Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Race-Neutral by Construction: A Transparent, Gaming-Resistant Mechanism for Disadvantage-Based College Admissions in the United States

Sanad Aburass

saburass@luther.edu

Luther College

Research Article

Keywords: algorithmic fairness, college admissions, affirmative action, transparency, mechanism design, strategyproofness, educational equity

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11027869/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Race-Neutral by Construction: A Transparent, Gaming-Resistant Mechanism for Disadvantage-Based College Admissions in the United States

Abstract

In *Students for Fair Admissions v. Harvard* (2023) the U.S. Supreme Court ended race-conscious college admissions, intensifying interest in race-neutral alternatives that select for disadvantage. The leading alternative — class-based admissions — is well argued in the policy literature but has repeatedly foundered on *operationalisation*: the College Board's opaque "adversity score" collapsed a person into a single number and was withdrawn, while naive additive indices are trivially gameable by advantaged applicants. We propose a transparent, closed-form admissions mechanism that (i) organises disadvantage factors by *verifiability* into three tiers, (ii) replaces raw achievement with a cross-fitted *context-adjusted residual* ("distance travelled") that is ability-informative but environment-neutral, (iii) combines these through a bounded, published index, and (iv) resists manipulation by anchoring the score on exogenous and verified data with a per-tier cap on self-attested factors. We prove that the manipulation gain any applicant can extract is bounded by a single published constant, and confirm the bound empirically. On a synthetic population the mechanism admits a class of *higher* latent ability than test-score ranking while roughly quadrupling the share of admits from disadvantaged (bottom-half) environments, and cuts the displacement of genuinely disadvantaged applicants by an adversary by ~85% relative to a naive index. On real data from the Stanford Education Data Archive (about 12,400 districts) we show that socioeconomic context explains 74% of achievement variance (grounding the approach), that the residual is environment-neutral (correlation with socioeconomic status falls from 0.77 to 0.02), and — decomposing the race–achievement gap — that context captures 80–89% of it while a real remainder persists, quantifying both the promise and the honest limit of class-based admissions. We discuss governance, contestability, and the ethical boundaries of adversity measurement.

Keywords: algorithmic fairness · college admissions · affirmative action · transparency · mechanism design · strategyproofness · educational equity

Introduction

On 29 June 2023 the United States Supreme Court held, in *Students for Fair Admissions, Inc. v. President and Fellows of Harvard College* and the companion case against the University of North Carolina, that the race-conscious admissions programmes of those institutions violate the Equal Protection Clause of the Fourteenth Amendment (Students for Fair Admissions, Inc. v. President and Fellows of Harvard College 2023). The decision overturned forty-five years of precedent reaching back to *Bakke* (Regents of the University of California v. Bakke 1978) and *Grutter* (Grutter v. Bollinger 2003), and for the first time barred private as well as public institutions nationwide from considering an applicant's race. Early enrolment figures reported sharp declines in the share of Black students at several selective institutions in the first post-decision cohorts. Whatever one's view of the ruling, its practical effect is settled: institutions that wish to enrol a socioeconomically and demographically broad class must now do so through *race-neutral* means.

The most developed race-neutral alternative is *class-based* (or *socioeconomic*, or *adversity-based*) affirmative action: rather than weighting an applicant's race, an admissions process gives weight to the disadvantage in the environment the applicant grew up in — low-income neighbourhoods, under-resourced schools, family disruption, and the absence of the enrichment opportunities that advantaged applicants take for granted. The intellectual case has been made for three decades, most prominently by Kahlenberg (Kahlenberg 1996, 2025), and simulation studies suggest class-based preferences can recover much of the socioeconomic and some of the racial diversity that race-conscious admissions produced (Carnevale and Rose 2004). Because the factors are applied to every applicant regardless of race, the approach is, by construction, non-racial: it advantages a poor rural white applicant and a poor urban Black applicant on the same terms, even where the population meeting the criteria is disproportionately of one group.

The idea is sound; the problem has always been *operationalisation*. Two failures are instructive. First, the College Board's "adversity score" (later rebranded *Landscape*) attempted to summarise a student's neighbourhood and school context in a single 1–100 number. It was withdrawn before full rollout amid criticism that a single composite is opaque, that it reduced a person's circumstances to a figure no admissions officer could interpret or contest, and that it tied a student's fate to their ZIP code while missing the family-level adversity that geography cannot see. Second, the obvious fix — a simple additive index that sums up disadvantage indicators — is trivially *gameable*: an advantaged family can relocate to a poorer tract for the senior year, or an applicant can assert hard-to-verify hardships, and collect the associated boost. Neither failure is a reason to abandon disadvantage-based admissions; each is a reason to take *measurement* seriously.

This paper takes measurement as the central problem and treats the admissions rule as an object of mechanism design. We ask: can a disadvantage-based admissions score be **transparent** (a published, closed-form function rather than a learned black box), **verifiable** (auditable and contestable by the applicant), and **manipulation-robust** (resistant to strategic misreporting), while remaining **race-neutral by construction** and preserving academic quality? Our contributions are:

1. **A verifiability-tiered factor taxonomy.** We organise disadvantage factors not by topic (neighbourhood, school, family) but by how they can be *verified* — exogenous administrative data, individually verified records, and self-attested claims — because verifiability, not topic, determines gameability (Section 3.2).
2. **A context-adjusted achievement residual.** We replace raw achievement with a cross-fitted residual, "distance travelled," that formalises what holistic review gestures at: how far an applicant has come given their environment. We show it is ability-informative yet environment-neutral (Sections 3.3, 4.2, 5.3).
3. **A bounded, published composite with a provable manipulation bound.** The mechanism is a closed-form weighted sum with per-tier caps and a neighbourhood-corroboration gate. We prove the score gain any applicant can extract through misreporting is bounded by a single published constant (Section 3.6).
4. **A two-part evaluation.** A synthetic testbed shows the mechanism behaves as designed — higher latent ability *and* more diversity, with manipulation and

displacement sharply reduced (Section 4). Real data from the Stanford Education Data Archive grounds the premise (context explains 74% of achievement), confirms the residual's environment-neutrality on about 12,400 districts, and decomposes the race–achievement gap to quantify honestly what class-based selection does and does not recover (Section 5).

We stress at the outset what the mechanism is *not*. It is not a claim that disadvantage-based admissions fully reproduce the outcomes of race-conscious admissions; our own real-data analysis shows a racial achievement gap that socioeconomic context does not fully explain. It is not a covert racial proxy; race is not an input and the design is auditable precisely so that this can be checked. And it does not address the binding real-world constraint of cost: admitting disadvantaged students without financial aid does not enrol them. We return to each limit in Sections 6 and 7.

Background and related work

The legal trajectory. Race-conscious admissions were permitted, under narrowing constraints, from *Bakke* (Regents of the University of California v. Bakke 1978) through *Grutter* (Grutter v. Bollinger 2003) and *Fisher* (Fisher v. University of Texas at Austin 2016), until *SFFA* (Students for Fair Admissions, Inc. v. President and Fellows of Harvard College 2023) ended them. Throughout, the Court treated race-neutral alternatives as relevant to whether race-conscious means were "narrowly tailored"; after *SFFA*, race-neutral means are not merely an alternative but the available path.

Class-based affirmative action. Kahlenberg's long argument (Kahlenberg 1996, 2025) is that selective institutions achieved racial diversity largely among affluent students and did little for socioeconomic diversity; class-based preferences, he contends, would produce more of the latter while recovering much of the former. Simulation evidence is mixed but consistent in direction: class-based substitution increases socioeconomic diversity substantially and recovers a majority, but not all, of the racial diversity of race-conscious admissions (Carnevale and Rose 2004). A persistent finding — which our Section 5.5 revisits on real data — is that a Black–White achievement gap remains even among students of similar socioeconomic status (Reardon 2011).

Percent plans. Texas-style "top X%" plans guarantee admission to students in the top decile of each high school, exploiting school segregation to produce diversity without explicit racial classification. They improve access but are blunt: they ignore individual circumstance and depend on the very segregation they route around (Long 2004).

Contextual scoring and its failures. The College Board's adversity score/*Landscape* is the most prominent contextual tool; its retreat from a single composite to a transparent two-number report is the cautionary tale that motivates our transparency requirement. Espenshade and Radford (Espenshade and Radford 2009) document the scale of admissions preferences and their socioeconomic skew; Gaertner and Hart (Gaertner and Hart 2013) report on a class-based pilot in Colorado.

Algorithmic fairness and mechanism design. Our framing draws on two adjacent literatures. From algorithmic fairness we take the concern with proxy discrimination, transparency, and contestability (Dwork et al. 2012; Hardt, Price, and Srebro 2016). From mechanism design we take the notion of strategyproofness — designing an allocation rule so that misreporting one's type yields little advantage — and adapt it to an admissions index by bounding the manipulation gain. Unlike much of the fairness literature, which studies learned classifiers, we deliberately restrict the mechanism to a *closed-form* function: the aim is an artefact an institution can publish, an applicant can contest, and an auditor can reproduce.

The gap. Existing work argues *for* class-based admissions or *evaluates* particular schemes; it does not provide a transparent, verifiable admissions mechanism with an explicit anti-gaming design and a stated robustness guarantee. That artefact is this paper's contribution.

The mechanism

Design principles

Four principles constrain the design, each a response to a documented failure.

- **Transparency.** The mechanism is a published closed-form function with stated weights, not a model learned from historical admissions or outcomes. A learned model inherits the biases in its training data and reproduces the opacity that ended *Landscape*; a closed form can be published, audited, and contested.
- **Verifiability.** Factors are admitted to the index according to how they can be checked, and the design's resistance to gaming rests on this rather than on secrecy.
- **Boundedness.** No single factor, and in particular no self-attested factor, can dominate the score; contributions are capped per tier.
- **Race-neutrality by construction.** Race is not an input. Any recovery of racial diversity is *incidental*, arising only through the correlation between race and disadvantage, and is therefore auditable.

A factor taxonomy organised by verifiability

Prior systems grouped factors by topic. We instead group them by verifiability, because verifiability determines gameability, and gameability is the crux (Table 1).

Table 1: Disadvantage factors organized by verifiability. The anti-gaming design follows directly: anchor the score on Tiers A and B, cap Tier C.

Tier	Description	Example factors	Gaming exposure
A	Exogenous administrative data, not self-reported	School Title I status, share on free/reduced-price lunch, per-pupil funding, AP/IB availability, neighbourhood poverty, unemployment, rural locale (Census/NCES)	Very low at the individual level; suffers ecological fallacy

Tier	Description	Example factors	Gaming exposure
B	Individually verified against existing records	Pell/financial-aid eligibility, first-generation status, foster/ward-of-court status, household size	Low: individual *and* verified
C	Self-attested, hard to verify, privacy-sensitive	Single-parent household, teen orphanhood, incarcerated parent/relative, caregiving burden, extracurricular "desert"	High: fabricable

Tier C contains exactly the factors that reach adversity nothing else captures — and exactly the factors an adversary can assert. The resolution is not to exclude Tier C (that would discard real hardship) but to *bound* it.

Context-adjusted achievement: "distance travelled"

Let y_i be applicant i's standardised observed achievement (a GPA/test composite) and C_i^A the vector of exogenous Tier-A context features. We fit an expected-achievement function using only exogenous context and, crucially, *cross-fit* it so that no applicant's expectation is fit to their own record: partitioning applicants into K folds, for each fold we fit $\hat{f}$ on the other folds and predict out-of-fold. The context-adjusted residual is

$$R_i = \frac{y_i - \hat{f}(C_i^A)}{\sigma_R}, \qquad \sigma_R = \mathrm{sd}\left(y - \hat{f}(C^A)\right) \quad (1)$$

R_i measures how far applicant i's achievement exceeds what their environment predicts — "distance travelled." Because $\hat{f}$ depends only on exogenous context, R is race-neutral by construction and, as we show empirically, retains the ability signal in y while shedding its environmental component.

A bounded adversity index

Each factor k is mapped to a contribution $\varphi_k \in [0,1]$ (a percentile for continuous factors, the indicator itself for binary factors), oriented so that higher means more disadvantaged. Writing $\bar{\varphi}_T(i)$ for the mean contribution over the factors in tier $T \in \{A, B, C\}$, the tier terms are capped:

$$A_i^T = \min(\,\mathrm{cap}_T,\ w_T\,\bar{\varphi}_T(i)\,), \qquad T \in \{A, B, C\} \quad (2)$$

The neighbourhood/school signal is *corroborated* by verified income, so that a high-income applicant in (or claiming) a poor tract gains little:

$$A_i^A \leftarrow A_i^A \cdot g(i), \qquad g(i) = \mathrm{clip}\big(\text{income-disadvantage}_i/\tau\,,\ 0,\ 1\big) \quad (3)$$

with τ the corroboration threshold (we use the median). The adversity index is the sum $A_i = A_i^A + A_i^B + A_i^C$, with the smallest cap placed on the self-attested tier, $\mathrm{cap}_C < \mathrm{cap}_A, \mathrm{cap}_B$.

The composite score

The admission signal is a published weighted sum,

$$S_i \;=\; \alpha\, z(y_i) \;+\; \beta\, R_i \;+\; \gamma\, A_i \quad (4)$$

where $z(\cdot)$ is the standardisation of observed achievement. Retaining $z(y_i)$ alongside the residual R_i anchors the score in measured attainment and prevents the residual term from rewarding advantaged under-performers — a known failure mode of "distance travelled" schemes that we neutralise explicitly. Applicants are admitted top-K by S. In our experiments $\alpha = 0.55,\ \beta = 0.65,\ \gamma = 0.90$; $w_A = 1.0,\ w_B = 1.3,\ w_C = 0.6$; $\text{cap}_A = 0.45,\ \text{cap}_B = 0.60,\ \text{cap}_C = 0.25$. Weights are published, not hidden; Section 3.7 discusses principled defaults and Section 4.5 shows results are robust to them.

Formal properties

The mechanism satisfies, by construction: **monotonicity** (increasing any φ_k never lowers S); **boundedness** ($A_i \le \text{cap}_A + \text{cap}_B + \text{cap}_C$, and no single tier can dominate); **transparency** (closed form, published weights); and **race-neutrality** (race is not an argument of S). The property that matters most for legitimacy is a bound on strategic manipulation.

Let M_i be applicant i's feasible manipulation set: the reported features they can change. Tier B is verified against records and cannot be fabricated; in the robust configuration the exogenous context used by both $\hat{f}$ and A^A is the applicant's verified, duration-weighted context, so late residential or school claims do not move it, and the corroboration gate $g(i)$ nullifies neighbourhood gains for high-income applicants. The only quantity an adversary can raise is the self-attested tier term A_i^C.

> **Proposition 1 (Manipulation bound).** Under verified anchoring of Tiers A and B, the score gain from any manipulation is bounded by the published constant $\gamma \cdot \text{cap}_C$:
>
> $$\max_{m \in M_i} S_i(m) - S_i \;\le\; \gamma\big(\text{cap}_C - A_i^C\big) \;\le\; \gamma\, \text{cap}_C \quad (5)$$

Proof. Under verified anchoring, $z(y_i)$, R_i, A_i^A and A_i^B are invariant to $m \in M_i$. Hence $S_i(m) - S_i = \gamma\,\big(A_i^C(m) - A_i^C\big)$. Since $A_i^C(m) = \min\big(\text{cap}_C, w_C \bar{\varphi}_C(m)\big) \le \text{cap}_C$ and $A_i^C \ge 0$, the difference is at most $\gamma\big(\text{cap}_C - A_i^C\big) \le \gamma\, \text{cap}_C$.

With our constants the bound is $0.90 \times 0.25 = 0.225$; Section 4.4 finds the 95th-percentile manipulation gain under an adversarial budget to be 0.225, exactly saturating the theoretical bound and confirming that verified anchoring — not obscurity — is what limits gaming.

Setting the weights

The one genuinely discretionary element is the weight vector. We recommend *outcome-anchored* weighting: set each w_k in proportion to the achievement deficit associated with

factor k net of ability, so that the boost restores expected loss rather than reflecting a panel's taste, and report a sensitivity analysis over weight perturbations (Section 4.5). Because the weights are published, the normative choice is exposed to scrutiny rather than buried in a model.

Evaluation I: simulation

Synthetic population

We generate $N = 20{,}000$ applicants and admit $K = 2{,}000$ (10%). Each applicant has a latent ability θ drawn *equally* across a majority and a minority group (share 27.8%); a latent environment E that is *lower* for the minority group, encoding the real correlation between group and disadvantage; observable Tier-A/B/C factors generated from E with independent noise; and observed achievement $y = \lambda_\theta \theta + \lambda_E E + \varepsilon$, so that advantaged environments inflate y above ability. We state plainly that this generative model *encodes* the premise that environment suppresses observed achievement; the simulation therefore tests whether the mechanism behaves as designed under that premise, and Section 5 tests the premise itself on real data.

E1 — the residual recovers ability and sheds environment

Raw achievement correlates 0.73 with latent ability θ but also 0.55 with environment E: it rewards advantage. The context-adjusted residual correlates 0.85 with θ — *higher* — while its correlation with E falls to 0.11 (Fig. 1). The residual is thus a better, environment-neutral ability signal, and this holds across every disadvantage tercile.

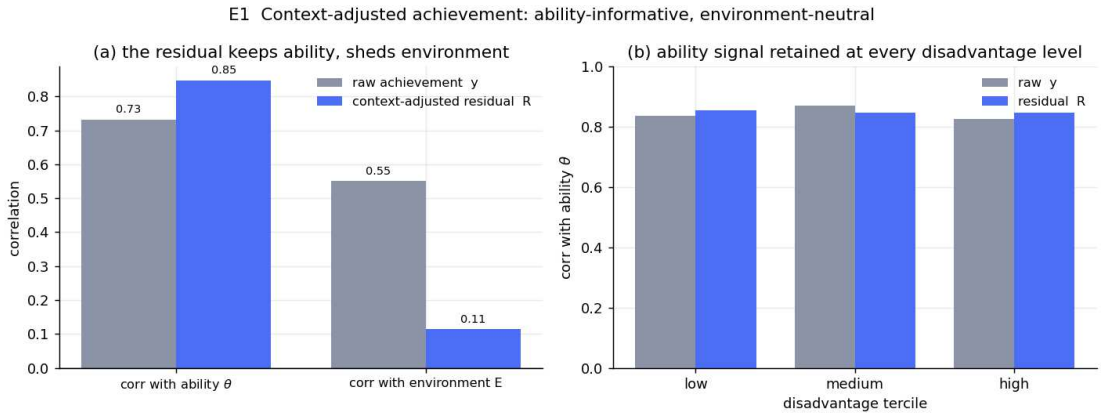


Fig. 1 The context-adjusted residual retains the ability signal while shedding the environment signal, across the disadvantage distribution.

E2 — composition: quality and diversity together

Table 2 and Fig. 2 report the admitted class under five rules. The mechanism admits the *highest-ability* class ($\theta = 1.47$), above test-score ranking (1.29), because the residual recovers able applicants whom raw scores, depressed by environment, rank too low.

Simultaneously it lifts the share from bottom-half environments from 13% to 49% and disadvantaged-tercile representation from 6% to 31%, and raises incidental minority share from 14% to 26% — about 58% of the race-conscious oracle's 45% (a ratio of 26 to 45), without using race. The oracle, by contrast, maximises minority share but has the lowest ability and weak socioeconomic diversity: it boosts race irrespective of disadvantage or attainment.

Table 2 Admitted-class composition by admission rule ($N = 20{,}000$, $K = 2{,}000$). The mechanism improves latent quality and diversity; the race oracle is an illegal reference point for the racial-diversity ceiling.

Rule	Mean θ	Bottom-half env.	Minority (incid.)	Disadv. tercile	Pred. success
Test-score	1.29	13%	14%	6%	0.80
Income-only	1.41	25%	18%	14%	0.82
Top-X% plan	1.41	43%	26%	30%	0.81
Mechanism	**1.47**	**49%**	**26%**	**31%**	**0.82**
Race oracle (ref.)	1.27	23%	45%	13%	0.79

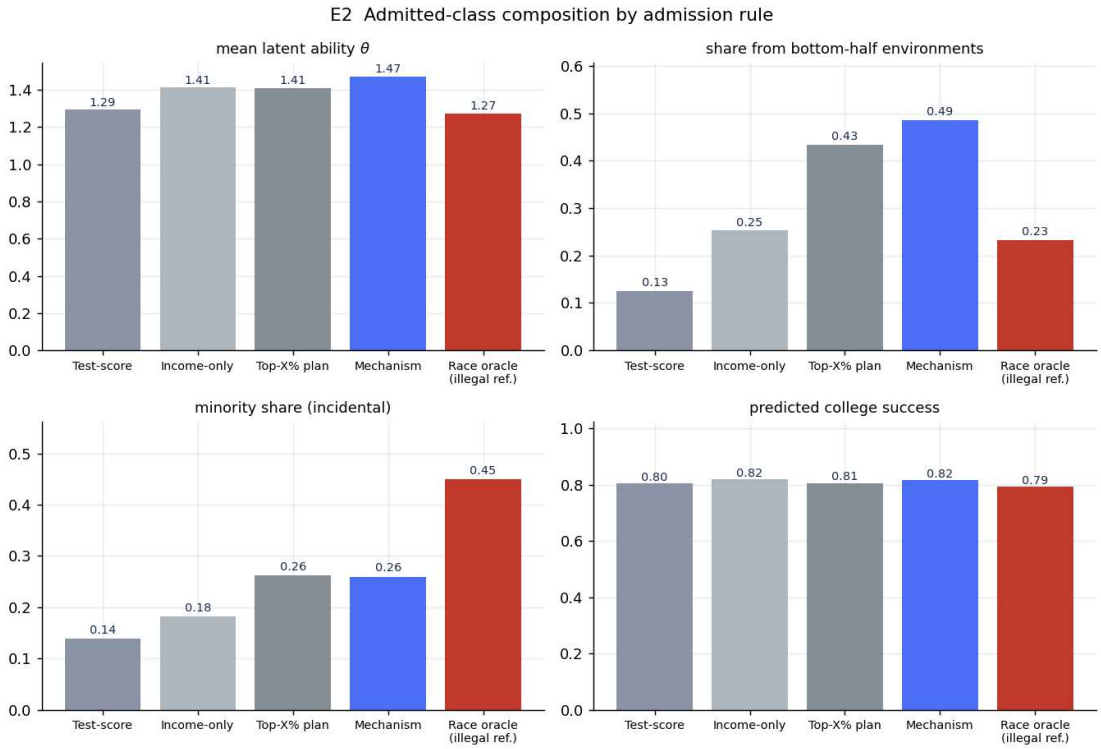


Fig. 2 Admitted-class composition. The mechanism (blue) achieves higher latent ability and greater socioeconomic diversity than test-score ranking, and recovers much of the oracle's incidental minority share without using race.

E3 — anti-gaming

We model advantaged applicants (top 40% by environment) who each flip up to three reported features to inflate their score; verified Tier-B records cannot be faked. We compare the robust mechanism against a *naive* additive index (standardised achievement plus an unweighted, uncapped, unverified sum of all reported factors). The naive index yields a mean manipulation gain of 2.97 (95th percentile 3.00), hands 489 seats to gamers and displaces 428 genuinely disadvantaged applicants. The robust mechanism holds the mean gain to 0.13 (95th percentile 0.225 — exactly the bound of Proposition 1), limits captured seats to 102 and displacement to 65 (Fig. 3): reductions of roughly 96% in manipulation gain, 79% in captured seats and 85% in displacement of the genuinely disadvantaged. Residual gaming is bounded but non-zero — an honest property, since no self-report regime is perfectly strategyproof; the cap is precisely what bounds it.

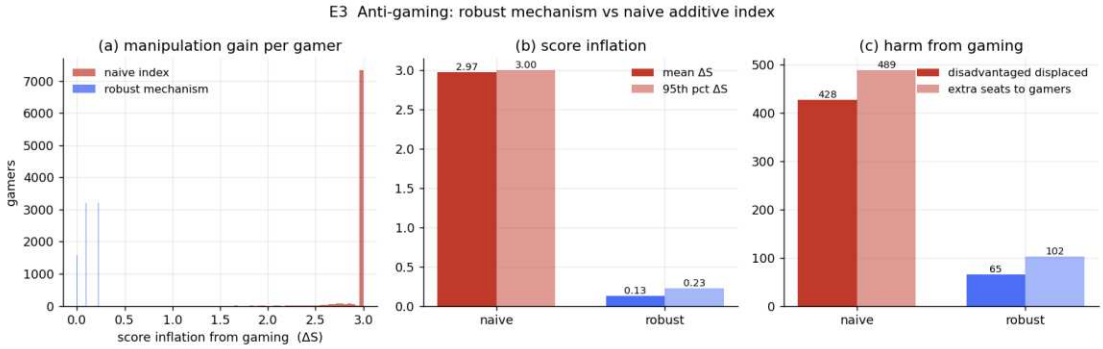


Fig. 3 Robust mechanism versus naive additive index under an adversarial manipulation budget. Left: per-gamer score inflation. Centre: mean and tail inflation. Right: seats captured by gamers and displacement of genuinely disadvantaged applicants.

E4 — sensitivity

Under $\pm$20% independent jitter on all nine weights (300 draws), the composition results are stable: bottom-half-environment share 0.48 (90% band [0.42,0.55]), incidental minority share 0.26 ([0.23,0.29]), and mean ability 1.47 ([1.45,1.48]) (Fig. 4). The conclusions are not artefacts of one hand-tuned weight vector.

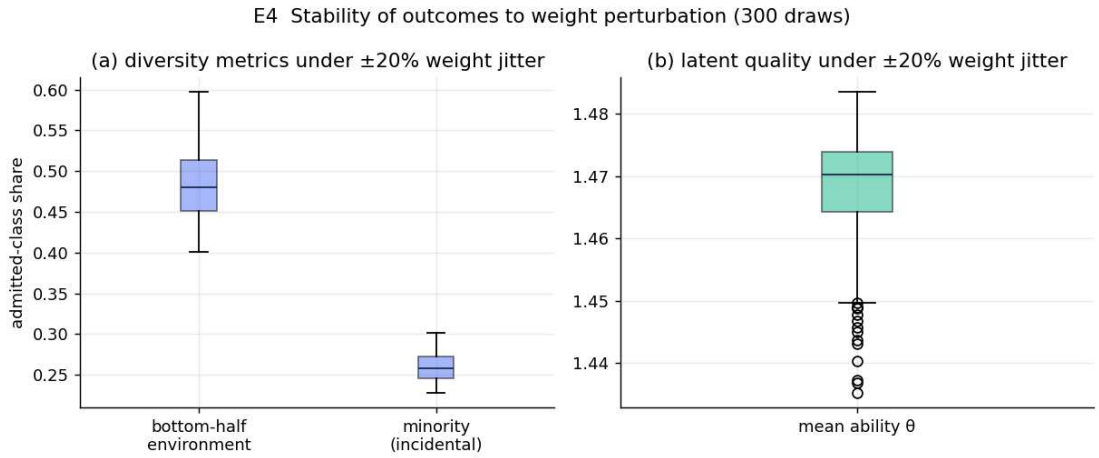


Fig. 4 Stability of admitted-class composition and latent quality to $\pm 20\%$ *perturbation of all mechanism weights.*

Evaluation II: real data

The simulation demonstrates internal validity under a stipulated premise. We now test the premise, and the residual construction, on real data.

Data

We use the Stanford Education Data Archive (SEDA, v4.0) (Reardon et al. 2021): district mean achievement and a socioeconomic-status composite for about 13,000 U.S. school districts, together with pooled 2009–2018 covariates — rural locale, free/reduced-lunch share, poverty, unemployment, SNAP receipt, single-parent share, educational attainment, and median income — obtained from a public mirror. After merging and complete-case filtering, $N = 12{,}379$ districts remain. SEDA is aggregate (district-level), not individual; it therefore validates the mechanism's *assumptions and construction*, not individual admission decisions. We are explicit about this throughout.

P1 — context explains most of achievement

Regressing district mean achievement on the socioeconomic-status composite alone yields $R^2 = 0.59$; on the full context factor set, $R^2 = 0.74$ (Fig. 5a). That roughly three-quarters of the variance in district achievement is explained by socioeconomic environment is the empirical justification for not ranking on raw achievement — and the real-world magnitude that the simulation had merely stipulated.

P2 — the residual is environment-neutral

Raw achievement correlates 0.77 with socioeconomic status; the context-adjusted residual correlates 0.02 (Fig. 5b) — a cleaner separation than in the simulation, because real context variables absorb the environmental signal well. "Distance travelled" is thus demonstrated on about 12,400 real districts.

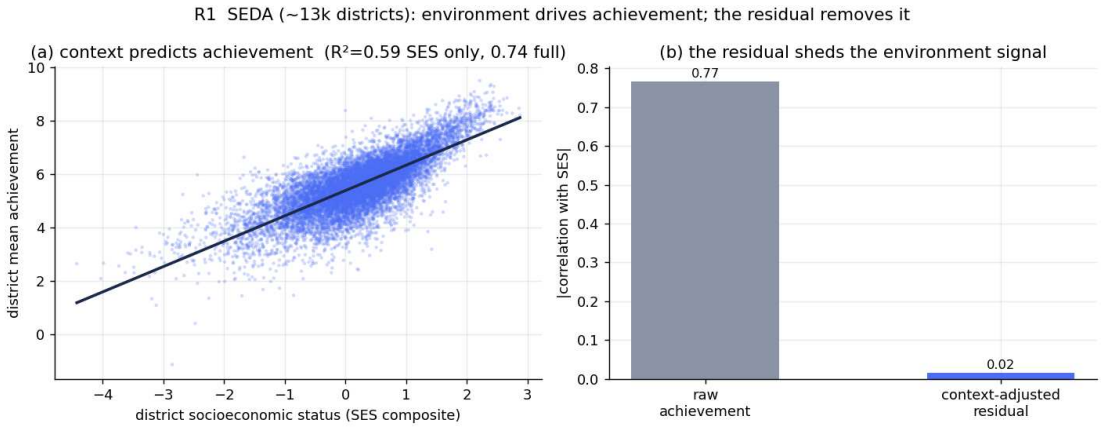


Fig. 5 On real SEDA districts, socioeconomic context predicts achievement (left), and the context-adjusted residual removes almost all of the environment signal that raw achievement carries (right).

Re-ranking: who reaches the top

Building a real adversity index from the SEDA factors (using the tier-and-cap structure of Section 3.4), we "admit" the top 15% of districts by raw achievement versus by the mechanism. Ranking by raw achievement selects affluent, largely suburban districts (mean socioeconomic status +1.27; 31% rural). The mechanism instead selects more disadvantaged, more rural districts (adversity +0.82 vs +0.45; socioeconomic status +0.43; 59% rural) that achieve far above their socioeconomic expectation (mean residual +1.45 vs +0.61) — the district-level image of what the applicant-level mechanism is designed to do (Fig. 6). The real adversity index correlates −0.72 with achievement, confirming it captures genuine disadvantage.

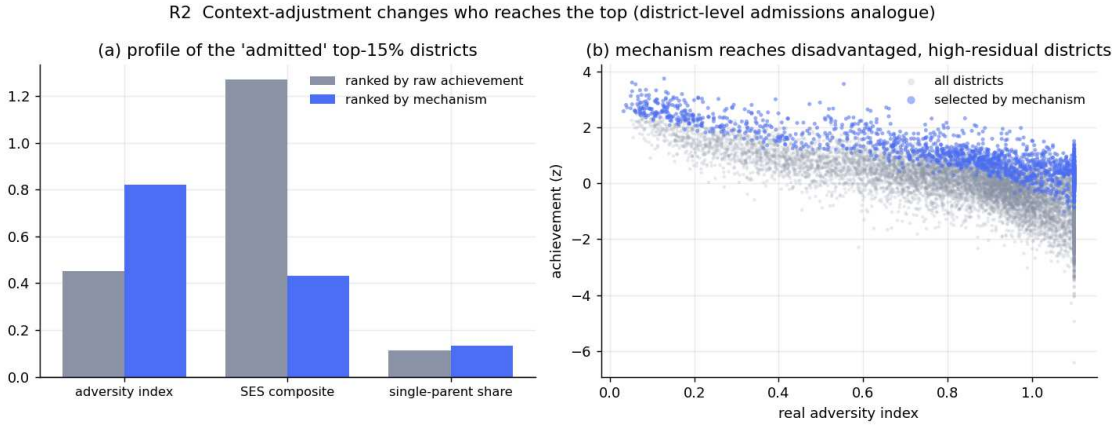


Fig. 6 Context-adjustment changes which districts reach the top. Left: profile of the "admitted" top-15% of districts. Right: the mechanism selects disadvantaged, high-residual districts.

The race-gap decomposition: promise and limit

Finally we quantify, on real data, how much of the achievement gap associated with race a class-based mechanism captures. In a nested regression we estimate the gradient of achievement on district racial composition, first alone (b_1) and then controlling for socioeconomic context (b_2); the share explained by context is $(b_1 - b_2)/b_1$. For the share Black, context explains 89% of the gradient (from $b_1 = -2.20$ to $b_2 = -0.25$ SD per 100% share); for the share Hispanic, 80% (from -1.72 to -0.35) (Table 3, Fig. 7). Equivalently, the correlation between district share Black and achievement falls from -0.34 to -0.04 once context is removed.

Table 3 Race-gap decomposition on about 12,400 SEDA districts. Context (which a class-based mechanism equalises) explains most of the gap; a real remainder persists.

Group	Total gradient b_1	After context b_2	Explained by context	Persists
Share Black	-2.20	-0.25	**89%**	11%
Share Hispanic	-1.72	-0.35	**80%**	20%

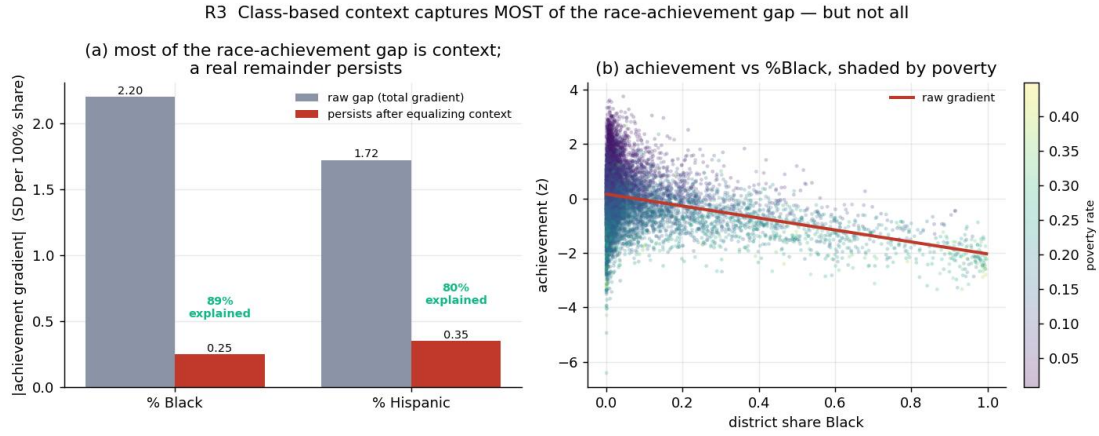


Fig. 7 A class-based (context) mechanism captures most — 80–89% — of the district-level race–achievement gap, but a real remainder persists after equalising context.

Two readings follow, and both belong in the paper. The **promise**: because the mechanism equalises context, it recovers the large majority of what race-conscious admissions targeted — the empirical reason the simulation's incidental minority share reached about 58% of the oracle's without using race. The **honest limit**: a real remainder (11% Black, 20% Hispanic) persists after context is equalised, and a class-based mechanism will not close it. This is the district-level echo of the individual-level finding that a racial achievement gap remains at equal socioeconomic status (Reardon 2011).

A critical caveat governs the *direction* of the inference. This is a district-level (ecological) decomposition; district aggregates are strongly collinear with composition and therefore tend to *over-attribute* the gap to context. The true individual-level remainder is, if anything, *larger* than 11–20%. The district result is thus a lower bound on the persistent racial component — which strengthens rather than weakens the paper's central honesty, and

motivates the individual-level study of Section 8. We do not report 89%/80% as an individual decomposition.

Discussion

What the mechanism does and does not do. It selects for disadvantage transparently, resists gaming with a stated bound, and — under our premise and on real aggregate data — improves measured quality while broadening the class. It recovers most, not all, of the racial diversity of race-conscious admissions; Section 5.5 quantifies the shortfall rather than eliding it. Institutions should adopt it, if at all, with clear eyes about that shortfall.

Transparency versus gaming. There is a genuine tension: a fully published rule is contestable but also, in principle, gameable. Our resolution is to make the *rule* public while resting robustness on *verification* rather than secrecy — Proposition 1 shows the exploitable surface is confined to the capped self-attested tier. This is the opposite of *Landscape*'s opacity-as-defence, and we believe it is the more legitimate design: an applicant can see, and challenge, exactly how they were scored.

Ethical boundaries of adversity measurement. Three ethical hazards deserve naming. (i) *Proxy discrimination*: a race-neutral index could in principle be engineered to track race; auditability is the safeguard, and race-neutrality by construction is checkable precisely because the mechanism is closed-form. (ii) *The boundary problem*: any list of counted adversities excludes some real hardship, and every exclusion is a fairness objection; a principled, published inclusion criterion — and a route to contest omissions — is preferable to a fixed list presented as complete. (iii) *Dignity and privacy*: the self-attested tier reaches into family circumstances (incarceration, orphanhood) that are sensitive; capping its weight limits both its gameability and the pressure on applicants to disclose, and verification regimes should minimise intrusion. These are not incidental; for an admissions mechanism they are the ethical core.

Governance and cost. The mechanism is an input to human review, not a replacement for it; the residual and the adversity index are meant to *surface* applicants for consideration, not to automate a decision. And it does not touch the binding constraint the policy literature stresses: disadvantaged admits need financial aid to enrol and succeed. A transparent selection mechanism is necessary but not sufficient for access.

Limitations

The synthetic population *encodes* the premise that environment suppresses observed achievement; its magnitudes are stipulated, and although Section 5 shows the premise holds on real data, the simulation's specific effect sizes are illustrative. SEDA is aggregate, not individual, so the real-data results validate assumptions and construction rather than individual admission decisions. The race-gap decomposition is ecological and, as argued, likely under-states the persistent racial component. Tier-C factors are modelled as fabricable booleans; real verification regimes sit between "free" and "impossible" and

would change the residual-gaming figure. The simulation uses a single stylised minority group with a fixed environment gap; real recovery of diversity depends on the true race–disadvantage joint distribution and varies by institution. Finally, admission is neither enrolment nor success; the mechanism targets the opportunity to be seen, not a guarantee of either.

Future work

The clear next study is an **individual-level validation** on restricted data such as NCES HSLS:09, constructing the adversity factors at the person level and validating the mechanism against real later college outcomes — the honest counterpart to the district analysis here. Beyond that: **institution-specific calibration** of weights via the outcome-anchored procedure; **field evaluation** of verification regimes for the self-attested tier, including their intrusiveness and error rates; **fairness audits** across multiple groups rather than a single stylised one; and formal work tightening Proposition 1 under partial (probabilistic) verification.

Conclusion

The end of race-conscious admissions has made the *measurement* of disadvantage the central problem in admissions policy, and it is a problem to which the tools of transparent, verifiable mechanism design apply directly. We have proposed a closed-form admissions mechanism that organises disadvantage by verifiability, replaces raw achievement with an environment-neutral "distance travelled" residual, bounds the influence of gameable self-reported factors, and carries a provable limit on manipulation. In simulation it improves quality and diversity together and blunts gaming; on real data it rests on a premise that holds strongly (context explains 74% of achievement), and it lets us state honestly — with numbers — both what class-based selection recovers and what it cannot. That honesty is not a concession but the point: a mechanism institutions can adopt with a clear account of its promise and its limits is more defensible, legally and ethically, than one that claims to reproduce what it cannot.

Declarations

Funding. The author declares that no funding was received for this work.

Competing interests. The author declares no competing interests.

Data availability. The real-data analysis uses the publicly available Stanford Education Data Archive. The synthetic-population generator, the mechanism implementation, and all analysis code are available from the author and will be deposited in a public repository upon acceptance.

Code availability. All simulation and analysis code (Python) is available from the author.

References

Carnevale, Anthony P., and Stephen J. Rose. 2004. "Socioeconomic Status, Race/Ethnicity, and Selective College Admissions." In *America's Untapped Resource: Low-Income Students in Higher Education*, edited by Richard D. Kahlenberg. New York: Century Foundation Press.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. "Fairness Through Awareness." In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, 214–26.

Espenshade, Thomas J., and Alexandria Walton Radford. 2009. *No Longer Separate, Not yet Equal: Race and Class in Elite College Admission and Campus Life*. Princeton, NJ: Princeton University Press.

Fisher v. University of Texas at Austin. 2016. "579 u.s. 365."

Gaertner, Matthew N., and Melissa Hart. 2013. "Considering Class: College Access and Diversity." *Harvard Law & Policy Review* 7: 367–403.

Grutter v. Bollinger. 2003. "539 u.s. 306."

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." In *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 29.

Kahlenberg, Richard D. 1996. *The Remedy: Class, Race, and Affirmative Action*. New York: Basic Books.

Kahlenberg, R. D. (2025). Class Matters: The Fight to Get Beyond Race Preferences, Reduce Inequality, and Build Real Diversity at America's Colleges. PublicAffairs.

Long, Mark C. 2004. "Race and College Admissions: An Alternative to Affirmative Action?" *The Review of Economics and Statistics* 86 (4): 1020–33.

Reardon, Sean F. 2011. "The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations." In *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*, edited by Greg J. Duncan and Richard J. Murnane. New York: Russell Sage Foundation.

Reardon, Sean F., Andrew D. Ho, Benjamin R. Shear, Erin M. Fahle, Demetra Kalogrides, Heewon Jang, Belen Chavez, and Richard DiSalvo. 2021. "Stanford Education Data Archive (Version 4.0)." Stanford University.

Regents of the University of California v. Bakke. 1978. "438 u.s. 265."

Students for Fair Admissions, Inc. v. President and Fellows of Harvard College. 2023. "600 u.s. 181."