\boldsymbol{v},\infty}$ and $S(z,\boldsymbol{v}) \to S_{z,\boldsymbol{v},\infty}$, so the correction term converges, $\Gamma \to \Gamma_\infty$. The function $x \mapsto \max\{0,x\}$ is continuous at every point of $\mathbb{R}$: The function equals x above zero and 0 below, and the two pieces agree at zero. A continuous function preserves the limit of a convergent nonrandom sequence, so $\max\{0,\Gamma\} \to \max\{0,\Gamma_\infty\}$.

The plug-in estimator $\widehat{\theta}$ is consistent for the limiting parameter.

Proposition S5 (Consistency of the plug-in estimator)**.** *Under Assumptions S1 and S5, together with Assumption 3 of the main text, define the finite-population estimand*

$$\theta_N \coloneqq \sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})\mu(1,\boldsymbol{v}) - \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})\mu(0,\boldsymbol{v})$$

— the estimand θ of (S13)*, written here with the subscript N to make explicit its dependence on the population size — and the limiting parameter*

$$\theta_\infty \coloneqq \sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})\mu_{1,\boldsymbol{v},\infty} - \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})\mu_{0,\boldsymbol{v},\infty}.$$

Then

$$\widehat{\theta} \xrightarrow{p} \theta_\infty \quad \text{as } N\to\infty,$$

where $\xrightarrow{p}$ denotes convergence in probability under the randomization distribution induced by complete random assignment, conditional on the potential outcomes.

Proof. The proof proceeds in four steps. Step 1 establishes the consistency of each cell mean for the corresponding population cell mean. Step 2 combines the cell means into the plug-in estimator and establishes the convergence in probability $\widehat{\theta} - \theta_N \xrightarrow{p} 0$. Step 3 establishes the nonrandom convergence $\theta_N \to \theta_\infty$. Step 4 combines the convergence in probability from Step 2 with the nonrandom convergence from Step 3.

Step 1: consistency of the cell means. Fix $(z,\boldsymbol{v}) \in \{0,1\}\times\mathcal{V}$. Under complete random assignment with fixed counts $n_{z,\boldsymbol{v}}$, the set $\{y_i(z,\boldsymbol{v}) : \boldsymbol{T}_i = (z,\boldsymbol{v})\}$ is a simple random sample without replacement of size $n_{z,\boldsymbol{v}}$ from the finite population $\{y_i(z,\boldsymbol{v}) : i = 1,\dots,N\}$, and $\hat{\mu}(z,\boldsymbol{v})$ is the corresponding sample mean. From the expression for the variance of $\hat{\mu}(z,\boldsymbol{v})$ derived in the proof of Proposition S4,

$$\operatorname{Var}\left[\hat{\mu}(z,\boldsymbol{v})\right] = \frac{N-n_{z,\boldsymbol{v}}}{N}\,\frac{S^2(z,\boldsymbol{v})}{n_{z,\boldsymbol{v}}}.$$

By Assumption S5(a), $n_{z,\boldsymbol{v}}/N \to \pi_{z,\boldsymbol{v},\infty} \in (0,1)$, so both $n_{z,\boldsymbol{v}}$ and $N - n_{z,\boldsymbol{v}}$ diverge and $(N-n_{z,\boldsymbol{v}})/N \to 1-\pi_{z,\boldsymbol{v},\infty} \in (0,1)$. By Assumption S5(b), $S^2(z,\boldsymbol{v})$ converges to the finite

limit $S^2_{z,\boldsymbol{v},\infty}$. There is therefore a constant $C_{z,\boldsymbol{v}} < \infty$ and an integer N_0 such that for all $N \geq N_0$,

$$\operatorname{Var}\left[\hat{\mu}(z,\boldsymbol{v})\right] \leq \frac{C_{z,\boldsymbol{v}}}{n_{z,\boldsymbol{v}}}.$$

The right-hand side tends to 0 because $n_{z,\boldsymbol{v}} \to \infty$, and the variance is nonnegative, so $\operatorname{Var}\left[\hat{\mu}(z,\boldsymbol{v})\right] \to 0$. Under complete random assignment, $\hat{\mu}(z,\boldsymbol{v})$ is unbiased for $\mu(z,\boldsymbol{v})$ (Proposition S3), so by Chebyshev's inequality, for any $\eta > 0$,

$$\Pr\left(|\hat{\mu}(z,\boldsymbol{v}) - \mu(z,\boldsymbol{v})| > \eta\right) \leq \frac{\operatorname{Var}\left[\hat{\mu}(z,\boldsymbol{v})\right]}{\eta^2} \to 0,$$

and therefore $\hat{\mu}(z,\boldsymbol{v}) - \mu(z,\boldsymbol{v}) \xrightarrow{p} 0$ for each profile $(z,\boldsymbol{v})$.

Step 2: consistency for the finite-population estimand. Write

$$\widehat{\theta} - \theta_N = \sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})\left\{\hat{\mu}(1,\boldsymbol{v}) - \mu(1,\boldsymbol{v})\right\} - \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})\left\{\hat{\mu}(0,\boldsymbol{v}) - \mu(0,\boldsymbol{v})\right\}.$$

Because $\mathcal{V}$ is finite and the weights $g_z(\boldsymbol{v})$ are fixed by Assumption S5(c), the right-hand side of the display above is a finite linear combination, with fixed coefficients, of terms that converge in probability to 0 by Step 1. A finite linear combination of such terms with fixed coefficients itself converges in probability to 0; hence

$$\widehat{\theta} - \theta_N \xrightarrow{p} 0.$$

Step 3: convergence of the finite-population estimand to its limit. By Assumption S5(b), $\mu(z,\boldsymbol{v}) \to \mu_{z,\boldsymbol{v},\infty}$ for each profile $(z,\boldsymbol{v})$. A finite sum of convergent sequences with fixed coefficients converges to the same sum of the limits, so

$$\theta_N \longrightarrow \sum_{\boldsymbol{v}\in\mathcal{V}} g_1(\boldsymbol{v})\mu_{1,\boldsymbol{v},\infty} - \sum_{\boldsymbol{v}\in\mathcal{V}} g_0(\boldsymbol{v})\mu_{0,\boldsymbol{v},\infty} = \theta_\infty.$$

Step 4: combining the limits. Adding and subtracting θ_N splits the error into a random

part and a nonrandom part,

$$\widehat{\theta} - \theta_\infty = \left(\widehat{\theta} - \theta_N\right) + (\theta_N - \theta_\infty).$$

The first term converges in probability to 0 by Step 2, and the second term is nonrandom and converges to 0 by Step 3. The sum therefore converges in probability to 0, which establishes $\widehat{\theta} \xrightarrow{p} \theta_\infty$. □

The plug-in estimator of the variance bound in (S53) is consistent for the scaled limiting bound $\overline{\sigma}^2_{\theta,\infty}$ in (S55).

Proposition S6 (Consistency of the plug-in estimator of the variance bound)**.** *Under Assumptions S1 and S5, together with Assumption 3 of the main text,*

$$N\,\widehat{\overline{\mathrm{Var}}}\left[\widehat{\theta}\right] \xrightarrow{p} \overline{\sigma}^2_{\theta,\infty}.$$

Proof. The proof establishes the consistency of each cell's sample variance and then assembles the scaled plug-in estimator of the variance bound by the continuous mapping theorem, which states that convergence in probability is preserved by continuous functions: If $X_N \xrightarrow{p} x$ and the function h is continuous at x, then $h(X_N) \xrightarrow{p} h(x)$.

Fix $(z, \boldsymbol{v})$. Under complete random assignment and SUTVA, the set $\{Y_i : \boldsymbol{T}_i = (z, \boldsymbol{v})\} = \{y_i(z, \boldsymbol{v}) : \boldsymbol{T}_i = (z, \boldsymbol{v})\}$ is a simple random sample without replacement of size $n_{z,\boldsymbol{v}}$ from the finite population $\{y_i(z, \boldsymbol{v}) : i = 1, \ldots, N\}$, and $\hat{S}^2(z, \boldsymbol{v})$ is the corresponding sample variance, which is unbiased for $S^2(z, \boldsymbol{v})$ (Cochran, 1977, Theorem 2.4). Two cases establish $\hat{S}^2(z, \boldsymbol{v}) - S^2(z, \boldsymbol{v}) \xrightarrow{p} 0$, according to whether the limiting variance $S^2_{z,\boldsymbol{v},\infty}$ of Assumption S5(b) is zero or positive. Both cases are possible: $S^2(z, \boldsymbol{v})$ measures the spread of the potential outcomes across units rather than a sampling variance, so the limit need not vanish as N grows, and Assumption S5 permits a zero limit for any single cell — condition (d) restricts only the limiting variance of $\widehat{\theta}$ as a whole. The zero case needs its own argument because the ratio consistency used in the positive case divides by $S^2(z, \boldsymbol{v})$.

If $S^2_{z,\boldsymbol{v},\infty} = 0$: Markov's inequality states that, for a nonnegative random variable X and

any $\eta > 0$, $\Pr(X > \eta) \leq \mathrm{E}[X]/\eta$. Applied to the nonnegative random variable $\hat{S}^2(z, \boldsymbol{v})$, the inequality gives

$$\Pr\left(\hat{S}^2(z, \boldsymbol{v}) > \eta\right) \leq \frac{\mathrm{E}\left[\hat{S}^2(z, \boldsymbol{v})\right]}{\eta} = \frac{S^2(z, \boldsymbol{v})}{\eta} \to 0,$$

because $\hat{S}^2(z, \boldsymbol{v})$ is unbiased for $S^2(z, \boldsymbol{v})$ and $S^2(z, \boldsymbol{v}) \to 0$. The display gives $\hat{S}^2(z, \boldsymbol{v}) \xrightarrow{p} 0$, and subtracting the nonrandom sequence $S^2(z, \boldsymbol{v}) \to 0$ leaves $\hat{S}^2(z, \boldsymbol{v}) - S^2(z, \boldsymbol{v}) \xrightarrow{p} 0$.

If $S^2_{z,\boldsymbol{v},\infty} > 0$: Proposition 1 of Li and Ding (2017) gives the consistency of the ratio

$$\hat{S}^2(z, \boldsymbol{v})/S^2(z, \boldsymbol{v}) \xrightarrow{p} 1,$$

provided Condition (4) of their Theorem 1 holds for the cell, namely

$$\frac{1}{\min(n_{z,\boldsymbol{v}}, N - n_{z,\boldsymbol{v}})} \cdot \frac{\max_{1 \leq i \leq N} \left[y_i(z, \boldsymbol{v}) - \mu(z, \boldsymbol{v})\right]^2}{S^2(z, \boldsymbol{v})} \to 0. \tag{S57}$$

To verify the condition (S57), factor its left-hand side into three pieces,

$$\frac{N}{\min(n_{z,\boldsymbol{v}}, N - n_{z,\boldsymbol{v}})} \cdot \frac{\max_{1 \leq i \leq N} \left[y_i(z, \boldsymbol{v}) - \mu(z, \boldsymbol{v})\right]^2}{N} \cdot \frac{1}{S^2(z, \boldsymbol{v})}.$$

The first factor converges to $1/\min(\pi_{z,\boldsymbol{v},\infty}, 1 - \pi_{z,\boldsymbol{v},\infty})$, a finite limit, by Assumption S5(a). The second factor converges to 0 by the Lindeberg-type condition of Assumption S5(b). The third factor converges to $1/S^2_{z,\boldsymbol{v},\infty}$, a finite limit, because $S^2_{z,\boldsymbol{v},\infty} > 0$ in the present case. A product of two factors with finite limits and one factor that converges to 0 converges to 0, so the condition (S57) holds. Since $S^2_{z,\boldsymbol{v},\infty} > 0$, the variance $S^2(z, \boldsymbol{v})$ is positive for all large N, and multiplying and dividing the estimation error by $S^2(z, \boldsymbol{v})$ gives

$$\hat{S}^2(z, \boldsymbol{v}) - S^2(z, \boldsymbol{v}) = S^2(z, \boldsymbol{v}) \left[\frac{\hat{S}^2(z, \boldsymbol{v})}{S^2(z, \boldsymbol{v})} - 1\right].$$

The factor in brackets converges in probability to 0 by the consistency of the ratio, and $S^2(z, \boldsymbol{v})$ is a nonrandom sequence with a finite limit, hence bounded. The product of a

bounded nonrandom sequence and a sequence that converges in probability to 0 converges in probability to 0, so $\hat{S}^2(z,\boldsymbol{v}) - S^2(z,\boldsymbol{v}) \xrightarrow{p} 0$.

In both cases $\hat{S}^2(z,\boldsymbol{v}) - S^2(z,\boldsymbol{v}) \xrightarrow{p} 0$, and $S^2(z,\boldsymbol{v}) \to S^2_{z,\boldsymbol{v},\infty}$ by Assumption S5(b), so $\hat{S}^2(z,\boldsymbol{v}) \xrightarrow{p} S^2_{z,\boldsymbol{v},\infty}$. Because the square-root function is continuous on $[0,\infty)$, the continuous mapping theorem then gives $\hat{S}(z,\boldsymbol{v}) \xrightarrow{p} S_{z,\boldsymbol{v},\infty}$.

Because $\mathcal{V}$ is finite, the vector $\left(\hat{S}^2(z,\boldsymbol{v}), \hat{S}(z,\boldsymbol{v})\right)_{z,\boldsymbol{v}}$ has finitely many coordinates, and convergence in probability coordinate by coordinate implies convergence in probability of the whole vector. The vector therefore converges in probability to $\left(S^2_{z,\boldsymbol{v},\infty}, S_{z,\boldsymbol{v},\infty}\right)_{z,\boldsymbol{v}}$. The scaled plug-in estimator of the variance bound

$$N\,\widehat{\overline{\mathrm{Var}}}\left[\widehat{\theta}\right] = \sum_{z,\boldsymbol{v}} g_z(\boldsymbol{v})^2 \hat{S}^2(z,\boldsymbol{v}) \frac{N}{n_{z,\boldsymbol{v}}} - \max\left\{0,\, \widehat{\Gamma}\right\},$$
$$\widehat{\Gamma} = 2\sum_{z,\boldsymbol{v}} g_z(\boldsymbol{v})^2 \hat{S}^2(z,\boldsymbol{v}) - \left(\sum_{\boldsymbol{v}} g_1(\boldsymbol{v})\hat{S}(1,\boldsymbol{v}) + \sum_{\boldsymbol{v}} g_0(\boldsymbol{v})\hat{S}(0,\boldsymbol{v})\right)^2,$$

is a finite composition of additions, multiplications by the fixed scalars $g_z(\boldsymbol{v})^2$ and the ratios $N/n_{z,\boldsymbol{v}} \to 1/\pi_{z,\boldsymbol{v},\infty}$, squared finite sums, and the function $x \mapsto \max\{0,x\}$, all of which are continuous operations. The ratios $N/n_{z,\boldsymbol{v}}$ are nonrandom, and a nonrandom sequence that converges to a limit also converges in probability to that limit, so the full vector of inputs — the sample variances, their square roots, and the ratios — converges in probability jointly. By the continuous mapping theorem, replacing $\hat{S}^2(z,\boldsymbol{v})$ by $S^2_{z,\boldsymbol{v},\infty}$, $\hat{S}(z,\boldsymbol{v})$ by $S_{z,\boldsymbol{v},\infty}$, and $N/n_{z,\boldsymbol{v}}$ by $1/\pi_{z,\boldsymbol{v},\infty}$ yields

$$N\,\widehat{\overline{\mathrm{Var}}}\left[\widehat{\theta}\right] \xrightarrow{p} \overline{\sigma}^2_{\theta,\infty},$$

which completes the proof. □

Finally, $\widehat{\theta}$ satisfies a finite-population central limit theorem, which together with the consistent plug-in estimator of the variance bound justifies the Normal-approximation confidence intervals reported in the application. The intervals require only the consistency just established. By Slutsky's theorem, standardizing $\widehat{\theta} - \theta_N$ by any scale that converges

in probability to a limit at least as large as $\sigma_{\theta,\infty}$ produces a statistic whose limiting distribution is Normal with variance at most one, so a confidence interval built from such a scale covers θ_N with limiting probability at least the interval's nominal level — coverage that is asymptotically conservative. Conservative coverage delivers valid inference: A test that rejects a hypothesized value of the estimand exactly when the interval excludes the hypothesized value rejects a true value with limiting probability at most the nominal error rate.

Proposition S7 (Finite population CLT for the plug-in estimator)**.** *Under Assumptions S1 and S5, together with Assumption 3 of the main text,*

$$\sqrt{N}\left\{\widehat{\theta} - \theta_N\right\} / \sigma_{\theta,N} \quad \xrightarrow{d} \quad \mathcal{N}(0,1),$$

where $\mathcal{N}(0,1)$ *denotes the Normal distribution with mean* 0 *and variance* 1*, and* $\xrightarrow{d}$ *denotes convergence in distribution under the randomization distribution induced by complete random assignment, conditional on the potential outcomes. The scaled variance* $\sigma^2_{\theta,N} = N \text{ Var}\left[\widehat{\theta}\right]$ *is given by Proposition S4, and* $\sigma^2_{\theta,N} \to \sigma^2_{\theta,\infty} \in (0,\infty)$*.*

Proof. The proof recasts the plug-in estimator of the estimand family, $\widehat{\theta}$, as a scalar contrast of the means of the arms of a multi-arm experiment — the setting of Theorem 5 of Li and Ding (2017), which states a finite-population central limit theorem for a fixed linear contrast of arm means under complete random assignment. Casting $\widehat{\theta}$ in that form lets the theorem apply directly: The cells $(z, \boldsymbol{v})$ become the arms, and the estimand weights become the contrast weights. The proof sets up the correspondence and then verifies the theorem's conditions.

To match the notation of Li and Ding (2017), index the treatment profiles $(z, \boldsymbol{v}) \in \mathcal{T}$ by $q = 1, \ldots, Q$, one arm per cell, and define scalar potential outcomes $Y_i(q) \coloneqq y_i(z, \boldsymbol{v})$. Let $\hat{Y}(q)$ and $\bar{Y}(q)$ denote the sample mean and finite-population mean in arm q. A contrast weight attaches to each arm the coefficient that the arm's mean carries in the estimand: The cells of racial condition 1 enter θ_N with weights $g_1(\boldsymbol{v})$, and the cells of racial condition

0 enter with weights $-g_0(\boldsymbol{v})$. Define the contrast weights accordingly,

$$c_q \coloneqq \begin{cases} g_1(\boldsymbol{v}) & \text{if arm } q \text{ indexes the profile } (1, \boldsymbol{v}), \\ -g_0(\boldsymbol{v}) & \text{if arm } q \text{ indexes the profile } (0, \boldsymbol{v}). \end{cases}$$

Substituting the arm notation into the definitions of $\widehat{\theta}$ in (S16) and of θ_N collects each into a single weighted sum over arms,

$$\widehat{\theta} = \sum_{q=1}^{Q} c_q \hat{Y}(q), \qquad \theta_N = \sum_{q=1}^{Q} c_q \bar{Y}(q),$$

which is the scalar contrast of Li and Ding (2017, Section 2) with a $1 \times Q$ contrast matrix whose entries are the c_q. The index q and the potential-outcome notation $Y_i(q)$ are local to this proof and follow Li and Ding (2017). The index q is unrelated to the joint PMF $q(z, \boldsymbol{v})$ and the conditionals q_z of the main text, and $Y_i(q)$ denotes the potential outcome $y_i(z, \boldsymbol{v})$ rather than the observed outcome Y_i.

Assumption S5(a)–(c) ensure that the conditions of Li and Ding (2017, Theorem 5) hold. As $N \to \infty$, the arm fractions n_q/N converge to positive limits by (a). The finite-population variances and covariances converge to finite limits as $N \to \infty$ by (b), and the Lindeberg-type condition on the maximum squared deviation also holds by (b). The contrast weights c_q do not vary with N by (c); the theorem takes the contrast matrix as fixed along the sequence, so weights that changed with N would change the estimand from one population to the next and fall outside the theorem. Assumption S5(d) is the nondegeneracy condition ensuring that the limiting variance $\sigma^2_{\theta,\infty} = \lim_{N\to\infty} N \operatorname{Var}[\widehat{\theta}]$ is strictly positive. With the conditions verified, Theorem 5 of Li and Ding (2017), applied to the scalar contrast $\widehat{\theta}$, gives

$$\sqrt{N}\left\{\widehat{\theta} - \theta_N\right\} \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\theta,\infty}\right).$$

The standardized display of the proposition follows by Slutsky's theorem. Write the stan-

dardized statistic as a product,

$$\frac{\sqrt{N}\left\{\widehat{\theta}-\theta_N\right\}}{\sigma_{\theta,N}}=\frac{\sqrt{N}\left\{\widehat{\theta}-\theta_N\right\}}{\sigma_{\theta,\infty}}\cdot\frac{\sigma_{\theta,\infty}}{\sigma_{\theta,N}}. \tag{S58}$$

The first factor on the right-hand side of (S58) converges in distribution to $\mathcal{N}(0,1)$ because dividing the limit $\mathcal{N}(0,\sigma^2_{\theta,\infty})$ by the constant $\sigma_{\theta,\infty}>0$ rescales the limit to unit variance. The second factor is a nonrandom sequence that converges to 1: $\sigma^2_{\theta,N}\to\sigma^2_{\theta,\infty}\in(0,\infty)$ by Assumption S5(d), and the square root and the reciprocal are continuous at the strictly positive limit. Applied to the two factors of (S58) — one converging in distribution and one converging to the constant 1 — Slutsky's theorem gives

$$\sqrt{N}\left\{\widehat{\theta}-\theta_N\right\}\,/\,\sigma_{\theta,N}\quad\xrightarrow{d}\quad\mathcal{N}(0,1),$$

which is the display of the proposition. □

S1.9 Confidence intervals for the NREC ratio

Section 7 of the main text estimates the NREC by the instrumental-variables ratio $\widehat{\mathrm{NREC}}=\widehat{\Delta}_Y/\widehat{\Delta}_Z$, the ratio of the reduced-form estimator (10) to the first-stage estimator (11) of the main text, with the cue as the instrument and perceived race as the treatment taken up; Section 6.2 of the main text develops how this instrumental-variables reading relates to the classical setting. A confidence interval for the ratio uses a delta-method approximation, justified for the finite population by Pashley (2022) together with a finite-population central limit theorem (Li and Ding, 2017). The interval is design-based, large-sample, and approximate rather than exact.

The justification runs along a sequence of finite populations of increasing size, indexed by N as in Assumption S5, with the cue in place of the configuration as the randomized assignment: n_w units receive cue value $w\in\{0,1\}$ under Assumption 4 of the main text, and the arm shares n_w/N converge to limits in $(0,1)$, the analogue of Assumption S5(a). Beyond

randomization of the cue, the delta-method approximation rests on three conditions, which specialize to the ratio the requirements of the finite-population delta method (Pashley, 2022, Theorem 2) and of the central limit theorem that the delta method draws on (Li and Ding, 2017, Theorem 5):

- The complier share converges to a positive limit, $|\mathcal{C}|/N \to \pi_{\mathcal{C}} > 0$. Under Assumption S3, the population first stage equals the complier share, $\Delta_Z = |\mathcal{C}|/N$, as in the proof of Corollary S1, so the positive limit keeps the denominator of the ratio away from zero along the sequence. The positive-limit condition strengthens Assumption S4, which requires only one complier at each population size N and so permits a complier share that vanishes as N grows.
- Under each value w of the cue, the finite-population means, variances, and covariances of the response $y_i(w)$ and of perceived race $z_i(w)$ converge to finite limits, and the response satisfies $\max_{1\le i\le N}\left[y_i(w) - \frac{1}{N}\sum_{j=1}^{N} y_j(w)\right]^2/N \to 0$. The two requirements are the analogues of Assumption S5(b) for the cue design. The maximum-squared-deviation condition needs no separate statement for perceived race: Because $z_i(w) \in \{0,1\}$, every squared deviation of $z_i(w)$ from its mean is at most 1, so the maximum divided by N is at most $1/N$, which tends to 0.
- The variance of the linearized statistic $\widehat{\Delta}_Y - \text{NREC}\cdot\widehat{\Delta}_Z$, the numerator of (S59) below, scaled by N, converges to a strictly positive limit, the analogue of Assumption S5(d).

The delta method reaches the ratio from a linear statistic. Subtracting the NREC from the ratio and placing the difference over the common denominator $\widehat{\Delta}_Z$ gives the exact identity

$$\widehat{\text{NREC}} - \text{NREC} = \frac{\widehat{\Delta}_Y - \text{NREC}\cdot\widehat{\Delta}_Z}{\widehat{\Delta}_Z}. \tag{S59}$$

The numerator is the *linearized statistic*: The numerator is linear in the pair $(\widehat{\Delta}_Y, \widehat{\Delta}_Z)$, and its mean over the randomization distribution of the cue is zero, because $\text{E}[\widehat{\Delta}_Y] = \Delta_Y$ and $\text{E}[\widehat{\Delta}_Z] = \Delta_Z$ — the two-arm analogue of Proposition S3 — and $\Delta_Y - \text{NREC}\cdot\Delta_Z = 0$

by Proposition 2. Two steps then deliver the Normal approximation for the ratio. First, the pair $(\widehat{\Delta}_Y, \widehat{\Delta}_Z)$ satisfies a joint finite-population central limit theorem (Li and Ding, 2017, Theorem 5). Second, the finite-population delta method (Pashley, 2022, Theorem 2) replaces the random denominator $\widehat{\Delta}_Z$ in (S59) by the nonrandom Δ_Z; the replacement is asymptotically negligible because $\widehat{\Delta}_Z$ converges in probability to the positive limit $\pi_{\mathcal{C}}$.

The variance of the linearized statistic has the two-arm Neyman form. For each unit i and cue value w, define the *composite outcome* $a_i(w) \coloneqq y_i(w) - \text{NREC} \cdot z_i(w)$: the response minus NREC times perceived race. Under the cue-level SUTVA (Assumption S2), the observed combination $Y_i - \text{NREC} \cdot Z_i$ equals $a_i(W_i)$. The linearized statistic is therefore the difference of the two cue arms' means of the observed composite outcomes, and the identity (S45), applied with one cell per arm and the composite outcome in place of the potential outcome, gives

$$\text{Var}\left[\widehat{\Delta}_Y - \text{NREC} \cdot \widehat{\Delta}_Z\right] = \frac{S_a^2(1)}{n_1} + \frac{S_a^2(0)}{n_0} - \frac{S^2(\tau_a)}{N}, \tag{S60}$$

where the variance Var is taken over the randomization distribution of the cue assignment, $S_a^2(w)$ is the finite-population variance of the composite outcomes $a_i(w)$ under cue value w, and $S^2(\tau_a)$ is the finite-population variance of the unit-level differences $\tau_{a,i} \coloneqq a_i(1) - a_i(0)$. As in Corollary S2, the subtracted term is unidentified — each unit reveals its composite outcome under one cue value only — and $S^2(\tau_a) \geq 0$, so dropping the subtracted term bounds the variance from above:

$$\text{Var}\left[\widehat{\Delta}_Y - \text{NREC} \cdot \widehat{\Delta}_Z\right] \leq \frac{S_a^2(1)}{n_1} + \frac{S_a^2(0)}{n_0}. \tag{S61}$$

The bound in (S61) follows Pashley (2022, Section 3.1), whose conservative variance keeps the identified within-arm variances and drops the subtracted term, and corresponds to the branch of Corollary S2 that discards the variance of the unit-level contrasts, (S46), with the composite outcome in place of the potential outcome.

The plug-in estimator of the bound replaces each population piece by its sample coun-

terpart, with the unknown NREC in the composite outcome replaced by the ratio $\widehat{\text{NREC}}$:

$$\widehat{V} \coloneqq \frac{\hat{S}^2_{\hat{a}}(1)}{n_1} + \frac{\hat{S}^2_{\hat{a}}(0)}{n_0}, \qquad \hat{S}^2_{\hat{a}}(w) \coloneqq \frac{1}{n_w - 1} \sum_{i\,:\,W_i = w} \left(\hat{A}_i - \frac{1}{n_w} \sum_{j\,:\,W_j = w} \hat{A}_j \right)^2, \tag{S62}$$

where $\hat{A}_i \coloneqq Y_i - \widehat{\text{NREC}} \cdot Z_i$ is the estimated composite outcome for unit i. Under the three conditions, $N\widehat{V}$ is consistent for the limit of N times the bound in (S61). Expanding the square shows that $\hat{S}^2_{\hat{a}}(w) = \hat{S}^2_Y(w) - 2\,\widehat{\text{NREC}}\,\hat{S}_{YZ}(w) + \widehat{\text{NREC}}^{\,2}\,\hat{S}^2_Z(w)$, where $\hat{S}^2_Y(w)$, $\hat{S}^2_Z(w)$, and $\hat{S}_{YZ}(w)$ are the within-arm sample variances and covariance of the observed outcome and perceived race. Each within-arm sample variance is consistent for its limit by the argument of Proposition S6, and the sample covariance follows from the same argument applied to the sum of the outcome and perceived race; $\widehat{\text{NREC}}$ is consistent for the NREC because its numerator and denominator converge in probability and the limiting denominator $\pi_{\mathcal{C}}$ is positive; and the expansion is a continuous function of the sample moments and the ratio. The reported confidence interval is

$$\widehat{\text{NREC}} \;\pm\; \Phi^{-1}(1 - \alpha/2)\, \frac{\sqrt{\widehat{V}}}{|\widehat{\Delta}_Z|},$$

where $\Phi^{-1}(1 - \alpha/2)$ is the standard Normal quantile at level $1 - \alpha/2$; the division by $|\widehat{\Delta}_Z|$ restores the denominator of the linearization in (S59), with $\widehat{\Delta}_Z$ in place of the unknown Δ_Z.

Divided by Δ_Z^2, the variance bound in (S61) is the finite-population analogue of the usual asymptotic variance of the instrumental-variables estimator (Angrist and Pischke, 2008, Section 4.2.1, pp. 138–140) — itself the variance of a linearization, with the first stage in its denominator. The interval therefore degrades when the cue moves perceived race only weakly (Angrist and Pischke, 2008, Section 4.6.4, pp. 205–216): A small complier share leaves $\widehat{\Delta}_Z$ small relative to its own sampling variability, and asymptotic standard errors and confidence intervals can then suggest that an unstable estimate is stable (Imbens and Rosenbaum, 2005, Section 1.4). Approaches built for weak instruments include the

permutation intervals of Imbens and Rosenbaum (2005) and the methods of Kang et al. (2018) and Aronow et al. (2026).

S2 Additional Material

This section collects the summary table of grounds from Section 5 of the main text (Section S2.1).

S2.1 Summary of grounds for choosing among members of the estimand family

Table S2 collects the grounds developed in Section 5 of the main text, ordered by which component of the estimand each ground bears on.

Ground	**Implication for the estimand**
Panel A. The contrast: which nonracial features are held fixed	
Classical conception of the category	Hold every nonracial feature fixed: the all-else-equal effect
Thick constructivist conception	Let the indexed features vary by race: the within-race effect
Isolating a causal mechanism	Provide the information the decision-maker could otherwise infer and hold it fixed: a mixed effect
Normative and legal commitments	Hold the allowable features fixed; let the non-allowable vary: a mixed effect
Panel B. The weighting: how the varying features are weighted	
External validity	Calibrate ρ or the q_z to the target population
Typicality	Weight each conditional distribution q_z by closeness to the group's prototype

Note: Panel A grounds select the contrast and, with it, the member type; Panel B grounds set the distributions the chosen contrast uses. Every member also fixes a subset of units, typically all N or, under the perception-based regime, the compliers $\mathcal{C}$ (Section 4).

Table S2: Grounds for choosing among the members of the estimand family

Disclosures

Funding. This research received no external funding.

Conflicts of interest. The author declares no conflicts of interest.

Use of artificial intelligence. The author used Claude (Anthropic; Claude Opus 4 models, via Claude Code, `https://claude.com/claude-code`) during 2026 as an editing and verification assistant — revising prose drafted by the author and checking notation, replication code, and figure code. All arguments, results, and interpretations are the author's own.

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables". In: *Journal of the American Statistical Association* 91.434, pp. 444–455.

Angrist, Joshua D. and Jörn-Steffen Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Aronow, Peter M., Haoge Chang, and Patrick Lopatto (2026). "Randomization-based Confidence Sets for the Local Average Treatment Effect". In: *Biometrika* 113.2, asag010.

Aronow, Peter M., Donald P. Green, and Donald K. K. Lee (2014). "Sharp Bounds on the Variance in Randomized Experiments". In: *The Annals of Statistics* 42.3, pp. 850–871.

Cochran, William G. (1977). *Sampling Techniques*. 3rd. Hoboken, NJ: John Wiley & Sons.

de la Cuesta, Brandon, Naoki Egami, and Kosuke Imai (2022). "Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution". In: *Political Analysis* 30.1, pp. 19–45.

Elder, Elizabeth Mitchell and Matthew Hayes (2023). "Signaling Race, Ethnicity, and Gender with Names: Challenges and Recommendations". In: *The Journal of Politics* 85.2, pp. 764–770.

Gaddis, S. Michael (2017). "How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies". In: *Sociological Science* 4, pp. 469–489.

Gärdenfors, Peter (2000). *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: The MIT Press.

— (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. Cambridge, MA: The MIT Press.

Gerber, Alan S. and Donald P. Green (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York, NY: W.W. Norton.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto (2014). "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments". In: *Political Analysis* 22.1, pp. 1–30.

Holland, Paul W. (1986). "Statistics and Causal Inference". In: *Journal of the American Statistical Association* 81.396, pp. 945–960.

Imbens, Guido W. and Joshua D. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects". In: *Econometrica* 62.2, pp. 467–475.

Imbens, Guido W. and Paul R. Rosenbaum (2005). "Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168.1, pp. 109–126.

Imbens, Guido W. and Donald B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY: Cambridge University Press.

Kang, Hyunseung, Laura Peck, and Luke Keele (2018). "Inference for Instrumental Variables: A Randomization Inference Approach". In: *Journal of the Royal Statistical Society. Series A: Statistics in Society* 181.4, pp. 1231–1254.

Kaufman, Aaron R., Jacob M. Grumbach, and Christopher Celaya (2026). "Improving Compliance in Experimental Studies of Discrimination". In: *The Journal of Politics* 88.1.

Landgrave, Michelangelo and Nicholas Weller (2022). "Do Name-Based Treatments Violate Information Equivalence? Evidence from a Correspondence Audit Experiment". In: *Political Analysis* 30.1, pp. 142–148.

Leavitt, Thomas and Viviana Rivera-Burgos (2024). "Audit Experiments of Racial Discrimination and the Importance of Symmetry in Exposure to Cues". In: *Political Analysis* 32.4, pp. 445–462.

— (2026). "Navigating the Mismeasurement of Intermediary Variables in Message-Based Experiments". In: *Political Science Research and Methods*. First View, pp. 1–14.

Li, Xinran and Peng Ding (2017). "General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference". In: *Journal of the American Statistical Association* 112.520, pp. 1759–1769.

Neyman, Jersey (1923). "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes". In: *Roczniki Nauk Rolniczych* 10, pp. 1–51.

Nosofsky, Robert Mark (1986). "Attention, Similarity, and the Identification-Categorization Relationship". In: *Journal of Experimental Psychology: General* 115.1, pp. 39–61.

Pashley, Nicole E. (2022). "Note on the Delta Method for Finite Population Inference with Applications to Causal Inference". In: *Statistics & Probability Letters* 188, p. 109540.

Shepard, Roger N. (1987). "Toward a Universal Law of Generalization for Psychological Science". In: *Science* 237.4820, pp. 1317–1323.

Zárate, Marques G., Enrique Quezada-Llanes, and Angel D. Armenta (2024). "Se Habla Español: Spanish-Language Appeals and Candidate Evaluations in the United States". In: *The American Political Science Review* 118.1, pp. 363–379.