

TILTED MONDRIAN: TRIAGE AND MITIGATION OF FINE-GRAINED DISPARITIES IN CONFORMAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Fairness in conformal prediction is often pursued through Mondrian calibration, which provides finite-sample coverage guarantees separately across groups. These guarantees, however, are coarse: substantial disparities can remain within finer strata. This coarse-to-fine fairness issue raises two questions: how severe can such disparities be, and how can they be mitigated? Aggregate disparities can depend on how groups are distributed across strata. For severity assessment, we introduce propensity standardization to examine the effect of population composition, and show how group composition can induce disagreement between aggregate and finer-stratum comparisons. For mitigation, we develop Tilted Mondrian to improve fairness compared to group-Mondrian at a finer level without additional model training. Tilted Mondrian retains groupwise coverage guarantees and mitigates coverage and prediction set size disparities within finer strata. Finally, we derive a method-agnostic bound that quantifies how much reduction in finer-stratum coverage disparity can be achieved.

1 INTRODUCTION

Conformal prediction (CP) (Vovk et al., 2005) attaches prediction sets to fitted classifiers with finite-sample coverage guarantees. A natural way to address groupwise fairness is group-Mondrian calibration (Vovk et al., 2003), which guarantees coverage across prespecified groups. Yet this guarantee is only group conditional. Specifically, substantial disparities may remain within finer strata of the groups. At a fixed coverage level, aggregate group comparisons can obscure behavior within strata. Equalized groupwise coverage guarantees can coexist with large disparities within finer strata, while for set size, group composition can alter the aggregate disparity. As a result, we view both as manifestations of the same coarse-to-fine problem. We first assess the impact of composition, that is, the distribution across finer strata within each group, by comparing set size disparities under the observed and standardized populations. We then adjust the nonconformity score to mitigate disparities within finer strata while preserving groupwise coverage.

Given a CP coverage level, we study the question of how much ambiguity remains, as measured by prediction set size, and how it is distributed across populations. Coverage controls how often the true label is included, but not how many labels a user needs to consider in downstream decision-making. For example, prediction sets of size two and eight can both be valid at the same nominal level yet impose very different burdens on a downstream review process (Cresswell et al., 2024; 2025). We summarize this burden by a set size curve, which records expected prediction set size as a function of coverage level. Group curves can mix difficulty with population composition. A group may receive smaller prediction sets overall simply because its members are concentrated in easier-to-predict subpopulations even when the within-subpopulation comparison goes in the opposite direction. Interpreting finer-grained disparities requires separating within-stratum differences from population composition. We therefore introduce propensity score standardization, which leaves the label predictor, nonconformity scores, calibration thresholds, prediction sets, and coverage guarantees unchanged while reaveraging set sizes over a standardized population. By placing groups on a standardized distribution, the set size curves assess how sensitive an aggregate difference is to population composition.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

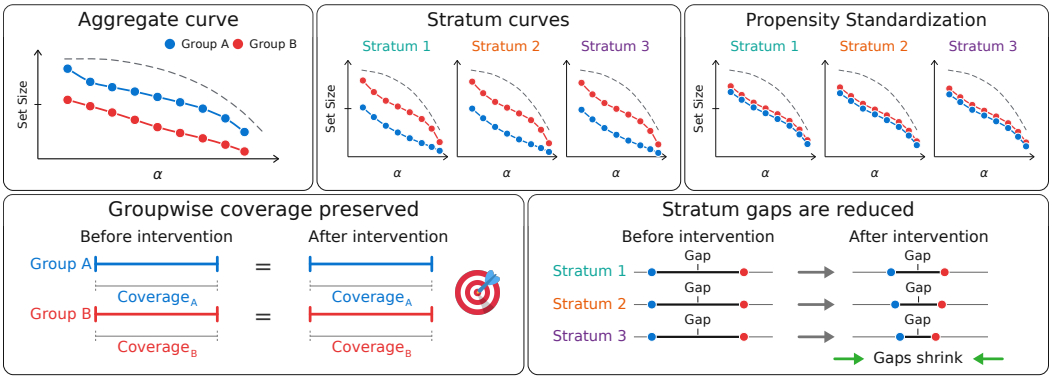


Figure 1: Top: Propensity standardization evaluates the aggregate–stratum set size disagreement. Bottom: Tilted Mondrian mitigates finer strata disparities and maintains groupwise coverage.

We next turn to score intervention when coarse groupwise coverage masks disparities across finer strata. For example, even when group-Mondrian calibration equalizes coverage across female and male groups, substantial disparities may persist within finer-grained strata, such as occupations. We introduce Tilted Mondrian CP, a score adjustment framework for mitigating fine-grained disparities while preserving groupwise coverage guarantees. When targeting coverage, it seeks to reduce stratum coverage disparities while measuring set size as an efficiency cost. The same framework can also target stratum set size disparities and track changes in coverage.

Our two main contributions, summarized in Figure 1, are as follows.

- **Severity assessment of composition via propensity standardization.** We introduce propensity standardization that evaluates composition within finer stratum disparities, revealing how sensitive aggregate–stratum misalignment is to population composition.
- **Tilted Mondrian and mitigation limits.** We develop Tilted Mondrian as a generic mitigation approach that can be instantiated at any Mondrian level and can target finer-stratum disparities in coverage and set size. For coverage mitigation, we derive a method-agnostic bound characterizing the mitigation limits under equalized group coverage.

The remainder of this paper is organized as follows: Section 2 reviews related work; Sections 3–4 develop the composition assessment; Section 5 studies Tilted Mondrian mitigation; Section 6 presents the experimental evidence on assessment and mitigation; and Section 7 concludes.

2 Related Work

Conformal prediction gives distribution-free set-valued coverage under exchangeability (Vovk et al., 2005; Angelopoulos & Bates, 2023). Average set size is a standard efficiency measure for set-valued predictors. Methods such as APS, RAPS, and SAPS construct adaptive classification sets that aim to minimize average set size (Romano et al., 2020b; Angelopoulos et al., 2021; Huang et al., 2023). We formalize set size efficiency as a curve across miscoverage levels and read its height as residual ambiguity. Related work studies direct tradeoffs between set size and conditional coverage (Bach, 2025), and connects CP set size to information theory (Correia et al., 2024). We study set size and use classical entropy–error relations (Fano, 1961; Cover & Thomas, 2006; Feder & Merhav, 1994) to explain when large sets may be unavoidable.

Fairness-oriented CP studies equalized coverage, adaptively selected groups, demographic-parity quantiles, equal-opportunity coverage, counterfactual fairness, and domain-specific fair predictors (Romano et al., 2020a; Zhou & Sesia, 2024; Vadlamani et al., 2025; Liu et al., 2022; Wang et al., 2023; Guldogan et al., 2025; Lu et al., 2022). Complementary work examines human decisions and substantive fairness when set-size burdens differ across groups (Cresswell et al., 2024; 2025; Liu et al., 2026; Gao et al., 2026). Our objective is more specific: a descriptive ambiguity comparison for a fixed predictor, fixed calibration rule, and stated evaluation population.

We connect population imbalance to Simpson's paradox (Simpson, 1951). Our standardization step draws on propensity reweighting, covariate balancing, and overlap populations (Rosenbaum & Ru-

bin, 1983; Imai & Ratkovic, 2014; Li et al., 2018). Related methods pursue conditional coverage via multivalid or conditional calibration (Jung et al., 2022; Gibbs et al., 2025; Bairaktari et al., 2025), or score transformations (Xie et al., 2024; Plassier et al., 2025). Prior methods primarily ask how to achieve conditional coverage or set efficiency. Our objective is to mitigate stratum-level disparities under explicit groupwise coverage and stratum constraints.

3 Aggregate Stratum Disagreement in Set Size Curves

Consider a multiclass problem with input X, label $Y \in \mathcal{Y}$ where $\mathcal{Y}$ is a finite label space of size $K \geq 2$, and a nonconformity score $s(x, y)$, where smaller values indicate that y is more compatible with x. Given an exchangeable calibration sample $D_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{n_{\text{cal}}}$, write $S_i = s(X_i, Y_i)$, $i = 1, \ldots, n_{\text{cal}}$, for the calibration scores. Split CP computes the empirical $(1-\alpha)$-quantile

$$\hat{q}_\alpha = \text{Quantile}\big(1-\alpha; \{S_i\}_{i=1}^{n_{\text{cal}}} \cup \{\infty\}\big), \tag{1}$$

equivalently the $\lceil (n_{\text{cal}}+1)(1-\alpha) \rceil$-th order statistic of the augmented multiset. The calibrated prediction set at level $1-\alpha$ is

$$\mathcal{C}_\alpha(x) = \{y \in \mathcal{Y} : s(x, y) \leq \hat{q}_\alpha\}. \tag{2}$$

Under exchangeability between the calibration and the test point, split CP guarantees

$$\mathbb{P}\{Y_{\text{test}} \in \mathcal{C}_\alpha(X_{\text{test}})\} \geq 1-\alpha. \tag{3}$$

Coverage controls true-label inclusion, but not set size.

Definition 1 (Set size curve)**.** Let $M_\alpha(X) = |\mathcal{C}_\alpha(X)|$ denote the size of the prediction set. Given a population P, the set size curve is $V^P(\alpha) := \mathbb{E}_{X \sim P} M_\alpha(X)$, the expected set size at level α.

Here, P indexes the population over which the post-calibration set sizes are averaged. Thus, $V^P(\alpha)$ summarizes the CP set efficiency at each miscoverage level. We omit the superscript P when the population is clear from context. A set size curve should be interpreted against intrinsic label ambiguity. Appendix B discusses that CP sets cannot be arbitrarily small. When the label Y remains hard to predict from the input X, CP must use larger prediction sets on average to maintain the target $1-\alpha$ coverage level. We now turn to the curves $V(\alpha)$ and examine how the associated ambiguity burden of set size is distributed across groups and strata of X.

Let $G \in \{a, b\}$ denote the comparison group, e.g. binary gender. The raw group curves are

$$V_g(\alpha) = \mathbb{E}\{M_\alpha(X) \mid G = g\}, \qquad g \in \{a, b\}, \tag{4}$$

and the raw aggregate gap is $\Delta(\alpha) = V_a(\alpha) - V_b(\alpha)$. The estimand $V_g(\alpha)$ indicates how many labels are delivered for each group $G = g$. It is a meaningful population summary, but its fairness interpretation depends on whether the observed mixture is the comparison population of interest.

To expose what the aggregate mixes, in addition to G, we consider another prespecified reporting variable $Z \in \mathcal{Z}$, where $\mathcal{Z}$ is a finite set, such as an occupation, education bucket, or employment category. We refer to each group–stratum pair (g, z) as a cell. Define $V_{g,z}(\alpha) = \mathbb{E}\{M_\alpha(X) \mid G = g, Z = z\}$ and $\Delta_z(\alpha) = V_{a,z}(\alpha) - V_{b,z}(\alpha)$. Let $\pi_g(z) = \mathbb{P}(Z = z \mid G = g)$. For every α, by the law of total expectation, one can obtain the set size gap decomposition

$$\Delta(\alpha) = \sum_{z \in \mathcal{Z}} \pi_a(z) \Delta_z(\alpha) + \sum_{z \in \mathcal{Z}} \{\pi_a(z) - \pi_b(z)\} V_{b,z}(\alpha). \tag{5}$$

If groups a and b have the same cell mixture $\pi_a(z) = \pi_b(z)$, then the aggregate gap $\Delta(\alpha)$ reduces to the weighted average of the stratum gaps $\sum_{z \in \mathcal{Z}} \pi_a(z) \Delta_z(\alpha)$, but this is not generally the case. The first term in Equation (5) averages stratum gaps under group a's mixture. The second isolates composition, which can make the aggregate gap $\Delta(\alpha)$ have the opposite sign from a stratum gap $\Delta_z(\alpha)$. This is an analog of Simpson's paradox (Simpson, 1951) realized in set sizes.

When an aggregate group comparison differs from the corresponding finer-stratum comparisons, we refer to this as aggregate–stratum disagreement. On the supported (z, α) grid, define

$$\text{Reversal for } z\text{:} \quad \Delta(\alpha)\Delta_z(\alpha) < 0, \quad \text{Conflict strength for } z\text{:} \quad \Gamma_z(\alpha) = [-\Delta(\alpha)\Delta_z(\alpha)]_+, \tag{6}$$

respectively. We interpret a reversal jointly with conflict strength. A reversal records when aggregate and stratum gaps point in opposite directions, whereas conflict strength accounts for the magnitudes of both gaps, so a reversal with negligible conflict strength carries little weight. We use these quantities to report reversal counts and conflict strength in the experiments in Section 6.

From Equation (6), $\Gamma_z(\alpha) > 0$ identifies an aggregate–stratum sign conflict. Equation (5) shows that such disagreement can arise from $\pi_a \neq \pi_b$. The next section therefore reaverages the same $M_\alpha(X)$ over a standardized comparison population to assess the severity of the composition effect.

4 Propensity Score Standardization for Disparity Assessment

The raw group curves average the set size over the group-specific population distributions. To separate this population mixture component from disparities in set size, we apply propensity score reweighting to place both groups on a standardized distribution. The fitted predictor and CP procedure remain fixed; only the population over which their set sizes are averaged is changed.

Let $H = H(X)$ denote the *profile*, a prespecified vector of balancing variables defined as a function of X, which could include probabilities from a classifier, margins, entropy, and top-label indicators. The profile H excludes the group variable G, realized labels, realized set sizes, calibration quantities, and final curve summaries. The profile H defines the target population to be standardized, whereas Z defines the finer strata used to assess aggregate–stratum alignment. Including Z among the balancing variables would instead standardize the joint (H, Z) distribution and thus define a different estimand. The balancing variables are fixed before the final curve analysis.

Let P_H be the marginal distribution of H, and define the group propensity score (Li et al., 2018)

$$e(h) = \mathbb{P}(G = a \mid H = h). \tag{7}$$

The propensity score describes how profiles are distributed across the two groups and determines the standardization weights. Comparing the raw and standardized curves then assesses how the aggregate–stratum relationship changes under the standardized population.

One option is to standardize both groups to the marginal profile distribution P_H. Up to normalizing constants,

$$\mathbb{P}_{H|G=a} \propto e(h)P_H, \qquad \mathbb{P}_{H|G=b} \propto \{1 - e(h)\}P_H. \tag{8}$$

Standardizing both groups to P_H applies the inverse propensity weights $1/e(H)$ for group a and $1/\{1 - e(H)\}$ for group b. This inverse probability weighting (IPW) retains the marginal profile population, but the weights can become unstable when $e(H)$ approaches 0 or 1, making the comparison sensitive to limited overlap. For the main analysis, we apply propensity reweighting

$$Q_H = \frac{e(H)\{1 - e(H)\}P_H}{\mathbb{E}[e(H)\{1 - e(H)\}]} \quad \text{provided that } \mathbb{E}[e(H)\{1 - e(H)\}] > 0\,. \tag{9}$$

The corresponding weights are $w_a(H) = 1 - e(H)$ and $w_b(H) = e(H)$, which map both groups to the standardized target distribution Q_H. Equivalently, H and G are independent in the weighted population. The target emphasizes profiles represented in both groups and downweights profiles with small overlap. The propensity reweighting method also avoids the extreme weighting that can arise under classical IPW. Thus, we identify Q_H as the target population for the standardized curves. For group $g \in \{a, b\}$, define the standardized curve by

$$\tilde{V}_g(\alpha) = \mathbb{E}_{Q_H}\,\mathbb{E}\{M_\alpha(X) \mid G = g, H\}. \tag{10}$$

The raw curve $V_g(\alpha)$ averages over the current data distribution, whereas $\tilde{V}_g(\alpha)$ averages over the standardized distribution Q_H. Empirically, the standardized curve $\tilde{V}_g(\alpha)$ is estimated by a normalized Hájek mean with propensity weights (Hájek, 1971). The same propensity reweighting induces a group-specific distribution over the stratum variable Z:

$$\tilde{\pi}_g(z) = \frac{\mathbb{E}\{\mathbf{1}(G = g)w_g(H)\mathbf{1}(Z = z)\}}{\mathbb{E}\{\mathbf{1}(G = g)w_g(H)\}} = \mathbb{E}_{Q_H}[\mathbb{P}(Z = z \mid G = g, H)]. \tag{11}$$

For any cell (g, z) with $\tilde{\pi}_g(z) > 0$, define the standardized stratum curve

$$\tilde{V}_{g,z}(\alpha) = \frac{1}{\tilde{\pi}_g(z)}\mathbb{E}_{Q_H}[\mathbb{E}[\mathbf{1}(Z = z)M_\alpha(X) \mid G = g, H]]. \tag{12}$$

Then $\tilde{V}_g(\alpha) = \sum_z \tilde{\pi}_g(z)\tilde{V}_{g,z}(\alpha)$. Define the standardized aggregate and stratum gaps by

$$\tilde{\Delta}(\alpha) = \tilde{V}_a(\alpha) - \tilde{V}_b(\alpha), \qquad \tilde{\Delta}_z(\alpha) = \tilde{V}_{a,z}(\alpha) - \tilde{V}_{b,z}(\alpha). \tag{13}$$

The next proposition shows that both reweighted groups follow the standardized distribution Q_H. *Complete proofs for all theoretical results are given in Appendix A.*

Proposition 2 (Propensity score reweighting)**.** *For every integrable function f of H,*

$$\frac{\mathbb{E}\{f(H)\mathbf{1}(G=a)(1-e(H))\}}{\mathbb{E}\{\mathbf{1}(G=a)(1-e(H))\}} = \frac{\mathbb{E}\{f(H)\mathbf{1}(G=b)e(H)\}}{\mathbb{E}\{\mathbf{1}(G=b)e(H)\}} = \mathbb{E}_{Q_H}[f(H)]. \tag{14}$$

Take-away. Propensity standardization puts both groups on the same population distribution, making an assessment of set sizes more comparable.

To relate the standardized aggregates to stratum gaps, let $\mathbb{E}_Z$ and Var_Z denote expectation and variance under the symmetric raw distribution $\{\pi_a(z) + \pi_b(z)\}/2$, and let $\mathbb{E}_{\tilde{Z}}$ and $\mathrm{Var}_{\tilde{Z}}$ denote the corresponding quantities under $\{\tilde{\pi}_a(z) + \tilde{\pi}_b(z)\}/2$. Define $R(\alpha) = \Delta(\alpha) - \mathbb{E}_Z\Delta_Z(\alpha)$ and $\tilde{R}(\alpha) = \tilde{\Delta}(\alpha) - \mathbb{E}_{\tilde{Z}}\tilde{\Delta}_Z(\alpha)$. Thus, $R(\alpha)$ is the composition effect in the raw aggregate gap, and $\tilde{R}(\alpha)$ is its standardized counterpart. Equations (25) and (26) give the explicit form of $\tilde{R}(\alpha)$ and show that it is controlled by the difference between the two standardized distributions $\tilde{\pi}_a$ and $\tilde{\pi}_b$.

To quantify aggregate–stratum alignment, we formulate the following alignment losses

$$\mathcal{A}_{\mathrm{raw}}(\alpha) = \mathbb{E}_Z\{\Delta(\alpha) - \Delta_Z(\alpha)\}^2, \qquad \mathcal{A}_{\mathrm{std}}(\alpha) = \mathbb{E}_{\tilde{Z}}\{\tilde{\Delta}(\alpha) - \tilde{\Delta}_Z(\alpha)\}^2. \tag{15}$$

A smaller $\mathcal{A}$ means that the aggregate gap is closer, on average, to the corresponding stratum gaps.

For mass functions p, q on the stratum grid, we define the total variation $\mathrm{TV}(p,q) := \frac{1}{2}\sum_z |p(z) - q(z)|$. Although propensity reweighting equalizes the distribution of H, the conditional distribution of Z given H may still differ across groups. We quantify this remaining difference by

$$\eta_H = \mathbb{E}_{Q_H}[\mathrm{TV}\{\mathbb{P}(Z \in \cdot \mid G=a, H), \mathbb{P}(Z \in \cdot \mid G=b, H)\}]\,. \tag{16}$$

Thus, η_H computes the average conditional difference between the two stratum distributions at the same profile H. In particular, η_H vanishes exactly when $\mathbb{P}(Z \in \cdot \mid G=a, H=h) = \mathbb{P}(Z \in \cdot \mid G=b, H=h)$ for Q_H-almost every h. Next, define

$$\Delta_{h,z}(\alpha) := \mathbb{E}\{M_\alpha(X) \mid G=a, H=h, Z=z\} - \mathbb{E}\{M_\alpha(X) \mid G=b, H=h, Z=z\}. \tag{17}$$

This quantity compares the set size between groups at a fixed profile h and stratum z. When η_H vanishes, we say that composition dominance holds at level α if

$$\operatorname*{ess\,sup}_{h,z} |\Delta_{h,z}(\alpha) - \mathbb{E}\{\Delta_{H,Z}(\alpha)\}| < |R(\alpha)|\,, \tag{18}$$

where the essential supremum and expectation are taken under the joint distribution induced by $H \sim Q_H$ and the conditional distribution of $Z \mid H$.

Theorem 3 (Alignment under standardization)**.** *Given a fixed level α, the propensity score standardization satisfies $\mathrm{TV}(\tilde{\pi}_a, \tilde{\pi}_b) \le \eta_H, |\tilde{R}(\alpha)| \le K\eta_H$. Suppose η_H vanishes, equivalently $\mathbb{P}(Z \in \cdot \mid G=a, H=h) = \mathbb{P}(Z \in \cdot \mid G=b, H=h)$ for Q_H-almost every h, then $\tilde{R}(\alpha) = 0$ and*

$$\mathcal{A}_{\mathrm{std}}(\alpha) \le \mathrm{Var}\{\Delta_{H,Z}(\alpha)\}\,. \tag{19}$$

Furthermore, suppose the composition dominance holds (Equation (18)), then

$$\mathcal{A}_{\mathrm{std}}(\alpha) < \mathcal{A}_{\mathrm{raw}}(\alpha)\,. \tag{20}$$

In particular, suppose $\Delta_{h,z}(\alpha)$ is constant for every h, z pair, then $\mathcal{A}_{\mathrm{std}}(\alpha) = 0$.

Take-away. Comparing the current population with a balanced population reveals how severely the composition affects the aggregate–stratum misalignment. A larger reduction indicates that propensity standardization removes a more severe composition effect. The remaining misalignment reflects heterogeneity across the strata of the data.

Operationally, in Theorem 3, comparing $\mathcal{A}_{\mathrm{raw}}$ and $\mathcal{A}_{\mathrm{std}}$ quantifies how sensitive aggregate–stratum misalignment is to the population used for averaging. A large reduction indicates that the observed raw aggregate is strongly shaped by the population mixture of the observed dataset. Algorithm 1 summarizes the propensity standardization procedure. Table 8 in Appendix C displays the decomposition with empirical results. Our analysis here is descriptive, and we make no causal claims.

Algorithm 1: Propensity Standardization
Input: $\mathcal{D}_{\text{prop}}, \mathcal{D}_{\text{cal}}, \mathcal{D}_{\text{test}}$; group G, stratum Z, profile rule H; grid $\boldsymbol{\alpha}$.

1: Construct H from the frozen outputs. On $\mathcal{D}_{\text{prop}}$, fit $\hat{e}(h) = \hat{\mathbb{P}}(G = a \mid H = h)$.
2: Define standardized target Q_H by Equation (9); freeze weights/support.
3: On $\mathcal{D}_{\text{cal}}$, perform split CP for all α.
4: On $\mathcal{D}_{\text{test}}$, set $\hat{e}_i = \hat{e}(H_i)$, $w_{a,i} \propto 1 - \hat{e}_i$, and $w_{b,i} \propto \hat{e}_i$; bound/normalize by group.
5: **for** $\alpha \in \boldsymbol{\alpha}$ **do**
6: Form each $\mathcal{C}_{\alpha,i}$; set $M_{\alpha,i} = |\mathcal{C}_{\alpha,i}|$.
7: From $M_{\alpha,i}$/cells, compute raw (unweighted) and standardized (Equation (10)) group/cell curves.
8: Compare aggregate/strata gap signs; record reversals/strengths (Equation (6)).
9: **end for**
10: **return** raw/standardized reports from fixed prediction sets.

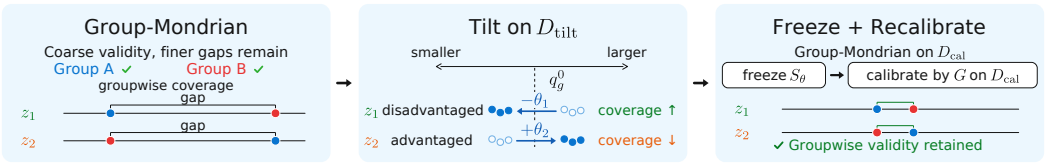


Figure 2: Tilted Mondrian mechanism for finer strata disparity mitigation.

5 Tilted Mondrian and Mitigation Limits

5.1 Tilted Mondrian CP

We now turn to mitigation and propose Tilted Mondrian CP, a generic framework built on a chosen Mondrian calibration level. The intervention acts through a learned adjustment of the nonconformity score. The algorithm shifts coverage across finer strata to reduce disparities while maintaining groupwise coverage guarantees. We identify coverage, the reliability quantity directly guaranteed by CP, as the primary mitigation target.

For each $z \in \mathcal{Z}$, let $c^0_{g,z}$ denote the baseline cellwise coverage for group g and stratum z. Let $g^+(z)$ and $g^-(z)$ be the groups with higher and lower baseline coverage, respectively, and define the baseline gap $d_z = c^0_{g^+(z),z} - c^0_{g^-(z),z}$. Tilted Mondrian uses nonnegative score adjustments θ^+_z and θ^-_z. It raises scores in $g^+(z)$ by θ^+_z and lowers scores in $g^-(z)$ by θ^-_z. Let S_θ denote the transformed score, with $\theta = 0$ as the baseline. Let $c^\theta_{g,z}$ denote the resulting cellwise coverage, and define $d^\theta_z = |c^\theta_{g^+(z),z} - c^\theta_{g^-(z),z}|$ and $m^\theta_z = d_z - d^\theta_z$. Here, d^θ_z is the coverage gap between the two groups in stratum z after the adjustment, and m^θ_z is its reduction from the baseline gap d_z; $m^\theta_z > 0$ indicates an improvement. A score shift is feasible only if the coverage loss in every initially advantaged cell is at most ϵ, coverage does not decrease in any initially disadvantaged cell, and stratum coverage gaps are non-increasing. During the tilt procedure, we preserve each group's baseline empirical coverage. We assume that $\mathcal{Z}$ partitions each group population and that $\pi_{g^-(z)}(z) > 0$ for all $z \in \mathcal{Z}$ (recall $\pi_g(z) = \mathbb{P}(Z = z \mid G = g)$). Figure 2 illustrates the score adjustment and subsequent groupwise recalibration. Algorithm 2 summarizes Tilted Mondrian CP.

Theorem 4 (Validity of Tilted Mondrian)**.** *Let θ be selected on a tilt split independent of calibration. If group-Mondrian split conformal calibration is applied to S_θ to generate prediction sets $\mathcal{C}_{\alpha,\theta}(X)$ as in Equation (2), and calibration and test examples are exchangeable conditional on $G = g$, then $\mathbb{P}\{Y \in \mathcal{C}_{\alpha,\theta}(X) \mid G = g\} \geq 1 - \alpha$ for each group $g \in \{a, b\}$.*

Take-away. Tilted Mondrian maintains the group-Mondrian CP coverage guarantee.

The cellwise post-processing adjustment of Tilted Mondrian also allows improvements to occur jointly across strata. When the same group has lower coverage in one cell and higher coverage in another, the two gaps can be reduced together. Let q^0_g denote the baseline threshold for group g. The following theorem formalizes this case.

Theorem 5 (Local mitigation)**.** *Assume $g = g^-(z_i) = g^+(z_j)$ for supported z_i, z_j with $d_{z_i}, d_{z_j} > 0$ and $\epsilon > 0$. Suppose the population cellwise nonconformity score distributions admit densities that are locally positive at the threshold q^0_g. Then the Tilted Mondrian family admits a nonzero feasible tilt θ preserving group coverage such that $d^\theta_{z_i} < d_{z_i}$ and $d^\theta_{z_j} < d_{z_j}$, equivalently $m^\theta_{z_i} > 0$ and $m^\theta_{z_j} > 0$, with all other stratum gaps unchanged.*

Algorithm 2: Tilted Mondrian score adjustment
Input: Baseline nonconformity score s; $\mathcal{D}_{\text{tilt}}, \mathcal{D}_{\text{cal}}, \mathcal{D}_{\text{test}}$; G, Z; error level α; tolerance ϵ.

On $\mathcal{D}_{\text{tilt}}$, compute baseline group-Mondrian thresholds; record group coverage levels and cellwise coverage $c^0_{g,z}$.
For each z, let $g^+(z)/g^-(z)$ be the higher/lower-coverage groups and d_z their gap.
for each deterministic candidate shift θ **do**
    Shift candidate scores by $+\theta^+_z$ in $g^+(z)$, $-\theta^-_z$ in $g^-(z)$; set $z = y$ for $Z = Y$, else $z = Z(x)$.
    Reselect group thresholds on $\mathcal{D}_{\text{tilt}}$ to keep group coverage levels; compute $c^\theta_{g,z}, d^\theta_z$.
    Set gap reduction $m^\theta_z = d_z - d^\theta_z$.
    Retain θ iff $c^0_{g^+(z),z} - c^\theta_{g^+(z),z} \le \epsilon$, $c^\theta_{g^-(z),z} \ge c^0_{g^-(z),z}$, and $d^\theta_z \le d_z$ for all z.
end for
Select in order: min worst gap; min mean gap; max total mitigation; min shift norm; and, only for remaining ties, minimum tilt split average set size.
Freeze score; group-Mondrian recalibrate on $\mathcal{D}_{\text{cal}}$; evaluate once on $\mathcal{D}_{\text{test}}$.
return selected score and test report.

Take-away. Under these conditions, Tilted Mondrian can strictly improve group-Mondrian by reducing coverage disparities without changing group coverage.

When finer-stratum fairness and coverage guarantees are themselves the objective, Mondrian calibration can be performed separately within each cell, rather than just at the group level. However, when working with a finite calibration set that becomes finely partitioned, cell-wise coverage suffers from high variance (Angelopoulos & Bates, 2023). Our setting is different. We ask whether disparities within Z can be mitigated while calibration remains only at the coarser group level G. Tilted Mondrian produces thresholds at the group level after score shifting, avoiding fine partitions of the calibration data, and mitigates disparities at the cell level with score adjustments. Tables 18 and 19 in Appendix E.1 compare Tilted Mondrian to the alternative: partitioning at the cell level and applying Mondrian for cell-level thresholds. Overall, Tilted Mondrian offers a practical mechanism for fairness improvement beyond the group level.

Tilted Mondrian can also target set size disparities. Let $\bar{M}^0_{g,z}$ denote the baseline mean prediction set size and $L^0_{z,S} = |\bar{M}^0_{a,z} - \bar{M}^0_{b,z}|$; define $L^\theta_{z,S}$ and $m^\theta_{z,S} = L^0_{z,S} - L^\theta_{z,S}$ analogously after the mitigation. Tilted Mondrian raises nonconformity scores in the group with the larger baseline mean set size and lowers them in the group with the smaller mean set size. Candidate tilts preserve group coverage level on $\mathcal{D}_{\text{tilt}}$, keep set size gaps and larger-side means non-increasing, and limit smaller-side increases. The adjusted score is frozen before group-Mondrian calibration on $\mathcal{D}_{\text{cal}}$. Details and coverage trade-offs about Tilted Mondrian for set size are provided in Appendix G.

We also considered a covariate-adaptive alternative inspired by boosted CP (Xie et al., 2024). Tilted Mondrian often achieves larger worst-gap reductions and lower mean cell disparities than the boosted variant (Appendix E.2). The remaining question is how much local improvement is possible when group coverage and cellwise harm are held fixed.

5.2 A Mitigation Limit under Fixed Group Coverage

Tilted Mondrian leaves the cell composition $\pi_g(z)$ unchanged and instead modifies scores to balance cell-level coverage. Because group coverage is the π_g-weighted average of cell coverage, preserving group coverage requires $\sum_{z\in\mathcal{Z}} \pi_g(z)(c^\theta_{g,z} - c^0_{g,z}) = 0$. Consequently, a coverage gain in one cell must be offset elsewhere within the same group, motivating the following mitigation limit.

Theorem 6 (Mitigation bound)**.** *Consider a class of disparity mitigation methods satisfying the same constraints as Tilted Mondrian: they preserve each group's baseline coverage, lose no more than ϵ coverage for each advantaged cell, do not reduce coverage in any disadvantaged cell, and do not increase stratum gaps. Assume $\pi_{g^-(z)}(z) > 0$ for every $z \in \mathcal{Z}$. Then*

$$\max_{z\in\mathcal{Z}} m^\theta_z \le \max_{z\in\mathcal{Z}} \min\left\{ d_z, \epsilon\left(1 + \frac{\sum_{r:g^+(r)=g^-(z)} \pi_{g^-(z)}(r)}{\pi_{g^-(z)}(z)}\right)\right\}. \tag{21}$$

Take-away. Disparities can be reduced by giving up some coverage in advantaged strata. The bound quantifies the maximum improvement achievable by any disparity mitigation strategy satisfying these constraints, including Tilted Mondrian.

6 EMPIRICAL EVIDENCE

We examine propensity standardization for aligning aggregate set size reports with finer-stratum and Tilted Mondrian for reducing finer-stratum disparities in coverage and set size. Groupwise coverage guarantees are retained throughout the mitigation procedures.

Datasets and tasks. The four real-world experimental settings are Bias in Bios occupation prediction with gender groups G and occupation strata Z (De-Arteaga et al., 2019); Folktables occupation prediction with gender groups G and education strata Z across five ACS years (Ding et al., 2021); Folktables income quintile prediction with race groups G and education strata Z; ATUS time-use activity prediction with gender groups G and employment strata Z (U.S. Bureau of Labor Statistics, 2025). We evaluate propensity standardization on set size disparities, and Tilted Mondrian mitigation of finer-stratum coverage disparities, along with its set size variant.

Models and scores. We begin with training, calibration, and test data splits. $\mathcal{D}_{\text{train}}$ fits DistilBERT (Sanh et al., 2019) for Bias in Bios, year-specific histogram gradient boosting for Folktables, and regularized histogram gradient boosting for ATUS. We test with four score functions based on model probabilities $\hat{p}(x)$: SIMPLE, $s(x, y) = 1 - \hat{p}_y(x)$, APS (Romano et al., 2020b), RAPS (Angelopoulos et al., 2021), and SAPS (Huang et al., 2023). The calibration data is divided into two disjoint parts; one remains $\mathcal{D}_{\text{cal}}$, and the other is used as $\mathcal{D}_{\text{prop}}$ or $\mathcal{D}_{\text{tilt}}$. $\mathcal{D}_{\text{prop}}$ fits propensity weights on profile H using a feed-forward network for Bias in Bios, or logistic regression for other datasets. $\mathcal{D}_{\text{tilt}}$ selects the score adjustments θ_z. $\mathcal{D}_{\text{cal}}$ calibrates the group-Mondrian thresholds. Finally, $\mathcal{D}_{\text{test}}$ is used for evaluation. Each split is random from the overall pool of data, making data in each set exchangeable. Appendix C gives the detailed experiment protocols and model specifications.

Metrics. For propensity standardization, Table 1 reports the number of reversals, high-conflict reversals (conflict strength > 0.020), and conflict strength from Equation (6). The normalized effective sample size (ESS) (Kish, 1992) is $(\sum_i w_i)^2 / \sum_i w_i^2$, divided by $|\mathcal{D}_{\text{test}}|$.

For coverage disparity mitigation with Tilted Mondrian, Table 2 reports the worst stratum coverage gap before and after adjustment, $L^0_{\max} = \max_z d_z$ and $L^\theta_{\max} = \max_z d^\theta_z$, together with the largest stratum gap reduction $\max_z m^\theta_z$. Harm is the largest coverage loss among initially advantaged cells using tolerance $\epsilon = 0.01$, and $\Delta|C|$ is the change in average test set size. Similarly, for set size disparity mitigation, let $L^0_{z,S}$ and $L^\theta_{z,S}$ denote the stratum set size gaps before and after adjustment, and define $m^\theta_{z,S} = L^0_{z,S} - L^\theta_{z,S}$. Table 3 reports $L^0_{\max,S} = \max_z L^0_{z,S}$, $L^\theta_{\max,S} = \max_z L^\theta_{z,S}$, and the largest stratum gap reduction $\max_z m^\theta_{z,S}$. Harm is the largest set size increase among initially smaller set cells using tolerance $\epsilon = 0.02$, and ΔCov is the change in coverage.

6.1 EVIDENCE FOR PROPENSITY STANDARDIZATION

The experiments quantify the composition sensitivity of prediction sets by examining the relevant metrics before and after propensity standardization is applied. Reductions in conflict strength reflect better aggregate–stratum alignment, while the ESS fraction records how much comparable support remains. Table 1 shows that standardization reduces conflict strength across the reported settings and usually reduces the number of reversals, revealing a strong composition effect in the observed data. For Bias in Bios and ATUS, conflict strength decreases greatly across all score variants. The Folktables occupation task shows that high conflict strength reversals are nearly eliminated in every year. These reductions reflect Theorem 3 and imply that raw aggregate–stratum disagreement is strongly sensitive to the composition effect, so raw set-size curves should be interpreted accordingly. Table 8 in Appendix C reports the corresponding alignment loss decomposition. Variability across data splits is examined in Table 22 in Appendix F.

6.2 EVIDENCE FOR TILTED MONDRIAN DISPARITY MITIGATION

Table 2 reports the results for the SIMPLE score function at $\alpha = 0.1$. Tilted Mondrian reduces the test set's worst stratum-level gap in Bias in Bios and both Folktables tasks (year 2018) with only small average set-size changes; ATUS is an exception where very little disparity existed initially. Table 15 in Appendix E reports the remaining score functions, where again Tilted Mondrian reduces disparities on the three datasets that initially display significant differences. These reductions are consistent with Theorem 5, which shows that Tilted Mondrian can reduce finer-stratum coverage disparities. Table 23 in Appendix F shows variability across data splits.

Table 1: Metrics before and after propensity standardization is applied. Top: Bias in Bios/ATUS across score functions $s(x, y)$. Bottom: Folktables with SIMPLE.

Bias in Bios			
Score	Reversals	Conflict strength	ESS
SIMPLE	$20 \to 2$	$0.101 \to 0.006$	0.833
APS	$41 \to 7$	$0.120 \to 0.012$	0.833
RAPS	$56 \to 1$	$0.151 \to 0.008$	0.833
SAPS	$38 \to 0$	$0.131 \to 0.005$	0.833

ATUS			
Score	Reversals	Conflict strength	ESS
SIMPLE	$36 \to 10$	$0.464 \to 0.003$	0.982
APS	$72 \to 51$	$1.635 \to 0.125$	0.982
RAPS	$72 \to 54$	$1.228 \to 0.086$	0.982
SAPS	$69 \to 45$	$0.039 \to 0.007$	0.982

Occupation task (SIMPLE)				
Year	Reversals	High conflict	Strength	ESS
2014	182→114	$94 \to 1$	9.378→0.092	0.666
2015	182→160	$87 \to 0$	7.809→0.319	0.611
2016	191→137	$80 \to 0$	7.751→0.111	0.702
2017	192→102	$70 \to 0$	6.822→0.544	0.629
2018	183→101	$69 \to 0$	7.904→0.543	0.621

Income quintile task (SIMPLE)				
Year	Reversals	High conflict	Strength	ESS
2014	182→107	$9 \to 0$	2.059→0.069	0.950
2015	182→156	$30 \to 0$	2.400→0.085	0.954
2016	194→108	$1 \to 0$	2.104→0.102	0.953
2017	182→202	$27 \to 0$	2.675→0.144	0.949
2018	182→144	$18 \to 0$	2.652→0.109	0.949

Table 2: Tilted Mondrian for coverage disparity mitigation.

Experiment	Score	$L^0_{\max} \to L^\theta_{\max}$	$\max_z m^\theta_z$	Harm	$\Delta\lvert C\rvert$
Bias in Bios	SIMPLE	$0.161 \to 0.090$	0.071 (surgeon)	0.010	0.012
Folktables occupation	SIMPLE	$0.037 \to 0.014$	0.024 (Less than high school)	0.010	0.007
Folktables income	SIMPLE	$0.033 \to 0.024$	0.009 (Bachelor)	0.003	0.001
ATUS	SIMPLE	$0.010 \to 0.013$	0.001 (employed)	0.009	0.000

Table 3: Tilted Mondrian for set size disparity mitigation.

Experiment	Score	$L^0_{\max,S} \to L^\theta_{\max,S}$	$\max_z m^\theta_{z,S}$	Harm	$\Delta\mathrm{Cov}(\mathrm{pp})$
Bias in Bios	SIMPLE	$0.065 \to 0.063$	0.005 (psychologist)	0.000	-0.041
Folktables occupation	SIMPLE	$0.180 \to 0.142$	0.058 (High school or GED)	0.020	0.012
Folktables income	SIMPLE	$0.188 \to 0.146$	0.041 (Graduate or professional)	0.003	0.007
ATUS	SIMPLE	$0.240 \to 0.238$	0.002 (not in labor force)	0.002	0.000

Table 3 demonstrates that Tilted Mondrian decreases the worst set size disparity for all four datasets. On Folktables we see relatively large reductions in the worst set size disparity. These reductions empirically support the set-size extension of Tilted Mondrian, showing that the algorithm can mitigate finer-stratum set-size disparities. Table 24 in Appendix G shows variability across data splits.

7 Discussion and Conclusion

Conclusion. This paper studies a coarse-to-fine fairness problem in group-Mondrian CP. Aggregate group disparities can be shaped by population composition. Coarse group coverage guarantees may still conceal large differences within finer strata. Propensity standardization evaluates the composition effect by reweighting the set sizes over a standardized distribution. The resulting change in aggregate–stratum misalignment between the observed and standardized population reveals how strong the composition effect is on the disagreement. More broadly, this assessment captures how population mixture interacts with the disparity structure, information that cannot be easily obtained from base rates alone. Moving from assessment to mitigation, Tilted Mondrian intervenes directly at the score level without requiring model retraining. It provides a generic, modular refinement layer on top of a chosen Mondrian calibration scheme, mitigating fine-grained disparities while preserving groupwise coverage. The same principle applies to finer Mondrian hierarchies. Importantly, Tilted Mondrian does not require dividing the calibration data into increasingly small cells, offering the practitioner flexibility to choose a suitable grouping level of interest and still direct the intervention toward disparities at a finer level. A method-agnostic bound further quantifies the limits of such fine-grained fairness improvement. The Tilted Mondrian framework also extends to set size, mitigating disparities across finer strata while retaining groupwise coverage.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, and clean and reformat data. We have not used generative AI tools to generate synthetic datasets, assist with translation, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, or interpret results. Qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for language editing and presentation to improve readability, identifying relevant literature, formatting references, organizing experimental outputs, preparing the graphical abstract, suggesting experimental parameters, and routine coding assistance such as debugging and code-file organization. We have reviewed all AI-assisted work. AI-assisted code was reviewed and tested, and AI-assisted mathematical and textual content was independently checked by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses existing research datasets and involves no new human-subject data collection or intervention. Demographic attributes are used solely to study prediction-set disparities and methodological behavior. The proposed evaluations and interventions are not intended as normative definitions of fairness for deployment.

REPRODUCIBILITY STATEMENT

The appendices provide complete proofs and detailed experimental protocols, including data splits, parameter selection, prediction methods, standardization, and intervention procedures. Source code and scripts are provided as supplementary material.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Peter C Austin and Elizabeth A Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679, 2015.

Francis Bach. A convex loss function for set prediction with optimal trade-offs between size and conditional coverage. *arXiv preprint arXiv:2512.19142*, 2025.

Konstantina Bairaktari, Jiayun Wu, and Zhiwei Steven Wu. Kandinsky conformal prediction: Beyond class-and covariate-conditional coverage. *arXiv preprint arXiv:2502.17264*, 2025.

Alvaro H Correia, Fabio V Massoli, Christos Louizos, and Arash Behboodi. An information theoretic perspective on conformal prediction. *Advances in Neural Information Processing Systems*, 37:101000–101041, 2024.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2 edition, 2006.

Jesse C. Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. Conformal prediction sets improve human decision making. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Jesse C. Cresswell, Bhargava Kumar, Yi Sui, and Mouloud Belbahri. Conformal prediction sets can cause disparate impact. In *International Conference on Learning Representations*, volume 2025, pp. 19803–19822, 2025.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128, 2019.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

R.M. Fano. *Transmission of Information: A Statistical Theory of Communication*. MIT Press Classics. MIT Press, 1961. ISBN 9780262561693. URL https://books.google.ca/books?id=2PMbtQEACAAJ.

Meir Feder and Neri Merhav. Relations between entropy and error probability. *IEEE Transactions on Information theory*, 40(1):259–266, 1994.

Ziang Gao, Pengqi Liu, Archer Yi Yang, Mouloud Belbahri, Jesse C. Cresswell, and Masoud Asgharian. On the burden of achieving fairness in conformal prediction. *arXiv preprint arXiv:2605.14260*, 2026.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1100–1126, 2025.

Ozgur Guldogan, Neeraj Sarna, Yuanyuan Li, and Michael Berger. Counterfactually fair conformal prediction. *arXiv preprint arXiv:2510.08724*, 2025.

Jaroslav Hájek. Comment on "an essay on the logical foundations of survey sampling by d. basu". *Foundations of statistical inference*, pp. 236, 1971.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. *arXiv preprint arXiv:2310.06430*, 2023.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):243–263, 2014.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *arXiv preprint arXiv:2209.15145*, 2022.

Leslie Kish. Weighting for unequal pi. *Journal of Official Statistics*, 8(2):183, 1992.

Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

Meichen Liu, Lei Ding, Dengdeng Yu, Wulong Liu, Linglong Kong, and Bei Jiang. Conformalized fairness via quantile regression. *Advances in Neural Information Processing Systems*, 35:11561–11572, 2022.

Pengqi Liu, Zijun Yu, Mouloud Belbahri, Arthur Charpentier, Masoud Asgharian, and Jesse C. Cresswell. Beyond procedure: Substantive fairness in conformal prediction. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 12008–12016, 2022.

Vincent Plassier, Alexander Fishkov, Victor Dheur, Mohsen Guizani, Souhaib Ben Taieb, Maxim Panov, and Eric Moulines. Rectifying conformity scores for better conditional coverage. *arXiv preprint arXiv:2502.16336*, 2025.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2(2), 2020a.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020b.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

U.S. Bureau of Labor Statistics. American Time Use Survey—2024 Microdata Files. https://www.bls.gov/tus/data/datafiles-2024.htm, 2025. Released June 26, 2025; accessed June 12, 2026.

Aditya Vadlamani, Anutam Srinivasan, Pranav Maneriker, Ali Payani, and Srinivasan Parthasarathy. A generic framework for conformal fairness. In *International Conference on Learning Representations*, volume 2025, pp. 16451–16483, 2025.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. *Technical Report*, 2003.

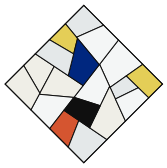
Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Fangxin Wang, Lu Cheng, Ruocheng Guo, Kay Liu, and Philip S Yu. Equal opportunity of coverage in fair regression. *Advances in Neural Information Processing Systems*, 36:7743–7755, 2023.

Ran Xie, Rina F Barber, and Emmanuel J Candès. Boosted conformal prediction intervals. *Advances in Neural Information Processing Systems*, 37:71868–71899, 2024.

Yanfei Zhou and Matteo Sesia. Conformal classification with equalized coverage for adaptively selected groups. *Advances in Neural Information Processing Systems*, 37:108760–108823, 2024.

APPENDIX CONTENTS



- Appendix A: Proofs and technical discussions
- Appendix B: A product form set–error uncertainty principle
- Appendix C: Experiment specific protocols and evaluation details
- Appendix D: Standardization targets and profile choice
- Appendix E: Tilted Mondrian details
- Appendix F: Finite sample robustness
- Appendix G: Tilted Mondrian for set size and fundamental tradeoffs

A PROOFS AND TECHNICAL DISCUSSIONS

A.1 PROOFS OF THE MAIN RESULTS

This appendix records the proofs and technical discussions behind the curve summaries. Throughout the appendix the predictor is fixed. The random variable being re-averaged is the delivered prediction-set size $M_\alpha(X) = |\mathcal{C}_\alpha(X)|$.

Table 4: Notation often used in the main text and appendices. Tildes denote quantities induced by the profile target distribution Q_H; they do not change the delivered prediction sets.

Symbol	Meaning
$X, Y, \mathcal{Y}, K$	Input, label, label space, and number of classes
$s(x, y), \hat{q}_\alpha$	Nonconformity score and split-conformal threshold
$\mathcal{C}_\alpha(x), M_\alpha(X)$	Delivered prediction set and set size
$\boldsymbol{\alpha}$	Grid of nominal miscoverage levels used for curve reporting
$\alpha^\star$	Selected information efficiency anchor
$M_\alpha^+(X)$	Nonempty set-size version used in Proposition 7
$\mathcal{H}, \mathcal{H}_{\text{bin}}$	Entropy and binary entropy
$G \in \{a, b\}, Z, \mathcal{Z}, z$	Comparison group, stratum variable, space, and a realized stratum
$\pi_g(z), \tilde{\pi}_g(z)$	Raw and Q_H-induced stratum distributions
$V_g, V_{g,z}, \tilde{V}_g, \tilde{V}_{g,z}$	Raw and standardized group and stratum curves
$\Delta, \Delta_z, \tilde{\Delta}, \tilde{\Delta}_z$	Raw and standardized aggregate and stratum gaps
$\mathbb{E}_Z, \mathbb{E}_{\tilde{Z}}$	Expectations under the symmetric raw and standardized distributions
$\text{Var}_Z, \text{Var}_{\tilde{Z}}$	Variance under the symmetric raw and standardized distributions
$R, \tilde{R}, \Gamma_z$	Raw residual, standardized residual, and conflict strength
$H, P_H, e(h), Q_H$	Profile, marginal distribution, propensity, and standardized distribution
$w_a(H), w_b(H)$	Standardization factors $1 - e(H)$ and $e(H)$
$\mathcal{A}_{\text{raw}}, \mathcal{A}_{\text{std}}$	Raw and standardized alignment losses
η_H	Average conditional reporting-stratum imbalance under Q_H
$\Delta_{h,z}(\alpha)$	Conditional group set-size gap at profile h and stratum z
$g^+(z), g^-(z)$	Groups with higher and lower baseline coverage in stratum z
θ_z^+, θ_z^-	Nonnegative shifts for the higher/lower coverage groups in stratum z
$S_\theta, \mathcal{C}_{\alpha,\theta}$	Score after intervention and resulting set; $\mathcal{C}_{\alpha,0}$ is the baseline set
$c_{g,z}^0, c_{g,z}^\theta$	Baseline and intervened cell coverage
$d_z, d_z^\theta, m_z^\theta$	Baseline stratum coverage gap, intervened gap, and gap reduction
ϵ	Harm budget for initially advantaged cells
Data splits	$\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{cal}}$, $\mathcal{D}_{\text{prop}}/\mathcal{D}_{\text{tilt}}$, and $\mathcal{D}_{\text{test}}$

Proof of Proposition 2.

Proposition 2 (Propensity score reweighting)**.** *For every integrable function f of H,*

$$\frac{\mathbb{E}\{f(H)\mathbf{1}(G=a)(1-e(H))\}}{\mathbb{E}\{\mathbf{1}(G=a)(1-e(H))\}} = \frac{\mathbb{E}\{f(H)\mathbf{1}(G=b)e(H)\}}{\mathbb{E}\{\mathbf{1}(G=b)e(H)\}} = \mathbb{E}_{Q_H}[f(H)]. \tag{14}$$

Proof. Let $e(h) = \mathbb{P}(G = a \mid H = h)$. For group a,

$$\begin{aligned}&\mathbb{E}\{f(H)\mathbf{1}(G = a)(1 - e(H))\}\\ &\quad = \int f(h)e(h)\{1 - e(h)\}dP_H,\end{aligned} \tag{22}$$

and for group b,

$$\begin{aligned}&\mathbb{E}\{f(H)\mathbf{1}(G = b)e(H)\}\\ &\quad = \int f(h)\{1 - e(h)\}e(h)dP_H.\end{aligned} \tag{23}$$

The normalizing constants are identical. Both normalized weighted distributions are therefore Q_H. □

An equivalent symmetric form of Equation (5), in Section 4, is obtained by adding and subtracting the same cellwise terms. The raw aggregate gap can be written as

$$\begin{aligned}\Delta(\alpha) = &\frac{1}{2}\sum_z\{\pi_a(z) + \pi_b(z)\}\Delta_z(\alpha)\\ &+ \frac{1}{2}\sum_z\{\pi_a(z) - \pi_b(z)\}\{V_{a,z}(\alpha) + V_{b,z}(\alpha)\}.\end{aligned} \tag{24}$$

The first term is $\mathbb{E}_Z\Delta_Z(\alpha)$ under the symmetric raw stratum distribution, and the second term is $R(\alpha)$. The same calculation after standardization gives the corresponding identity for the Q_H-induced stratum distributions and the cell curves defined around Equation (11):

$$\begin{aligned}\tilde{\Delta}(\alpha) = &\frac{1}{2}\sum_z\{\tilde{\pi}_a(z) + \tilde{\pi}_b(z)\}\tilde{\Delta}_z(\alpha)\\ &+ \frac{1}{2}\sum_z\{\tilde{\pi}_a(z) - \tilde{\pi}_b(z)\}\{\tilde{V}_{a,z}(\alpha) + \tilde{V}_{b,z}(\alpha)\},\end{aligned} \tag{25}$$

which is the displayed identity $\tilde{\Delta}(\alpha) = \mathbb{E}_{\tilde{Z}}\tilde{\Delta}_Z(\alpha) + \tilde{R}(\alpha)$; Equation (25) records the residual explicitly. For the residual bound, write $r_z = \tilde{\pi}_a(z) - \tilde{\pi}_b(z)$ and $s_z = \tilde{V}_{a,z}(\alpha) + \tilde{V}_{b,z}(\alpha)$. Since $\sum_z r_z = 0$ and $0 \le s_z \le 2K$,

$$\begin{aligned}|\tilde{R}(\alpha)| &= \frac{1}{2}\left|\sum_z r_z\{s_z - K\}\right|\\ &\le \frac{K}{2}\sum_z |r_z| = K\ \mathrm{TV}(\tilde{\pi}_a, \tilde{\pi}_b).\end{aligned} \tag{26}$$

This is the residual bound used in Section 4. The centering step sharpens the direct range bound and shows that the remaining standardized residual is controlled by the total-variation distance between the two Q_H-induced stratum distributions.

Proof of Theorem 3.

Theorem 3 (Alignment under standardization)**.** *Given a fixed level α, the propensity score standardization satisfies $\mathrm{TV}(\tilde{\pi}_a, \tilde{\pi}_b) \le \eta_H, |\tilde{R}(\alpha)| \le K\eta_H$. Suppose η_H vanishes, equivalently $\mathbb{P}(Z \in \cdot \mid G = a, H = h) = \mathbb{P}(Z \in \cdot \mid G = b, H = h)$ for Q_H-almost every h, then $\tilde{R}(\alpha) = 0$ and*

$$\mathcal{A}_{\mathrm{std}}(\alpha) \le \mathrm{Var}\{\Delta_{H,Z}(\alpha)\}\ . \tag{19}$$

Furthermore, suppose the composition dominance holds (Equation (18)), then

$$\mathcal{A}_{\mathrm{std}}(\alpha) < \mathcal{A}_{\mathrm{raw}}(\alpha)\ . \tag{20}$$

In particular, suppose $\Delta_{h,z}(\alpha)$ is constant for every h, z pair, then $\mathcal{A}_{\mathrm{std}}(\alpha) = 0$.

Proof. Write $p_g(z \mid h) = \mathbb{P}(Z = z \mid G = g, H = h)$. By Equation (11), both propensity reweighted groups use the same distribution Q_H, so

$$\tilde{\pi}_g(z) = \int p_g(z \mid h) dQ_H.$$

The triangle inequality under this common mixture gives

$$\begin{aligned} \mathrm{TV}(\tilde{\pi}_a, \tilde{\pi}_b) &= \frac{1}{2} \sum_z \left| \int \{p_a(z \mid h) - p_b(z \mid h)\} dQ_H \right| \\ &\leq \int \frac{1}{2} \sum_z |p_a(z \mid h) - p_b(z \mid h)| dQ_H = \eta_H. \end{aligned} \tag{27}$$

Combining this with Equation (26) yields

$$|\tilde{R}(\alpha)| \leq K \ \mathrm{TV}(\tilde{\pi}_a, \tilde{\pi}_b) \leq K \eta_H.$$

Now suppose $\eta_H = 0$. Since the integrand defining η_H is nonnegative,

$$p_a(\cdot \mid h) = p_b(\cdot \mid h)$$

for Q_H-almost every h; write the common conditional distribution as $p(\cdot \mid h)$. Hence

$$\tilde{\pi}_a(z) = \tilde{\pi}_b(z) = \int p(z \mid h) dQ_H =: \tilde{\pi}(z),$$

so Equation (25) immediately gives $\tilde{R}(\alpha) = 0$.

For brevity within the proof, let

$$m_g(h, z) := \mathbb{E}\{M_\alpha(X) \mid G = g, H = h, Z = z\},$$

so that $\Delta_{h,z}(\alpha) = m_a(h, z) - m_b(h, z)$. From the definition of $\tilde{V}_{g,z}(\alpha)$ and the common conditional cell distribution,

$$\begin{aligned} \tilde{\Delta}_z(\alpha) &= \tilde{V}_{a,z}(\alpha) - \tilde{V}_{b,z}(\alpha) \\ &= \frac{\int p(z \mid h) \Delta_{h,z}(\alpha) dQ_H}{\tilde{\pi}(z)} \\ &= \mathbb{E}\{\Delta_{H,Z}(\alpha) \mid Z = z\}, \end{aligned} \tag{28}$$

where the expectations and variances are under the joint distribution obtained by first drawing $H \sim Q_H$ and then $Z \sim p(\cdot \mid H)$. Therefore, using Equation (15) and $\tilde{R}(\alpha) = 0$,

$$\begin{aligned} \mathcal{A}_{\mathrm{std}}(\alpha) &= \mathrm{Var}_{\tilde{Z}}\{\tilde{\Delta}_Z(\alpha)\} \\ &= \mathrm{Var}[\mathbb{E}\{\Delta_{H,Z}(\alpha) \mid Z\}]. \end{aligned} \tag{29}$$

The law of total variance now gives

$$\begin{aligned} \mathrm{Var}\{\Delta_{H,Z}(\alpha)\} &= \mathrm{Var}[\mathbb{E}\{\Delta_{H,Z}(\alpha) \mid Z\}] \\ &\quad + \mathbb{E}[\mathrm{Var}\{\Delta_{H,Z}(\alpha) \mid Z\}] \\ &\geq \mathcal{A}_{\mathrm{std}}(\alpha), \end{aligned} \tag{30}$$

which proves $\mathcal{A}_{\mathrm{std}}(\alpha) \leq \mathrm{Var}\{\Delta_{H,Z}(\alpha)\}$.

Now suppose composition dominance holds. By Equation (18),

$$\begin{aligned} \mathrm{Var}\{\Delta_{H,Z}(\alpha)\} &= \mathbb{E}\left[\{\Delta_{H,Z}(\alpha) - \mathbb{E}[\Delta_{H,Z}(\alpha)]\}^2\right] \\ &\leq (\operatorname*{ess\,sup}_{h,z} |\Delta_{h,z}(\alpha) - \mathbb{E}[\Delta_{H,Z}(\alpha)]|)^2 \\ &< R^2(\alpha). \end{aligned} \tag{31}$$

By the definition $R(\alpha) = \Delta(\alpha) - \mathbb{E}_Z \Delta_Z(\alpha)$,

$$\begin{aligned}\mathcal{A}_{\text{raw}}(\alpha) &= \mathbb{E}_Z\left[\{R(\alpha) + \mathbb{E}_Z\Delta_Z(\alpha) - \Delta_Z(\alpha)\}^2\right] \\ &= R^2(\alpha) + \text{Var}_Z\{\Delta_Z(\alpha)\} \\ &\geq R^2(\alpha). \end{aligned} \tag{32}$$

Combining Equations (31) and (32) gives

$$\mathcal{A}_{\text{std}}(\alpha) \leq \text{Var}\{\Delta_{H,Z}(\alpha)\} < R^2(\alpha) \leq \mathcal{A}_{\text{raw}}(\alpha), \tag{33}$$

and hence

$$\mathcal{A}_{\text{std}}(\alpha) < \mathcal{A}_{\text{raw}}(\alpha). \tag{34}$$

If $\Delta_{h,z}(\alpha)$ is constant for every (h, z) pair, then $\text{Var}\{\Delta_{H,Z}(\alpha)\} = 0$ and $\mathcal{A}_{\text{std}}(\alpha) = 0$. □

Proof of Theorem 4.

Theorem 4 (Validity of Tilted Mondrian)**.** *Let θ be selected on a tilt split independent of calibration. If group-Mondrian split conformal calibration is applied to S_θ to generate prediction sets $\mathcal{C}_{\alpha,\theta}(X)$ as in Equation (2), and calibration and test examples are exchangeable conditional on $G = g$, then $\mathbb{P}\{Y \in \mathcal{C}_{\alpha,\theta}(X) \mid G = g\} \geq 1 - \alpha$ for each group $g \in \{a, b\}$.*

Proof. Condition on the tilt split. Then θ is fixed before calibration, so S_θ is an ordinary nonconformity score. Within each group g, the calibration true label scores and the future true label score are exchangeable conditional on $G = g$. The standard group Mondrian split conformal rank argument gives $\mathbb{P}\{S_\theta(X, Y) \leq \tilde{q}_g \mid G = g\} \geq 1 - \alpha$, where $\tilde{q}_g$ is the group specific conformal quantile computed on the calibration split after score shifting. Since $Y \in \mathcal{C}_{\alpha,\theta}(X)$ exactly when $S_\theta(X, Y) \leq \tilde{q}_G$, the claim follows. □

Proof of Theorem 5.

Theorem 5 (Local mitigation)**.** *Assume $g = g^-(z_i) = g^+(z_j)$ for supported z_i, z_j with $d_{z_i}, d_{z_j} > 0$ and $\epsilon > 0$. Suppose the population cellwise nonconformity score distributions admit densities that are locally positive at the threshold q_g^0. Then the Tilted Mondrian family admits a nonzero feasible tilt θ preserving group coverage such that $d_{z_i}^\theta < d_{z_i}$ and $d_{z_j}^\theta < d_{z_j}$, equivalently $m_{z_i}^\theta > 0$ and $m_{z_j}^\theta > 0$, with all other stratum gaps unchanged.*

Proof. Let $F_{g,z}$ be the baseline true-label score CDF, q_g^0 the baseline group threshold, and write $\pi_i = \pi_g(z_i)$ and $\pi_j = \pi_g(z_j)$. Since both cells are supported, $\pi_i, \pi_j > 0$. Choose $\eta > 0$ small enough that

$$\delta_i = \eta\pi_j < d_{z_i}, \qquad \delta_j = \eta\pi_i < \min\{d_{z_j}, \epsilon\}.$$

Local positivity gives $\theta_i, \theta_j > 0$ such that lowering scores in (g, z_i) by θ_i increases its coverage by δ_i, while raising scores in (g, z_j) by θ_j decreases its coverage by δ_j; set all other shifts to zero. Since

$$\pi_i\delta_i = \pi_j\delta_j,$$

group coverage is unchanged. Hence $d_{z_i}^\theta = d_{z_i} - \delta_i$ and $d_{z_j}^\theta = d_{z_j} - \delta_j$, with all other gaps unchanged. The chosen bounds prevent gap reversal and satisfy the harm constraint, proving feasibility. □

Proof of Theorem 6.

Theorem 6 (Mitigation bound)**.** *Consider a class of disparity mitigation methods satisfying the same constraints as Tilted Mondrian: they preserve each group's baseline coverage, lose no more than ϵ coverage for each advantaged cell, do not reduce coverage in any disadvantaged cell, and do not increase stratum gaps. Assume $\pi_{g^-(z)}(z) > 0$ for every $z \in \mathcal{Z}$. Then*

$$\max_{z\in\mathcal{Z}} m_z^\theta \leq \max_{z\in\mathcal{Z}} \min\left\{d_z, \epsilon\left(1 + \frac{\sum_{r:g^+(r)=g^-(z)} \pi_{g^-(z)}(r)}{\pi_{g^-(z)}(z)}\right)\right\}. \tag{21}$$

Proof. Recall that $\pi_g(z) = \mathbb{P}(Z = z \mid G = g)$, and that $\mathcal{Z}$ partitions each group population. For each stratum z, let

$$h_z = c^0_{g^+(z),z} - c^\theta_{g^+(z),z}, \qquad u_z = c^\theta_{g^-(z),z} - c^0_{g^-(z),z}.$$

Thus h_z is the coverage loss on the initially advantaged side and u_z is the coverage gain on the initially disadvantaged side. Since $d^\theta_z \geq 0$, the single-cell mitigation satisfies $m^\theta_z = d_z - d^\theta_z \leq d_z$. The no-loss condition on disadvantaged cells gives $u_z \geq 0$. The no-gap-increase condition gives $0 \leq h_z + u_z \leq 2d_z$, because the intervened signed gap is $d_z - h_z - u_z$. Hence

$$m^\theta_z = d_z - |d_z - h_z - u_z| \leq h_z + u_z.$$

The harm constraint gives $h_z \leq \epsilon$. Fixed group coverage gives, for every group g,

$$\sum_{r:g^-(r)=g} \pi_g(r)u_r = \sum_{r:g^+(r)=g} \pi_g(r)h_r \leq \epsilon \sum_{r:g^+(r)=g} \pi_g(r).$$

Because all disadvantaged-cell gains are nonnegative, every stratum z satisfies

$$u_z \leq \frac{\epsilon \sum_{r:g^+(r)=g^-(z)} \pi_{g^-(z)}(r)}{\pi_{g^-(z)}(z)}.$$

Combining this with $h_z \leq \epsilon$ gives

$$m^\theta_z \leq h_z + u_z \leq \epsilon \left(1 + \frac{\sum_{r:g^+(r)=g^-(z)} \pi_{g^-(z)}(r)}{\pi_{g^-(z)}(z)}\right).$$

Together with $m^\theta_z \leq d_z$, this yields the displayed cellwise bound. Taking the maximum over strata proves the result. □

A.2 Finite-sample reporting stability and technical discussions

Limitations. Standardization requires adequate overlap and a profile that reflects the relevant composition; weak support, omitted information, task heterogeneity, or a weak classifier can result in lingering disagreement. The current framework is stated for binary group contrasts, and a multigroup formulation remains future work. Our Tilted Mondrian framework is interpretable in a fairness context but relies on a finite heuristic approach. The tilt split also trades the calibration sample size for score adjustment selection and calibration. Future work may focus on more efficient constrained optimization for establishing fine-grained fairness with explicit cell-level costs and trade-offs. Evaluation across additional real-world settings would further assess the framework's generality.

As noted in Section 6, the normalized Effective Sample Size (ESS) serves as a diagnostic for our standardization. The behavior of the ESS exposes a limit of identifiability. If profile features strongly separate the demographic groups, overlap weights vanish for the majority of samples, causing the ESS to drop. This warns practitioners that the standardized gap is driven by a thin subpopulation. In the extreme case where profile features perfectly proxy group membership (i.e., $e(h) \in \{0, 1\}$), all overlap weights become zero, rendering the standardized disparity mathematically undefined.

Fixed-policy reporting. All reporting identities are applied after a thresholding policy has already been chosen and the prediction sets have been formed. The policy may be pooled or group-conditional. The raw and standardized curves then re-average the same delivered set sizes under different reporting populations. Thus the algebra below does not require the same thresholding policy in every experiment; it requires that, within each experiment, the thresholding policy be fixed before reporting weights and curve summaries are computed.

Split-conformal rank argument. After the score rule is specified, the calibration scores $U_i = s(X_i, Y_i)$ and the test score $U_{n+1} = s(X, Y)$ are exchangeable. With the usual conservative split-conformal order statistic, the event $Y \notin \mathcal{C}_\alpha(X)$ is the event that U_{n+1} exceeds the selected calibration threshold. Exchangeability implies that this event has probability at most α, up to the standard finite-sample conservativeness from ties and order-statistic rounding. The groupwise statement applies the same rank argument after conditioning on $G = g$ and using only calibration examples from that group.

Finite-grid Hájek stability. For group g, the population standardized curve is

$$\tilde{V}_g(\alpha) = \frac{\mathbb{E}\{\mathbf{1}(G=g)w_g(H)M_\alpha(X)\}}{\mathbb{E}\{\mathbf{1}(G=g)w_g(H)\}}, \tag{35}$$

where $w_a(H) = 1-e(H)$ and $w_b(H) = e(H)$. On the test split, the corresponding stabilized Hájek report is the finite-sample ratio in Equation (54), denoted by $\tilde{V}'_g(\alpha)$. Cellwise reports use the same ratio after multiplying numerator and denominator by $\mathbf{1}(Z_i = z)$. This normalized ratio form is the classical Hájek style estimator (Hájek, 1971); the bounded weights used here follow the propensity reweighting construction of Li et al. (2018).

Suppose the finite-sample propensity standardized factors are bounded, the corresponding population denominator is bounded away from zero, and the operating grid Λ is finite. Since $0 \le M_\alpha(X_i) \le K$, concentration for bounded averages (Hoeffding, 1963) gives uniform convergence of the numerator and denominator over $\alpha \in \Lambda$. The ratio map is Lipschitz on denominators bounded away from zero, so

$$\sup_{\alpha\in\Lambda} \left|\tilde{V}'_g(\alpha) - \tilde{V}_g(\alpha)\right| = o_p(1) \tag{36}$$

when the fitted weights converge to the population standardized weights and the stabilized denominator remains nondegenerate. The same argument applies to cellwise reports with positive weighted denominators. This is a stability statement for the reported curve estimates rather than a coverage argument.

Plug-in propensity estimates. The standardized target in the population is written in terms of $e(H) = \mathbb{P}(G = a \mid H)$, whereas the finite-sample reports use a fitted propensity $e'(H)$ learned on the propensity split and then held fixed on the test split. This is the standard plug-in use of estimated propensity scores in weighting estimators (Hirano et al., 2003). In our setting the claim is descriptive rather than causal: the fitted propensity defines the standardized report used to re-average already delivered prediction-set sizes. If the stabilized weight is a Lipschitz function of the fitted propensity on the clipped range and $\mathbb{E}|e'(H) - e(H)| = r_n$, then the population weighted curve formed with e' differs from the ideal standardized curve by at most a constant multiple of Kr_n, provided the group denominators are bounded away from zero. This follows by writing the difference between two ratio functionals, bounding the numerator and denominator perturbations by the weight perturbation, and using the boundedness of M_α. The fixed bounded-weight rule and Hájek normalization make this perturbation controlled in finite samples, while balance checks assess whether the learned weighted report is stable (Austin & Stuart, 2015).

Reporting cell interpretation. Propensity standardization aligns the groups in the coordinates of the chosen prediction profile H (Rosenbaum & Rubin, 1983; Li et al., 2018). Under the shared profile distribution, reporting-cell balance is governed by the within-profile distributions $\mathbb{P}(Z \in \cdot \mid G, H)$. Theorem 3 summarizes their remaining discrepancy through η_H and its composition effect through $\tilde{R}$. A standardized cell curve applies the learned reweighting within each test stratum $Z = z$. This design keeps H as the frozen standardization profile and Z as the reporting coordinate for cellwise Simpson-type disagreement. Exact induced distribution equality follows under $Z \perp G \mid H$. We therefore report both profile balance and residual reporting-cell mixture summaries.

Post-conformal reporting note. The delivered set $\mathcal{C}_\alpha(x)$ is determined by the specified score and the split conformal calibration quantile. Under the separated reporting protocol in Section 6, the profile map, propensity fit, stabilization rule, and Hájek averages are applied after these sets have been fixed. The score, order statistic, and membership rule in Equation (2) remain unchanged, so the event $Y \in \mathcal{C}_\alpha(X)$ is identical before and after reporting standardization.

B A Product Form Set–Error Uncertainty Principle

We first give the full Fano argument (Correia et al., 2024) underlying the intrinsic ambiguity scale presented in Section 3. Throughout, all logarithms are natural. Write $\mathcal{H}(Y \mid X)$ for the conditional entropy of Y given X, let $\mathcal{H}_{\text{bin}}(u) = -u\log u - (1-u)\log(1-u)$, and denote $[t]_+ = \max(t, 0)$.

Proposition 7 (Fano set-size floor). *For the split CP sets in Equation (2). Let $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ and the fresh test point (X, Y) be i.i.d. draws from the same distribution with the nonempty set size convention $M_\alpha^+(X) = M_\alpha(X) \vee 1$. If $\alpha < 1/2$, then*

$$\mathbb{E}M_\alpha^+(X) \geq \exp\Big\{ \big[\mathcal{H}(Y \mid X) - \mathcal{H}_{\text{bin}}(\alpha) - \alpha \log(K-1)\big]_+ \Big\}. \tag{37}$$

Proof. Let $\mathcal{C}_\alpha^+(X)$ be the nonempty version of the set and let $E = \mathbf{1}\{Y \notin \mathcal{C}_\alpha^+(X)\}$. Because the nonempty replacement only adds a label when the original set is empty, split-conformal validity gives $\mathbb{P}(E=1) \leq \alpha$. Since $D_{\text{cal}} \perp (X, Y)$ and E is determined by (X, Y, D_{cal}),

$$\begin{aligned} \mathcal{H}(Y \mid X) &= \mathcal{H}(Y \mid X, D_{cal}) \\ &= \mathcal{H}(E \mid X, D_{cal}) + \mathcal{H}(Y \mid X, D_{cal}, E) \\ &\leq \mathcal{H}_{\text{bin}}\big(\mathbb{P}(E=1)\big) + \mathbb{E}\big[(1-E)\log M_\alpha^+(X)\big] + \mathbb{P}(E=1)\log(K-1) \\ &\leq \mathcal{H}_{\text{bin}}\big(\mathbb{P}(E=1)\big) + \mathbb{E}\log M_\alpha^+(X) + \mathbb{P}(E=1)\log(K-1). \end{aligned} \tag{38}$$

Since $\mathbb{P}(E=1) \leq \alpha < 1/2$,

$$\mathcal{H}(Y \mid X) \leq \mathcal{H}_{\text{bin}}(\alpha) + \mathbb{E}\log M_\alpha^+(X) + \alpha \log(K-1). \tag{39}$$

Because $M_\alpha^+(X) \geq 1$, the positive part may be taken. Jensen's inequality then gives $\mathbb{E}M_\alpha^+(X) \geq \exp\{\mathbb{E}\log M_\alpha^+(X)\}$, which yields Equation (37). □

Take-away. CP sets cannot be arbitrarily small. When the label Y remains hard to predict from the input X, the procedure must use larger sets on average.

Price of increasing coverage. Write

$$C := \exp\{\mathcal{H}(Y \mid X)\}$$

and define

$$L_K(\alpha) := C \exp\{\alpha \log(\alpha) + (1-\alpha)\log(1-\alpha) - \alpha \log(K-1)\}. \tag{40}$$

For $\alpha < 1/2$, Proposition 7 gives $\mathbb{E}[M_\alpha^+(X)] \geq L_K(\alpha)$ whenever this is the nontrivial branch of the floor. Substituting nominal coverage $\gamma = 1 - \alpha \in (1/2, 1)$, define

$$G_K(\gamma) := L_K(1-\gamma). \tag{41}$$

Thus $G_K(\gamma)$ is the information-theoretic lower bound at nominal coverage γ. For $K \geq 2$ and $1/2 < \gamma < 1$,

$$G_K'(\gamma) = G_K(\gamma)\left\{\log\left(\frac{\gamma}{1-\gamma}\right) + \log(K-1)\right\} > 0, \tag{42}$$

and

$$G_K''(\gamma) = G_K(\gamma)\left[\left\{\log\left(\frac{(K-1)\gamma}{1-\gamma}\right)\right\}^2 + \frac{1}{\gamma(1-\gamma)}\right] > 0. \tag{43}$$

Therefore, $G_K(\gamma)$ is strictly increasing and strictly convex in nominal coverage. Increasing nominal coverage raises the lower bound set size burden, and this burden rises progressively faster as coverage approaches one.

The first derivative measures the instantaneous increase in the lower bound per unit increase in nominal coverage. Define the proportional price

$$r_K(\gamma) := \frac{G_K'(\gamma)}{G_K(\gamma)} = \log\left(\frac{(K-1)\gamma}{1-\gamma}\right). \tag{44}$$

For a small coverage increase $d\gamma > 0$,

$$\frac{G_K(\gamma + d\gamma) - G_K(\gamma)}{G_K(\gamma)} \approx r_K(\gamma) d\gamma. \tag{45}$$

Hence a one percentage point coverage increase, $d\gamma = 0.01$, corresponds approximately to a $r_K(\gamma)\%$ increase in the lower-bound burden. Moreover,

$$r_K'(\gamma) = \frac{1}{\gamma(1-\gamma)} > 0, \qquad \lim_{\gamma \to 1} r_K(\gamma) = \infty. \tag{46}$$

Thus the proportional lower-bound price becomes arbitrarily large as nominal coverage approaches one. For example, the same one-percentage-point increase in coverage is cheaper (in terms of the set-size lower bound) when moving from 80% to 81% than when moving from 95% to 96%.

Finite coverage increments. Let $0 < \delta < \alpha < 1/2$. Define the relative lower-bound inflation associated with increasing nominal coverage by δ as

$$R_{K,\delta}(\alpha) = \frac{L_K(\alpha - \delta) - L_K(\alpha)}{L_K(\alpha)}. \tag{47}$$

For fixed K and δ, $R_{K,\delta}(\alpha)$ is strictly decreasing in α. Suppose a practitioner wants to increase nominal coverage by δ, equivalently decreasing the error rate from α to $\alpha - \delta$, and is willing to tolerate at most a proportional lower-bound inflation $r > 0$. The change is within tolerance exactly when

$$\begin{aligned} R_{K,\delta}(\alpha) \le r &\iff \frac{L_K(\alpha - \delta)}{L_K(\alpha)} \le r + 1 \\ &\iff \log L_K(\alpha - \delta) - \log L_K(\alpha) \le \log(r + 1). \end{aligned} \tag{48}$$

Because $R_{K,\delta}(\alpha)$ is strictly decreasing in α, there is at most one interior cutoff α^* satisfying

$$\log L_K(\alpha^* - \delta) - \log L_K(\alpha^*) = \log(r + 1). \tag{49}$$

Whenever such a cutoff exists,

$$\alpha \ge \alpha^* \implies R_{K,\delta}(\alpha) \le r, \qquad \alpha < \alpha^* \implies R_{K,\delta}(\alpha) > r. \tag{50}$$

Now, the question is when does the cutoff exist? For a fixed δ, the possible current error rates satisfy $\delta < \alpha < 1/2$. Since $R_{K,\delta}(\alpha)$ is strictly decreasing in α, its smallest possible value is

$$R_{\min} = \lim_{\alpha \to 1/2} R_{K,\delta}(\alpha) = \frac{L_K(1/2 - \delta)}{L_K(1/2)} - 1,$$

and its largest possible value is

$$R_{\max} = \lim_{\alpha \to \delta} R_{K,\delta}(\alpha) = \frac{L_K(0)}{L_K(\delta)} - 1 = \exp\{\mathcal{H}_{\text{bin}}(\delta) + \delta \log(K - 1)\} - 1.$$

Therefore,

- If $r \le R_{\min}$, the desired coverage gain is never affordable within $0 < \alpha < 1/2$.
- If $r \ge R_{\max}$, it is affordable for every admissible starting α.
- If $R_{\min} < r < R_{\max}$, there is a unique cutoff α^*. This cutoff can be calculated by one-dimensional root finding for Equation (49).

The above computation illustrates an important operating rule. It provides a useful guide on coverage that one hopes to attain and inflation in set size that one expects to tolerate. For a proposed coverage increase δ and a tolerated proportional inflation r in the set-size lower bound, one can solve Equation (49) once and compare the current error level α with α^*. If $\alpha \ge \alpha^*$, the proposed coverage increase remains within the tolerance; if $\alpha < \alpha^*$, the same coverage increase violates that tolerance, indicating that one must either request a smaller coverage increment or accept a larger set-size inflation.

The expected set size curve records the arithmetic burden delivered to a user. Multiplying the floor in Equation (37) by the corresponding error channel factor gives

$$\underbrace{\mathbb{E}M_\alpha^+(X)}_{\substack{\text{expected set-size} \\ \text{burden}}} \times \underbrace{\exp\{\mathcal{H}_{\text{bin}}(\alpha)\}\,(K-1)^\alpha}_{\substack{\text{effective error-channel} \\ \text{uncertainty}}} \ge \underbrace{\exp\{\mathcal{H}(Y \mid X)\}}_{\substack{\text{intrinsic label perplexity}}}. \tag{51}$$

The two factors represent uncertainty retained for downstream selection and uncertainty delegated to tolerated misses. Thus they cannot be simultaneously compressed below intrinsic label perplexity. In this product form sense, Equation (51) has a trade-off interpretation. That is, reducing the burden carried by one channel requires the other to represent the remainder.

Proposition 8 (Information efficiency anchor)**.** *For the split CP family in Equations (1) and (2), suppose that $\alpha < 1/2$ for every $\alpha \in \boldsymbol{\alpha}$, and define*

$$\alpha^\star \in \arg\min_{\alpha \in \boldsymbol{\alpha}} \left\{\mathbb{E}M_\alpha^+(X) \exp\{\mathcal{H}_{\text{bin}}(\alpha)\}\,(K-1)^\alpha\right\}. \tag{52}$$

Then $\alpha^\star$ is the valid grid level with the smallest multiplicative excess above intrinsic label perplexity, and hence comes closest to saturating Equation (51).

Proof. The entropy floor in Equation (51) is constant across the finite grid $\boldsymbol{\alpha}$. Hence minimizing the product in Equation (52) also minimizes its multiplicative excess above that floor. □

Take-away. Selecting coverage levels trades smaller prediction sets against more tolerated errors. The proposition identifies the level that comes closest to the best possible balance between the two.

Operational implications. In a shortlist-and-fallback system, the set channel retains candidates for human or algorithmic selection, whereas tolerated misses are delegated to escalation, additional measurement, or exception handling. Therefore, the anchor $\alpha^\star$ provides a reference for the least redundant allocation between these uncertainty channels.

The principle also identifies when set valued output is necessary. If the error channel is smaller than $\exp\{\mathcal{H}(Y \mid X)\}$, then Equation (51) forces $\mathbb{P}\{M_\alpha^+(X) > 1\} > 0$. That is, a singleton output cannot represent the intrinsic uncertainty. In addition, the minimized product can also compare procedures, with a smaller excess indicating less redundant uncertainty representation. These uses do not make $\alpha^\star$ a universal deployment optimum, because cost of misses, review, and escalation remain application dependent. For a broader information theoretic perspective on conformal prediction, see Correia et al. (2024).

C Experiment Protocols and Evaluation Details

This appendix collects experimental details and evaluation summaries for the empirical studies in Sections 6 and 5. The experiments use Bias in Bios (De-Arteaga et al., 2019), Folktables/ACS (Ding et al., 2021), and ATUS (U.S. Bureau of Labor Statistics, 2025). In the standardization table, an arrow denotes raw to standardized re-averaging on the same test prediction sets. The score, conformal threshold, and prediction set membership rule are fixed before reporting; reporting weights enter only the curve averages on the test split.

For a supported grid point (z, α) on the test split, a reversal is recorded when the aggregate and stratum gaps have opposite signs. Its raw conflict strength is $\Gamma'_z(\alpha) = [-\Delta'(\alpha)\Delta'_z(\alpha)]_+$, and the standardized version uses the gaps analogously. Total conflict strength sums this quantity over the supported test grid. High conflict strength is a supported reversal with $\Gamma'_z(\alpha) > 0.020$. This threshold is descriptive only; the continuous conflict strength remains the main magnitude summary. Effective sample size is the Kish ESS of all test reporting weights divided by the total test sample size. Balance is reported in the coordinates of the fixed profile H and, when relevant, in the reporting stratum mixture.

Finite sample standardization estimates use stabilized Hájek averages. On $\mathcal{D}_{\text{test}}$, let $e'_i = e'(H_i)$ be the group propensity fitted on $\mathcal{D}_{\text{prop}}$ and then treated as fixed. The standardized factors are $r'_{a,i} = 1 - e'_i$ for group a and $r'_{b,i} = e'_i$ for group b. After the prespecified bounded-weight stabilization, the within group weights are

$$w'_{g,i} = \frac{r'_{g,i}}{\sum_{j \in \mathcal{D}_{\text{test}}} \mathbf{1}(G_j = g) r'_{g,j}}. \tag{53}$$

The aggregate standardized curve is estimated by

$$\tilde{V}'_g(\alpha) = \sum_{i \in \mathcal{D}_{\text{test}}} \mathbf{1}(G_i = g) w'_{g,i} M_\alpha(X_i). \tag{54}$$

Cellwise curves use the same standardized factors with the Hájek ratio restricted to $Z_i = z$.

Conformal score construction. Let $\hat{p}_y(x)$ denote the frozen fitted probability for candidate label y, let $\hat{p}_{(1)}(x) \geq \cdots \geq \hat{p}_{(K)}(x)$ be the sorted class probabilities, and let $r_y(x)$ be the rank of y. For randomized adaptive scores, let $U \sim \text{Unif}(0, 1)$ denote the auxiliary randomization, treated as part

Algorithm 3 Propensity Standardization for Disparity Assessment (detailed version of Algorithm 1)

Input: Disjoint splits $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{prop}}, \mathcal{D}_{\text{cal}}, \mathcal{D}_{\text{test}}$; group G, reporting stratum Z, profile rule H, score policy, thresholding policy, operating grid $\boldsymbol{\alpha}$, and supported-cell rule.

1: Fit classifier f on $\mathcal{D}_{\text{train}}$; freeze f, the score policy, and the induced class probabilities.
2: Construct H from the frozen outputs. On $\mathcal{D}_{\text{prop}}$, fit $e'(h) = \widehat{\mathbb{P}}(G = a \mid H = h)$; freeze the propensity, bounded-weight stabilization, and supported-cell rules.
3: On $\mathcal{D}_{\text{cal}}$, obtain the threshold(s) prescribed by the fixed thresholding policy for every $\alpha \in \boldsymbol{\alpha}$.
4: On $\mathcal{D}_{\text{test}}$, compute H_i, $e'_i = e'(H_i)$, and standardized factors $r'_{a,i} = 1 - e'_i$, $r'_{b,i} = e'_i$; apply the frozen stabilization and Hájek-normalize within group to obtain $w'_{g,i}$.
5: **for** $\alpha \in \boldsymbol{\alpha}$ **do**
6: Construct each $\mathcal{C}_{\alpha,i}$ once under the fixed thresholding policy; set $M_{\alpha,i} = |\mathcal{C}_{\alpha,i}|$.
7: Compute raw group and cell curves $V'_g(\alpha), V'_{g,z}(\alpha)$ by unweighted means on the frozen supported grid.
8: Compute $\tilde{V}'_g(\alpha)$ by Equation (54) and $\tilde{V}'_{g,z}(\alpha)$ by the same Hájek ratio restricted to $Z_i = z$.
9: Form raw and standardized aggregate/stratum gaps; record sign reversals and product strengths by Equation (6).
10: **end for**
11: Record ESS fractions and profile/cell-mixture balance for the frozen reporting weights.

Output: Raw and standardized group/cell curves; reversal counts; conflict strengths; effective sample size; balance checks.

Invariant: Both reports use the same test values $M_\alpha(X_i)$ and the same supported (z, α) grid. Reweighting changes neither f, the score, the calibrated threshold(s), nor $\mathcal{C}_\alpha(X_i)$.

Table 5: Exact split sizes for the empirical reports. The split columns are training, calibration, propensity/tilt, and testing.

Setting	Year(s)	Train	Cal.	Prop./Tilt	Test	Total
Bias in Bios	all	127,023	79,390	79,390	31,756	317,559
Folktables occupation	2014	163,527	102,204	102,205	40,882	408,818
Folktables occupation	2015	162,130	101,331	101,332	40,533	405,326
Folktables occupation	2016–2018 each	164,195	102,622	102,622	41,049	410,488
Folktables income quintile	2014–2018 each	164,195	102,622	102,622	41,049	410,488
ATUS	all	71,356	44,683	44,655	17,783	178,477

of the frozen score rule. All scores follow the convention that smaller values are more conforming:

$$s_{\text{SIMPLE}}(x, y) = 1 - \hat{p}_y(x), \tag{55}$$

$$s_{\text{APS}}(x, y; U) = \sum_{j < r_y(x)} \hat{p}_{(j)}(x) + U\hat{p}_y(x), \tag{56}$$

$$s_{\text{RAPS}}(x, y; U) = s_{\text{APS}}(x, y; U) + \lambda\{r_y(x) - k_{\text{reg}}\}_+, \tag{57}$$

$$s_{\text{SAPS}}(x, y; U) = \begin{cases} U\hat{p}_{(1)}(x), & r_y(x) = 1, \\ \hat{p}_{(1)}(x) + \lambda\{r_y(x) - 2 + U\}, & r_y(x) > 1. \end{cases} \tag{58}$$

The experiment-specific score parameters and thresholding policies are listed in Table 6. In every experiment, the score rule is fixed before split conformal calibration, and reporting weights enter neither the score nor the calibration quantile.

Prediction profile construction. For Bias in Bios and the two Folktables tasks, H is built from the frozen class-probability vector. For a K-class task, the profile contains one-hot indicators for the top first, second, and third predicted labels, the three largest probabilities $p_{(1)}, p_{(2)}, p_{(3)}$, normalized entropy, the two margins $p_{(1)} - p_{(2)}$ and $p_{(2)} - p_{(3)}$, and the top-three mass. Its dimension is therefore $3K + 7$, giving 37 features for $K = 10$ and 22 features for $K = 5$. ATUS uses four probability features because the final closed-world task has the top-10 activity vocabulary but the implemented profile rule omits top-label indicators. Across all experiments, H excludes the true group, true label,

Table 6: Score policies, operating grids, and support rules. All raw-to-standardized arrows use the same test prediction sets.

Setting	Threshold policy	Scores	α grid	Support rule
Bias in Bios	Gender-conditional thresholds	SIMPLE; APS; SAPS($\lambda = 0.2$); RAPS($\lambda = 0.2$, $k_{\text{reg}} = 5$)	0.020 to 0.100; step 0.005	At least 200 test observations per Gender group within each occupation stratum; the same supported grid is used for raw and standardized summaries.
Folktables occupation	Pooled threshold within year	SIMPLE	0.010 to 0.100; step 0.001	At least 200 test observations per group within each education stratum.
Folktables income quintile	Pooled threshold within year	SIMPLE	0.010 to 0.100; step 0.001	Same test education stratum support rule as Folktables occupation.
ATUS	Pooled threshold	SIMPLE primary; APS, RAPS($\lambda = 0.01$, $k_{\text{reg}} = 5$), and SAPS($\lambda = 0.2$) checks	0.020 to 0.250; step 0.010	Prespecified employment stratum support rule.

reporting stratum, realized set size, conformal thresholds, alpha-specific membership indicators, and final curve summaries.

Empirical alignment loss decomposition. Table 8 reports the realized component changes. Squared residuals decrease by 59.7–99.3% in every setting, and stratum-gap variance decreases by 55.0–92.6% for Bias in Bios and Folktables. ATUS has increased stratum-gap variance, while its composition reduction yields a 36.1–57.0% net alignment-loss decrease.

Table 8: Relative reductions (%) in grid-averaged alignment loss and its two components. Ranges are across scores or years; negative values denote increases.

Dataset	Alignment loss	Squared composition residual R^2	Stratum-gap variance
Bias in Bios	66.5–72.6	98.7–99.3	60.7–69.9
Folktables occupation	85.1–92.8	75.0–94.2	85.9–92.6
Folktables income	60.0–72.5	88.8–92.2	55.0–69.0
ATUS	36.1–57.0	59.7–60.9	−88.8–−5.9

C.1 Bias in Bios text classification

The text experiment uses the Hugging Face dataset `LabHC/bias_in_bios` to predict one of the top 10 occupations: attorney, dentist, journalist, nurse, photographer, physician, professor, psychologist, surgeon, and teacher. The comparison groups are Male and Female, and occupation is the reporting stratum for stratum curve summaries. The true occupation label is the prediction target and is used after prediction to define reporting cells. It is not included in the reporting profile used for standardization, and no realized set size, calibration threshold, or reversal indicator enters the profile.

The DistilBERT classifier is initialized from `distilbert-base-uncased` (Sanh et al., 2019). Bias in Bios uses groupwise conformal thresholds for the delivered sets and reports SIMPLE, APS (Romano et al., 2020b), SAPS (Huang et al., 2023), and RAPS (Angelopoulos et al., 2021) on the

Table 7: Prediction profiles and propensity models for the standardized report. The reporting weight model enters only the curve summaries on the test split.

Setting	Predictor and score	Profile H	Propensity model
Bias in Bios	DistilBERT top-10 occupation classifier; SIMPLE, APS, SAPS, and RAPS.	37-dimensional frozen probability profile.	Feed-forward neural-network propensity model with hidden layers $(128, 64)$ and ReLU activations; fitted on the propensity split.
Folktables occupation	Histogram gradient boosting on ACS person records, one model per year; SIMPLE.	37-dimensional frozen softmax profile; Gender and education are excluded from H.	Logistic propensity model with balanced class weights; fitted separately by year on the propensity split.
Folktables income quintile	Histogram gradient boosting for income quintiles, one model per year; SIMPLE.	22-dimensional frozen softmax profile; race and education are excluded from H.	Logistic propensity model with balanced class weights; fitted separately by year on the propensity split.
ATUS	Regularized histogram gradient boosting on diary, time, calendar, demographic, region, and household features; SIMPLE with APS, RAPS, and SAPS checks.	Four probability features: p_{top1}, p_{top2}, margin, and entropy.	Logistic propensity model fitted on the propensity split.

Notes. In the logistic rows, standardized means that the coordinates of H are centered and scaled before logistic fitting. All standardized reports use prespecified bounded weights and Hájek normalization. Exact training hyperparameters for all base and propensity models, including learning rates, regularization, batch sizes, epochs or iterations, and random seeds, are provided in the supplementary code.

Table 9: Bias in Bios balance evaluations for the standardized report on the test split.

Measure	Value
Effective sample size fraction	0.833
Profile mean standardized mean difference	$0.0924 \to 0.0133$
Profile maximum standardized mean difference	$0.4331 \to 0.0458$

grid in Table 6. The profile used for standardization consists of fixed probability-profile summaries derived from the classifier output. The same reporting profile and fit in Table 7 are used across score variants, so the score specific entries in Table 1 compare the same standardization design under different set construction rules.

All four Bias in Bios score variants follow the same test inclusion rule. We retain an occupation only when both Gender groups contain at least 200 test observations. For each retained occupation, we evaluate every α value on the prespecified grid, using exactly the same occupation–α pairs for the raw and standardized summaries. A reversal is counted when $\Delta(\alpha)\Delta_z(\alpha) < 0$, $|\Delta(\alpha)| \geq 0.01$, and $|\Delta_z(\alpha)| \geq 0.02$. The conflict strength remains $[-\Delta(\alpha)\Delta_z(\alpha)]_+$, summed over the same supported occupation–α pairs.

C.2 Folktables occupation prediction

The Folktables occupation task uses ACS person records from 2014 through 2018. Each year is treated as a separate problem. The prediction target is top 10 occupation, the comparison group is Gender, and the reporting stratum is education bucket. The classifier is a histogram gradient boosting model and the score is SIMPLE. Education can be a covariate for the classifier, but it is excluded

Table 10: Folktables occupation evaluation summary across 2014–2018.

Measure	Raw	Standardized
Reversals, summed	930	614
High conflict strength reversals	400	1
Conflict strength	39.664	1.609
Years with reversal count drop		5/5
Years with strength drop		5/5
Median cell total variation	0.124	0.045
Median maximum standardized mean difference	0.783	0.355
Effective sample size fraction, range		0.611–0.702
Effective sample size fraction, median		0.629

Table 11: Folktables income quintile evaluation summary across 2014–2018.

Measure	Raw	Standardized
Reversals, summed	922	717
High conflict strength reversals	85	0
Conflict strength	11.890	0.509
Years with reversal count drop		4/5
Years with strength drop		5/5
Median cell total variation	0.120	0.034
Median maximum standardized mean difference	0.290	0.037
Effective sample size fraction, range		0.949–0.954
Effective sample size fraction, median		0.950

from the profile H used for reporting standardization. This leaves education stratum balance as an evaluation of the standardized report.

The exact yearly split sizes are listed in Table 5. The test grid in Table 6 uses the resolved support rule of at least 200 test observations in each comparison group within the reporting stratum. The same support rule is applied to raw and standardized summaries before comparing reversal counts or conflict strengths.

These results support the tabular replication claim: all five years show lower reversal counts and lower total conflict strength after standardization, while high conflict reversals nearly vanish.

C.3 Folktables income quintile stress test

The income quintile stress test also uses ACS person records from 2014 through 2018, but changes both the target and comparison group. The target is income quintile, the comparison groups are White and Black, and the reporting strata are education buckets. This design separates supported reversal counts from the magnitude of aggregate and stratum disagreement.

The split roles, score, support rule, and raw to standardized comparison are the same as in Appendix C.2; exact yearly split sizes and grids are listed in Tables 5 and 6. The profile H uses fixed softmax summaries and excludes race and education. The restricted classifier feature policy removes direct income proxy features before prediction and propensity reporting, keeping the standardized comparison tied to prediction difficulty summaries.

The 2017 row explains why the paper reports more than reversal counts alone. Supported reversals increase from 182 to 202, while high conflict strength reversals change from 27 to 0 and conflict strength changes from 2.675 to 0.144. The standardized curve therefore reduces the magnitude of aggregate and stratum disagreement even though one discrete count rises.

Table 12: ATUS evaluation details for the primary SIMPLE score on the test split.

Measure	Raw	Standardized
SIMPLE reversals	36	10
SIMPLE conflict strength	0.463730	0.002697
Maximum profile standardized mean difference	0.0817	0.0101
Employment cell total variation	0.142	0.091
Effective sample size fraction	0.981913	

C.4 ATUS ACTIVITY PREDICTION

The ATUS experiment uses daily time use diary episodes. A closed world top 10 non sleep activity vocabulary is learned from training respondents, and episodes outside that vocabulary are removed in every split. The closed world vocabulary leaves 178,477 analyzed episodes, with 17,783 test observations for final curve evaluation, as shown in Table 5. The comparison group is Gender and the reporting stratum is employment status.

The classifier is a regularized histogram gradient boosting classifier trained on time, calendar, demographic, region, and household features available before reporting. The primary score is SIMPLE, with APS (Romano et al., 2020b), RAPS (Angelopoulos et al., 2021), and SAPS (Huang et al., 2023) as robustness checks on the grid in Table 6. The standardization profile uses the four fixed probability features in Table 7. The summaries report profile balance, employment stratum mixture balance, and effective sample size for the standardized comparison.

D STANDARDIZATION TARGETS AND PROFILE CHOICE

The main text uses the standardized profile distribution Q_H because the empirical goal is a stable standardized comparison. This appendix compares the raw, marginal-profile, standardized, and uniform-cell reports using the notation of Appendix A. Here P_H denotes the marginal distribution of the prediction profile $H = H(X)$ in the evaluation population: $P_H(B) = \mathbb{P}\{H(X) \in B\}$. It is not group-specific; raw group reports average over $P(H \in \cdot \mid G = g)$, the marginal-profile report averages both groups over P_H, and the main standardized report averages both groups over Q_H.

Observed-mixture report. The raw curve uses the profile distribution observed within group g:

$$V_g(\alpha) = \mathbb{E}_{H \sim P(H \mid G=g)} \mathbb{E}\{M_\alpha(X) \mid G = g, H\}. \tag{59}$$

It therefore reports the mean set size delivered to group g under that group's observed mixture of prediction profiles.

Marginal profile report. Another option is to average both groups over the marginal profile distribution P_H:

$$V_g^{P_H}(\alpha) = \mathbb{E}_{H \sim P_H} \mathbb{E}\{M_\alpha(X) \mid G = g, H\}. \tag{60}$$

In the two-group case with $e(h) = \mathbb{P}(G = a \mid H = h)$, the corresponding inverse-probability factors are proportional to

$$\frac{1}{e(H)} \quad \text{for group } a, \qquad \frac{1}{1 - e(H)} \quad \text{for group } b. \tag{61}$$

These factors become large when $e(H)$ is close to 0 or 1, so the marginal-profile report can be sensitive to weak overlap.

Standardized report. The main analysis instead uses Q_H from Equation (9), giving the standardized curve $\tilde{V}_g(\alpha)$ in Equation (10). The corresponding factors are

$$w_a(H) = 1 - e(H), \qquad w_b(H) = e(H). \tag{62}$$

Both factors are bounded. Profiles dominated by one group therefore receive less weight, and $\tilde{V}_g(\alpha)$ compares the groups over profile regions represented in both.

Table 13: Several population indices for the same delivered set sizes. The paper uses the raw and standardized reports because together they show the observed burden and the standardized burden.

Report	Target population	Weight behavior	Interpretation
Raw group curve	Group-specific $P(H \mid G = g)$	No reweighting	Observed-mixture burden delivered to each group.
Marginal-profile curve	Marginal distribution P_H	$1/e(H)$ and $1/\{1 - e(H)\}$	Full-population standardized contrast; sensitive to weak support.
Standardized curve	Common standardized distribution Q_H	Bounded factors $1 - e(H)$ and $e(H)$	Propensity standardized contrast; bounded and stable.
Uniform-cell curve	Selected reporting cells equally weighted	Cell weights chosen by design	Cell-equal sensitivity summary; can overweight rare or unstable cells.

Comparing the reports. These contrasts differ in the distribution used to average the same delivered set sizes. The raw gap uses the group-specific distributions $P(H \mid G = g)$, the marginal-profile contrast $V_a^{P_H}(\alpha) - V_b^{P_H}(\alpha)$ uses P_H, and the standardized contrast $\tilde{\Delta}(\alpha) = \tilde{V}_a(\alpha) - \tilde{V}_b(\alpha)$ uses Q_H. They are therefore different population indices, not competing estimates of the same quantity. A difference between the raw and standardized gaps reflects sensitivity to the reporting population; the predictor, score, calibration threshold, and prediction sets remain unchanged.

Why standardize the prediction profile H rather than the cells. A distribution defined directly over the cells provides a useful alternative summary, but it does not serve the same purpose as propensity standardization. The variable Z defines where finer-stratum differences are examined, and its meaning and granularity vary across tasks. Matching Z controls differences in the proportions of the reported strata. Our goal is further to compare set sizes under similar prediction characteristics, which can remain different within the same stratum. In particular, H is constructed from frozen classifier summaries that are directly relevant to the delivered prediction sets. We use the standardized distribution Q_H because it emphasizes profile regions represented in both groups. The resulting comparison is based on prediction profiles that are both relevant to the delivered sets and well represented in the two groups, rather than only on matching the proportions of the reported strata.

Reporting cells. Standardization balances functions of H, not the reporting variable Z unless Z is included in H. In our experiments, Z is often excluded deliberately because it is the axis on which finer-stratum differences are reported. The remaining difference between $\tilde{\pi}_a$ and $\tilde{\pi}_b$ therefore records residual reporting-cell composition after profile standardization.

Choosing the prediction profile. The profile H determines which prediction characteristics are balanced before the group curves are compared. We construct H from frozen classifier summaries such as top probabilities, margins, entropy, and top-label indicators. We do not include realized labels, realized set sizes, calibration thresholds, or final curve summaries. The reporting variable Z is also excluded when it is the axis of comparison. Support, balance, and effective sample size are then checked for the fitted reweighting.

Finite-sample stabilization. The population definitions use the ideal propensity $e(H)$. In the experiments, the propensity is estimated on the propensity split and then frozen before test evaluation. Standardized curves are computed with normalized Hájek means after the prespecified bounded-weight stabilization. Effective sample size and balance checks accompany these estimates to summarize finite-sample support. The reporting weights do not enter the conformal calibration quantile.

E TILTED MONDRIAN DETAILS

E.1 TILTED MONDRIAN

A demonstration example. As an auxiliary 28 occupation illustration, we finetune DistilBERT for one epoch on all 28 Bias in Bios occupations, using a $40/25/25/10$ train/calibration/tilt/test split, SIMPLE scores at $\alpha = 0.10$, group-Mondrian calibration with $\epsilon = 0.01$. The intervention is selected using only $\mathcal{D}_{\text{tilt}}$, then frozen before recalibration and test evaluation. For the "model" occupation, baseline coverage on the tilt split was 37.7% in the Male group and 88.4% in the Female group, a gap of 50.7 percentage points, with a similarly large test gap of 48.8 percentage points, 41.1% versus 89.8%. For "paralegal", Tilted Mondrian increased Male tilt split coverage from 30.3% to 40.9% while Female coverage remained 55.3%, reducing the tilt split gap by 10.6 percentage points. That is, from 25.0 to 14.4 percentage points. After recalibration, the test gap decreased from 21.3 to 0.9 percentage points, a reduction of 20.4 percentage points.

The intervention report uses the same prespecified $G \times Z$ cells as Section 6. Bias in Bios has label-defined $Z = Y$, whereas Folktables and ATUS use covariate-defined intervention cells. Table 14 lists the group–stratum definitions.

Table 14: Intervention cells for Tilted Mondrian.

Experiment	Group and intervention stratum
Bias in Bios	Gender and occupation, with $Z = Y$
Folktables occupation	Gender and education bucket
Folktables income	Race group and education bucket
ATUS	Gender and employment status

The tilt split is used to choose the intervention, which is then frozen. The calibration split then supplies ordinary group Mondrian calibration for the frozen score S_θ, and the test split supplies the final evaluation. The Bias in Bios RAPS row uses $\lambda = 0.01$ as a separate robustness instance; its Section 6 curve report uses $\lambda = 0.2$. The search routine is finite and deterministic rather than exhaustive: it combines coordinate paths, critical threshold values, stratum-wise low dimensional block updates, beam search, and Halton search. Each stratum-wise block treats (θ_z^+, θ_z^-) jointly and considers advantaged-side, disadvantaged-side, and joint directional updates. The hyperparameters use grid $0, 0.025, \ldots, 3$, beam width 20, boundary projection grid with 41 points, and 1000 Halton candidates.

The feasibility checks in Algorithm 4 define the candidate set for the score search. On the tilt split, a retained directional score shift preserves each group's covered count, satisfies the ϵ harm limit for cells that were initially advantaged, maintains coverage in cells that were initially disadvantaged, and leaves every stratum gap weakly smaller. Theorem 6 turns these primitive checks into a tilt split result for the largest single stratum mitigation. The improvement is controlled by the original stratum gap and the same group coverage budget available in cells where that group was initially advantaged. Algorithm 4 chooses $\theta = (\theta^+, \theta^-)$, forms S_θ, recalibrates the frozen score, and evaluates the induced prediction sets on the test split.

Algorithm 4 Tilted Mondrian CP with directional score shifting (detailed version of Algorithm 2)

Input: Base nonconformity score S_0; binary group G and intervention stratum Z; tilt split $\mathcal{D}_{\text{tilt}}$, calibration split $\mathcal{D}_{\text{cal}}$, test split $\mathcal{D}_{\text{test}}$; nominal level α, harm budget ϵ, and fixed group-Mondrian rank/tie rule.

1: On $\mathcal{D}_{\text{tilt}}$, use the fixed rank/tie rule to compute baseline group-Mondrian thresholds q_g^0, sets $\mathcal{C}_{\alpha,0}$, and covered counts

$$k_g^0 = \sum_{i \in \mathcal{D}_{\text{tilt}}: G_i = g} \mathbf{1}\{Y_i \in \mathcal{C}_{\alpha,0}(X_i)\}.$$

2: For each audited stratum z, compute

$$c_{g,z}^0 = \frac{\sum_{i \in \mathcal{D}_{\text{tilt}}} \mathbf{1}\{G_i = g, Z_i = z, Y_i \in \mathcal{C}_{\alpha,0}(X_i)\}}{\sum_{i \in \mathcal{D}_{\text{tilt}}} \mathbf{1}\{G_i = g, Z_i = z\}},$$

set $g^+(z) = \arg\max_g c_{g,z}^0$, $g^-(z) = \arg\min_g c_{g,z}^0$, and $d_z = c_{g^+(z),z}^0 - c_{g^-(z),z}^0$.

3: For nonnegative θ_z^+, θ_z^-, define

$$S_\theta(x,y) = S_0(x,y) + \begin{cases} \theta_y^+ \mathbf{1}\{G = g^+(y)\} - \theta_y^- \mathbf{1}\{G = g^-(y)\}, & Z = Y, \\ \theta_{Z(x)}^+ \mathbf{1}\{G = g^+(Z(x))\} & \\ -\theta_{Z(x)}^- \mathbf{1}\{G = g^-(Z(x))\}, & Z = Z(X). \end{cases}$$

4: Build a finite deterministic candidate set from coordinate paths, critical threshold grids, stratum-wise updates of (θ_z^+, θ_z^-), block updates, beam, and Halton search.

5: **for** each candidate θ **do**

6: Reselect thresholds q_g^θ on $\mathcal{D}_{\text{tilt}}$ using the same rank/tie rule so that every group covered count remains exactly k_g^0.

7: Recompute $c_{g,z}^\theta$; set $d_z^\theta = |c_{a,z}^\theta - c_{b,z}^\theta|$ and $m_z^\theta = d_z - d_z^\theta$.

8: Retain θ iff, for every z,

$$c_{g^+(z),z}^0 - c_{g^+(z),z}^\theta \le \epsilon, \qquad c_{g^-(z),z}^\theta \ge c_{g^-(z),z}^0, \qquad d_z^\theta \le d_z.$$

9: **end for**

10: Select retained θ lexicographically by smallest worst stratum gap, smallest mean stratum gap, largest total mitigation, smallest directional shift norm, and, only for remaining ties, smallest prespecified average set size cost on the tilt split.

11: Record tilt split feasibility checks and the Theorem 6 result for the selected θ.

12: Freeze S_θ. On $\mathcal{D}_{\text{cal}}$, obtain final group-Mondrian thresholds at level α.

13: Evaluate once on $\mathcal{D}_{\text{test}}$; report test coverage gaps, harmed/helped cells, set sizes, selected shifts, and achieved mitigations.

Output: Frozen score S_θ, final calibrated group thresholds, tilt split feasibility report, and test coverage/set size report.

Invariant: $\mathcal{D}_{\text{test}}$ enters neither candidate generation, feasibility checking, ranking, nor calibration; it is evaluated once after the score and calibration thresholds are frozen.

Table 15: Test APS, RAPS, and SAPS directional score intervention results. Columns follow Table 2. Harm is the largest tilt split coverage loss among initially advantaged cells. Rounding values may conceal changes at higher precision.

Experiment	Score	$L^0_{\max} \to L^\theta_{\max}$	$\Delta L_{\max}$	$\max_z m^\theta_z$	Harm	$\Delta\lvert C\rvert$
Bias in Bios	APS	0.046→0.043	0.003	0.008 (nurse)	0.0077	0.0414
Bias in Bios ($\lambda = 0.01$)	RAPS	0.036→0.036	0.000	0.006 (nurse)	0.0050	0.3389
Bias in Bios	SAPS	0.178→0.083	0.095	0.173 (surgeon)	0.0082	0.0090
Folktables occupation	APS	0.105→0.090	0.015	0.019 (Bachelor)	0.0098	−0.1008
Folktables occupation	RAPS	0.101→0.084	0.017	0.017 (Less than high school)	0.0098	−0.0491
Folktables occupation	SAPS	0.028→0.013	0.015	0.022 (Bachelor)	0.0100	0.0078
Folktables income	APS	0.094→0.065	0.029	0.029 (Less than high school)	0.0096	−0.0031
Folktables income	RAPS	0.094→0.064	0.029	0.029 (Less than high school)	0.0100	−0.0039
Folktables income	SAPS	0.020→0.020	0.000	0.009 (Bachelor)	0.0040	0.0017
ATUS	APS	0.013→0.009	0.004	0.004 (unemployed)	0.0093	−0.0063
ATUS	RAPS	0.006→0.016	−0.010	0.001 (employed)	0.0073	0.0017
ATUS	SAPS	0.007→0.036	−0.028	0.002 (employed)	0.0093	0.0004

Because feasibility is enforced on the tilt split, a test stratum gap can increase after the intervention is frozen.

Table 16 reports the tilt split result from Theorem 6 for every dataset and score. The active term indicates whether the maximum improvement is controlled by the original stratum gap or by the harm budget. The test column reports the achieved test split mitigation for comparison.

Table 16: Tilt split result for Tilted Mondrian. The Theorem 6 bound is evaluated on the tilt split because the intervention feasibility constraints are checked there. The test column reports the corresponding achieved mitigation on the test split and is not used as a feasibility check.

Dataset	Score	Theorem 6 bound	Active term	Tilt split $\max_z m^\theta_z$	Test $\max_z m^\theta_z$
Bias in Bios	SIMPLE	0.156	gap d_z	0.125	0.071
	APS	0.028	ϵ-budget term	0.017	0.008
	RAPS ($\lambda = 0.01$)	0.041	gap d_z	0.040	0.006
	SAPS	0.166	gap d_z	0.164	0.173
Folktables occupation	SIMPLE	0.037	gap d_z	0.022	0.024
	APS	0.039	ϵ-budget term	0.021	0.019
	RAPS	0.039	ϵ-budget term	0.015	0.017
	SAPS	0.035	gap d_z	0.025	0.022
Folktables income	SIMPLE	0.005	gap d_z	0.005	0.009
	APS	0.054	ϵ-budget term	0.020	0.029
	RAPS	0.054	ϵ-budget term	0.022	0.029
	SAPS	0.009	gap d_z	0.009	0.009
ATUS	SIMPLE	0.009	gap d_z	0.001	0.001
	APS	0.024	gap d_z	0.020	0.004
	RAPS	0.017	gap d_z	0.016	0.001
	SAPS	0.020	gap d_z	0.009	0.002

Table 17 shows the sensitivity analysis of Tilted Mondrian mitigation limit. When the harm budget ϵ is the main factor to limit the mitigation, increasing ϵ enlarges the theoretical mitigation limit; once the original stratum gap d_z becomes the main factor, the bound saturates. Configurations that are already saturated at $\epsilon = 0.01$ therefore remain unchanged across the grid. Across all settings, the mitigation stays within the corresponding theoretical limit.

Table 17: Sensitivity to the harm budget ϵ. Each entry reports B_ϵ/M_ϵ, where B_ϵ is the Theorem 6 bound and $M_\epsilon = \max_z m_z^\theta$ is the achieved tilt split mitigation. A † indicates that the ϵ-budget term is active in the bound; otherwise the original stratum gap d_z is active.

Score	$\epsilon = .01$	$\epsilon = .02$	$\epsilon = .03$	$\epsilon = .04$	$\epsilon = .05$
Bias in Bios					
SIMPLE	0.1557/0.1245	0.1557/0.1197	0.1557/0.1268	0.1557/0.1341	0.1557/0.1363
APS	0.0276†/0.0169	0.0477/0.0169	0.0477/0.0169	0.0477/0.0169	0.0477/0.0169
RAPS ($\lambda = .01$)	0.0414/0.0399	0.0414/0.0399	0.0414/0.0399	0.0414/0.0399	0.0414/0.0399
SAPS	0.1656/0.1640	0.1656/0.1373	0.1656/0.1293	0.1656/0.1628	0.1656/0.1628
Folktables occupation					
SIMPLE	0.0370/0.0225	0.0370/0.0336	0.0370/0.0336	0.0370/0.0336	0.0370/0.0336
APS	0.0395†/0.0211	0.0543†/0.0284	0.0815†/0.0439	0.0963/0.0439	0.0963/0.0483
RAPS	0.0395†/0.0150	0.0543†/0.0280	0.0815†/0.0280	0.0875/0.0366	0.0875/0.0648
SAPS	0.0348/0.0253	0.0348/0.0273	0.0348/0.0282	0.0348/0.0282	0.0348/0.0282
Folktables income					
SIMPLE	0.0054/0.0049	0.0054/0.0049	0.0054/0.0049	0.0054/0.0049	0.0054/0.0049
APS	0.0540†/0.0198	0.0878/0.0309	0.0878/0.0381	0.0878/0.0404	0.0878/0.0404
RAPS	0.0540†/0.0222	0.0878/0.0324	0.0878/0.0371	0.0878/0.0475	0.0878/0.0475
SAPS	0.0092/0.0092	0.0092/0.0092	0.0092/0.0092	0.0092/0.0092	0.0092/0.0092
ATUS					
SIMPLE	0.0092/0.0009	0.0092/0.0009	0.0092/0.0009	0.0092/0.0009	0.0092/0.0009
APS	0.0242/0.0203	0.0242/0.0231	0.0242/0.0231	0.0242/0.0231	0.0242/0.0231
RAPS	0.0170/0.0157	0.0170/0.0157	0.0170/0.0157	0.0170/0.0157	0.0170/0.0157
SAPS	0.0195/0.0091	0.0195/0.0090	0.0195/0.0090	0.0195/0.0090	0.0195/0.0090

Stress test under direct finer-cell calibration. Direct finer-cell calibration is a natural and direct choice when coverage control is required in every (G, Z) cell. Tilted Mondrian is not intended to replace this construction. It instead acts as an additional score adapter on top of group-Mondrian calibration. In our setting, Z indexes the finer strata targeted for mitigation, for either coverage or set-size disparities, while the final calibration remains at G. Equalizing coverage does not necessarily equalize set size, and equalizing set size does not necessarily equalize coverage. Tilted Mondrian is used to reduce the finer-stratum disparities while satisfying prespecified constraints. When the finer strata are imbalanced, direct cellwise calibration estimates quantiles from unequal and sometimes small calibration samples, making these quantiles more discrete and more sensitive to sampling variation. This can also yield conservatively large prediction sets in small cells.

We additionally compare the distribution of delivered set sizes under Tilted Mondrian and direct (G, Z)-cell calibration on the same 20 resplits for coverage mitigation. Because the procedures calibrate at different granularities, this is a descriptive stress test rather than a like-for-like comparison. Furthermore, direct cellwise Mondrian does not impose the loss-control condition used by Tilted Mondrian. On the test split, we further record the largest cellwise coverage loss from the group-Mondrian over strata initially identified on $\mathcal{D}_{\text{tilt}}$. Using the 1-percentage-point harm budget, this loss exceeds 1 pp in $289/320$ resplits under direct cellwise calibration, compared with $103/320$ under Tilted Mondrian. Tables 18 and 19 report two dispersion measures of the final set sizes, the Gini coefficient and relative mean absolute deviation (RMAD), computed within each test split and then averaged across resplits. Tilted Mondrian yields lower dispersion in $15/16$ data–score configurations under both measures, and in $282/320$ and $296/320$ paired resplits, respectively.

E.2 Boosted Tilted Mondrian

Motivated by gradient boosted conformal score adaptation (Xie et al., 2024), we consider a covariate adaptive extension of Tilted Mondrian. It keeps the same mitigation target, exact constraints, and recalibration protocol as Algorithm 4, but replaces the stratum-wise constants $\theta_z^\pm$ by nonnegative functions $\theta_z^\pm(x)$. Using the same stratum convention as above,

$$S_\theta(x, y) = S_0(x, y) + \theta_z^+(x)\mathbf{1}\{G = g^+(z)\} - \theta_z^-(x)\mathbf{1}\{G = g^-(z)\}.$$

Table 18: Set size dispersion under direct finer-cell calibration: Gini coefficient. For each resplit, let $S_i = |\mathcal{C}(X_i)|$ and $\bar{S} = n^{-1}\sum_i S_i$. We compute $G = (2n^2\bar{S})^{-1}\sum_{i,j}|S_i - S_j|$. Entries are mean $\pm$ SD over the same 20 paired resplits. Lower is better; bold marks the lower value.

Score	Method	Bias in Bios	Folktables occupation	Folktables income	ATUS
SIMPLE	Tilted	$\mathbf{0.0436 \pm 0.0032}$	$\mathbf{0.2579 \pm 0.0009}$	$\mathbf{0.1915 \pm 0.0009}$	$\mathbf{0.1390 \pm 0.0007}$
	Cellwise	0.1118 ± 0.0031	0.2722 ± 0.0007	0.1930 ± 0.0008	0.1536 ± 0.0011
APS	Tilted	$\mathbf{0.3786 \pm 0.0246}$	$\mathbf{0.2789 \pm 0.0012}$	$\mathbf{0.2259 \pm 0.0012}$	0.1185 ± 0.0011
	Cellwise	0.4068 ± 0.0010	0.3005 ± 0.0012	0.2262 ± 0.0010	$\mathbf{0.1133 \pm 0.0015}$
RAPS	Tilted	$\mathbf{0.3036 \pm 0.0153}$	$\mathbf{0.2391 \pm 0.0012}$	$\mathbf{0.2259 \pm 0.0011}$	$\mathbf{0.0998 \pm 0.0009}$
	Cellwise	0.3687 ± 0.0009	0.2619 ± 0.0014	0.2262 ± 0.0010	0.1041 ± 0.0016
SAPS	Tilted	$\mathbf{0.0434 \pm 0.0035}$	$\mathbf{0.2643 \pm 0.0006}$	$\mathbf{0.2082 \pm 0.0009}$	$\mathbf{0.0380 \pm 0.0007}$
	Cellwise	0.1790 ± 0.0039	0.2729 ± 0.0008	0.2092 ± 0.0008	0.0886 ± 0.0014

Table 19: Set size dispersion under direct finer-cell calibration: relative mean absolute deviation. For each resplit, let $S_i = |\mathcal{C}(X_i)|$ and $\bar{S} = n^{-1}\sum_i S_i$. We compute $\mathrm{RMAD} = (n\bar{S})^{-1}\sum_i |S_i - \bar{S}|$. Entries are mean $\pm$ SD over the same 20 paired resplits. Lower is better; bold marks the lower value.

Score	Method	Bias in Bios	Folktables occupation	Folktables income	ATUS
SIMPLE	Tilted	$\mathbf{0.068 \pm 0.007}$	$\mathbf{0.397 \pm 0.001}$	$\mathbf{0.286 \pm 0.002}$	$\mathbf{0.203 \pm 0.001}$
	Cellwise	0.209 ± 0.006	0.417 ± 0.002	0.289 ± 0.002	0.222 ± 0.002
APS	Tilted	$\mathbf{0.564 \pm 0.040}$	$\mathbf{0.411 \pm 0.002}$	$\mathbf{0.350 \pm 0.002}$	0.172 ± 0.002
	Cellwise	0.609 ± 0.002	0.427 ± 0.002	0.352 ± 0.001	$\mathbf{0.165 \pm 0.002}$
RAPS	Tilted	$\mathbf{0.467 \pm 0.028}$	$\mathbf{0.372 \pm 0.002}$	$\mathbf{0.350 \pm 0.002}$	$\mathbf{0.145 \pm 0.002}$
	Cellwise	0.557 ± 0.001	0.390 ± 0.002	0.352 ± 0.001	0.152 ± 0.003
SAPS	Tilted	$\mathbf{0.086 \pm 0.007}$	$\mathbf{0.419 \pm 0.002}$	$\mathbf{0.337 \pm 0.001}$	$\mathbf{0.055 \pm 0.002}$
	Cellwise	0.312 ± 0.007	0.428 ± 0.002	0.339 ± 0.001	0.132 ± 0.003

Let $\mathcal{Z}$ denote the strata and $n_{g,z}$ the tilt split size of cell (g, z). Let $\sigma(u) = (1 + e^{-u})^{-1}$, and let $\rho_g > 0$ be the fixed group-specific smoothing scale. Using the baseline covered count k_g^0 defined above, define the soft group threshold $\xi_g(\theta)$ by

$$\sum_{i:G_i=g} \sigma\left(\frac{\xi_g(\theta) - S_\theta(X_i, Y_i)}{\rho_g}\right) = k_g^0.$$

The corresponding soft cell coverage and gap are

$$\tilde{c}_{g,z}(\theta) = \frac{1}{n_{g,z}} \sum_{i:G_i=g,Z_i=z} \sigma\left(\frac{\xi_g(\theta) - S_\theta(X_i, Y_i)}{\rho_g}\right), \qquad \tilde{d}_z(\theta) = \left|\tilde{c}_{g^+(z),z}(\theta) - \tilde{c}_{g^-(z),z}(\theta)\right|.$$

We perform boosting with the log-sum-exp surrogate

$$\tilde{L}(\theta) = \frac{1}{50}\log \sum_{z\in\mathcal{Z}} \exp\Big\{50\,\tilde{d}_z(\theta)\Big\}.$$

Thus, every stratum contributes to the gradient, with larger gaps receiving more weight. Let $\mathcal{B}$ denote the nonnegative amplitude grid. The smooth objective only proposes search directions: every candidate is re-evaluated with the exact tilt split thresholds and the original hard constraints. For the experiments, we use depth-three XGBoost trees, log-sum-exp coefficient 50, the amplitude grid $\mathcal{B} = \{0\} \cup \{2^{-16}, \ldots, 2^3\}$, and select among 0–10 boosting rounds by three-fold cross-validation within $\mathcal{D}_{\mathrm{tilt}}$.

Algorithm 5 Boosted Tilted Mondrian

Input: Base score S_0; group G and stratum Z; tilt, calibration, and test splits $\mathcal{D}_{\text{tilt}}$, $\mathcal{D}_{\text{cal}}$, and $\mathcal{D}_{\text{test}}$; level α; harm budget ϵ; fixed smoothing scales $\{\rho_g\}_g$; nonnegative amplitude grid $\mathcal{B}$; fixed group-Mondrian rank/tie rule.
On $\mathcal{D}_{\text{tilt}}$, compute the baseline covered counts k_g^0, cell coverages $c_{g,z}^0$, mitigation directions $g^{\pm}(z)$, and strata $\mathcal{Z}$.
Use three-fold cross-validation entirely within $\mathcal{D}_{\text{tilt}}$ to select the boosting round $T^\star$.
Refit on the full $\mathcal{D}_{\text{tilt}}$. Initialize the smooth boosting path at $\theta = 0$ and the best exact feasible adapter at $\theta^\star = 0$.
for $t = 1, \ldots, T^\star$ **do**
    Fit depth-three trees $h_{z,t}^+(x)$ and $h_{z,t}^-(x)$ to the negative gradients of $\tilde{L}(\theta)$ with respect to $\theta_z^+(x)$ and $\theta_z^-(x)$.
    For each $(\lambda^+, \lambda^-) \in \mathcal{B}^2$, form candidates

$$\theta_z'^+(x) = \left[\theta_z^+(x) + \lambda^+ h_{z,t}^+(x)\right]_+, \qquad \theta_z'^-(x) = \left[\theta_z^-(x) + \lambda^- h_{z,t}^-(x)\right]_+.$$

    For every candidate θ', recompute the exact tilt split thresholds under the fixed rank/tie rule and retain only candidates satisfying the original Tilted Mondrian constraints on $\mathcal{D}_{\text{tilt}}$.
    Update $\theta^\star$ to the lexicographically best adapter among the current $\theta^\star$ and all retained candidates, using the same exact ranking rule as Algorithm 4.
    Let $\theta_{\text{smooth}} \in \arg\min_{\theta' \text{ retained}} \tilde{L}(\theta')$.
    if $\tilde{L}(\theta_{\text{smooth}}) < \tilde{L}(\theta)$ **then**
        $\theta \leftarrow \theta_{\text{smooth}}$.
    else
        break.
    end if
end for
Freeze $S_{\theta^\star}$. On $\mathcal{D}_{\text{cal}}$, compute the modified true-label scores and obtain the final group-Mondrian thresholds $\hat{q}_g$ at level α.
On $\mathcal{D}_{\text{test}}$, form

$$\mathcal{C}_\alpha(x) = \{y : S_{\theta^\star}(x, y) \le \hat{q}_{G(x)}\},$$

and report test coverage gaps and set sizes.
Output: Frozen boosted score $S_{\theta^\star}$, calibrated group thresholds $\{\hat{q}_g\}_g$, and test coverage/set size report.

Tables 20–21 compare the Boosted Tilted Mondrian with Tilted Mondrian using the SIMPLE score.

Take-away. The covariate adaptivity of BTM does not translate into stronger mitigation. Tilted Mondrian often achieves larger worst-gap reductions and lower average remaining stratum disparities. This suggests that the structured search is better suited to the constrained mitigation.

Table 20: SIMPLE robustness comparison between Tilted Mondrian and Boosted Tilted Mondrian. Tilted Mondrian (TM) is summarized over 20 resplits; Boosted Tilted Mondrian (BTM) over the ten shared resplits. For either method, $\Delta L_{\max} = L^0_{\max} - L^{\text{final}}_{\max}$, where $L_{\max}$ is the worst supported fine-grained coverage gap; positive values indicate improvement. Continuous entries are mean $\pm$ SD. Null interventions count the unchanged solution.

Method	Tilt split worst-gap reduction $\Delta L_{\max}$	Test worst-gap reduction $\Delta L_{\max}$	Tilt split maximum harm	Test mean set size change	Null interventions
Bias in Bios					
TM (20)	0.0778 ± 0.0397	0.0680 ± 0.0345	0.0090 ± 0.0020	0.0038 ± 0.0034	$0/20$
BTM (10)	0.0145 ± 0.0144	0.0145 ± 0.0112	0.0041 ± 0.0027	0.0002 ± 0.0003	$0/10$
Folktables occupation					
TM (20)	0.0231 ± 0.0025	0.0209 ± 0.0060	0.0099 ± 0.0001	0.0040 ± 0.0023	$0/20$
BTM (10)	0.0099 ± 0.0049	0.0075 ± 0.0055	0.0076 ± 0.0035	0.0051 ± 0.0058	$0/10$
Folktables income					
TM (20)	0.0157 ± 0.0049	-0.0002 ± 0.0082	0.0074 ± 0.0018	-0.0002 ± 0.0015	$0/20$
BTM (10)	0.0071 ± 0.0095	0.0012 ± 0.0027	0.0035 ± 0.0035	0.0010 ± 0.0017	$3/10$
American Time Use Survey					
TM (20)	0.0031 ± 0.0052	-0.0067 ± 0.0119	0.0032 ± 0.0043	0.0015 ± 0.0024	$9/20$
BTM (10)	0.0029 ± 0.0064	-0.0032 ± 0.0068	0.0011 ± 0.0031	0.0005 ± 0.0012	$8/10$

Table 21: Final mean fine-grained coverage gap on the SIMPLE resplits. Entries are averages across resplits; lower is better. This complements Table 20, which reports worst-gap reduction, by measuring the average remaining disparity across strata.

	Tilt split final mean gap		Test final mean gap	
Dataset	Tilted Mondrian	Boosted Tilted Mondrian	Tilted Mondrian	Boosted Tilted Mondrian
Bias in Bios	**0.03291**	0.05111	**0.03353**	0.04918
Folktables occupation	**0.00812**	0.01615	**0.00994**	0.01845
Folktables income	**0.00067**	0.00398	**0.01127**	0.01132
American Time Use Survey	**0.00576**	0.00641	0.01294	**0.01047**

F Finite Sample Robustness

To assess finite-sample sensitivity, we freeze each trained predictor and its probability outputs and repeatedly repartition the remaining observations. For standardization, the resplits contain calibration, propensity-estimation, and test data; for mitigation, they contain tilt, calibration, and test data. Across 50 standardization resplits, conflict strength decreases in 899 of 900 configuration-resplit evaluations, with normalized test effective sample size fractions between 0.601 and 0.984 (Table 22). Across 20 mitigation resplits per dataset and score, all 320 mitigation problems satisfy the prespecified constraints, although test worst-gap improvements vary by dataset and score (Table 23).

Table 22: Fixed model robustness of propensity standardization. Each row summarizes 50 calibration/propensity/test resplits. Conflict strength and normalized test effective sample size are reported as mean ± SD; relative reduction is the median [P5, P95]. D/T/I counts resplits in which standardization decreases, ties, or increases the reversal count.

Setting	Raw → standardized strength	Relative reduction median [P5, P95] (%)	Reversal D/T/I	Test ESS
Bias in Bios				
SIMPLE	$0.118 \pm 0.034 \to 0.006 \pm 0.004$	$95.3\,[83.6, 99.5]$	$49/0/1$	0.824 ± 0.004
APS	$0.152 \pm 0.068 \to 0.012 \pm 0.008$	$92.5\,[77.1, 98.0]$	$48/1/1$	0.824 ± 0.004
SAPS	$0.151 \pm 0.038 \to 0.006 \pm 0.004$	$96.4\,[90.2, 98.7]$	$50/0/0$	0.824 ± 0.004
RAPS	$0.156 \pm 0.056 \to 0.010 \pm 0.006$	$93.3\,[81.6, 98.4]$	$49/1/0$	0.824 ± 0.004
Folktables occupation				
2014	$9.610 \pm 0.535 \to 0.121 \pm 0.071$	$99.0\,[97.5, 99.6]$	$50/0/0$	0.666 ± 0.003
2015	$8.054 \pm 0.463 \to 0.419 \pm 0.212$	$95.4\,[90.4, 98.6]$	$44/0/6$	0.609 ± 0.004
2016	$6.883 \pm 0.468 \to 0.264 \pm 0.121$	$96.3\,[92.9, 98.7]$	$49/0/1$	0.696 ± 0.007
2017	$6.616 \pm 0.350 \to 0.669 \pm 0.184$	$90.0\,[83.2, 94.7]$	$50/0/0$	0.640 ± 0.005
2018	$7.625 \pm 0.537 \to 0.483 \pm 0.159$	$93.7\,[89.8, 97.1]$	$50/0/0$	0.619 ± 0.005
Folktables income				
2014	$1.665 \pm 0.475 \to 0.072 \pm 0.039$	$96.0\,[91.1, 98.7]$	$41/0/9$	0.950 ± 0.002
2015	$1.800 \pm 0.593 \to 0.057 \pm 0.039$	$96.9\,[93.5, 99.0]$	$39/1/10$	0.954 ± 0.001
2016	$2.145 \pm 0.551 \to 0.130 \pm 0.069$	$94.2\,[88.3, 96.7]$	$43/0/7$	0.953 ± 0.001
2017	$2.317 \pm 0.607 \to 0.132 \pm 0.064$	$94.9\,[88.3, 97.6]$	$33/1/16$	0.950 ± 0.001
2018	$2.208 \pm 0.610 \to 0.145 \pm 0.073$	$93.9\,[88.8, 96.0]$	$42/0/8$	0.949 ± 0.001
ATUS				
SIMPLE	$0.280 \pm 0.077 \to 0.015 \pm 0.025$	$97.8\,[79.3, 99.2]$	$47/1/2$	0.982 ± 0.001
APS	$1.372 \pm 0.201 \to 0.198 \pm 0.149$	$87.3\,[69.4, 96.6]$	$32/16/2$	0.982 ± 0.001
RAPS	$1.016 \pm 0.152 \to 0.138 \pm 0.110$	$88.7\,[70.6, 97.0]$	$32/18/0$	0.982 ± 0.001
SAPS	$0.029 \pm 0.018 \to 0.014 \pm 0.014$	$61.7\,[15.6, 98.4]$	$47/2/1$	0.982 ± 0.001

Table 23: Fixed-model robustness of Tilted Mondrian for coverage. Each row summarizes 20 tilt/calibration/test resplits; continuous entries are mean ± SD. $L^0_{\max}$, $L^\theta_{\max}$, $\Delta L_{\max}$, and $\Delta|C|$ are evaluated on the test split. Harm is the largest tilt split coverage loss among initially advantaged cells. P/T/N counts positive, numerically tied, or negative test worst-gap reductions. All 320 mitigation problems satisfy the prespecified feasibility constraints.

Score	$L^0_{\max} \to L^\theta_{\max}$	$\Delta L_{\max}$	Harm	$\Delta\|C\|$	P/T/N
Bias in Bios					
SIMPLE	$0.172 \pm 0.031 \to 0.104 \pm 0.031$	0.0680 ± 0.0345	0.0090 ± 0.0020	0.0038 ± 0.0034	$20/0/0$
APS	$0.043 \pm 0.007 \to 0.040 \pm 0.005$	0.0025 ± 0.0030	0.0036 ± 0.0022	0.2955 ± 0.2220	$19/1/0$
RAPS	$0.041 \pm 0.009 \to 0.034 \pm 0.007$	0.0074 ± 0.0082	0.0046 ± 0.0012	0.4799 ± 0.0959	$12/7/1$
SAPS	$0.179 \pm 0.034 \to 0.084 \pm 0.016$	0.0951 ± 0.0328	0.0093 ± 0.0011	0.0071 ± 0.0036	$20/0/0$
Folktables occupation					
SIMPLE	$0.044 \pm 0.007 \to 0.023 \pm 0.006$	0.0209 ± 0.0060	0.0099 ± 0.0001	0.0040 ± 0.0023	$20/0/0$
APS	$0.103 \pm 0.007 \to 0.089 \pm 0.008$	0.0147 ± 0.0026	0.0093 ± 0.0014	-0.0915 ± 0.0128	$20/0/0$
RAPS	$0.096 \pm 0.007 \to 0.081 \pm 0.007$	0.0150 ± 0.0037	0.0098 ± 0.0002	-0.0455 ± 0.0063	$20/0/0$
SAPS	$0.043 \pm 0.008 \to 0.023 \pm 0.009$	0.0197 ± 0.0074	0.0091 ± 0.0015	0.0057 ± 0.0033	$20/0/0$
Folktables income					
SIMPLE	$0.020 \pm 0.006 \to 0.020 \pm 0.010$	-0.0002 ± 0.0082	0.0074 ± 0.0018	-0.0003 ± 0.0015	$10/1/9$
APS	$0.091 \pm 0.017 \to 0.069 \pm 0.021$	0.0213 ± 0.0117	0.0088 ± 0.0028	-0.0053 ± 0.0051	$20/0/0$
RAPS	$0.091 \pm 0.017 \to 0.068 \pm 0.021$	0.0229 ± 0.0104	0.0096 ± 0.0018	-0.0057 ± 0.0049	$20/0/0$
SAPS	$0.021 \pm 0.009 \to 0.023 \pm 0.012$	-0.0024 ± 0.0067	0.0063 ± 0.0022	0.0008 ± 0.0011	$7/1/12$
ATUS					
SIMPLE	$0.019 \pm 0.011 \to 0.026 \pm 0.017$	-0.0067 ± 0.0119	0.0033 ± 0.0043	0.0015 ± 0.0024	$2/10/8$
APS	$0.026 \pm 0.015 \to 0.028 \pm 0.015$	-0.0018 ± 0.0136	0.0064 ± 0.0038	-0.0048 ± 0.0046	$8/5/7$
RAPS	$0.020 \pm 0.014 \to 0.025 \pm 0.017$	-0.0050 ± 0.0133	0.0052 ± 0.0035	-0.0005 ± 0.0035	$6/3/11$
SAPS	$0.020 \pm 0.007 \to 0.027 \pm 0.016$	-0.0071 ± 0.0145	0.0040 ± 0.0039	0.0009 ± 0.0016	$6/8/6$

G TILTED MONDRIAN FOR SET SIZE AND FUNDAMENTAL TRADEOFFS

We use coverage as the primary mitigation target in the main text because it is the reliability quantity guaranteed by CP. Here we apply the same Tilted Mondrian construction to set size and report the resulting coverage changes. For cell (g, z), let $\bar{M}_{g,z}$ denote its mean prediction-set size. The baseline set-size gap in stratum z is

$$L^0_{z,S} = |\bar{M}^0_{a,z} - \bar{M}^0_{b,z}|,$$

and after a score adjustment θ it is

$$L^\theta_{z,S} = |\bar{M}^\theta_{a,z} - \bar{M}^\theta_{b,z}|.$$

The set-size-targeted search uses these gaps in place of the coverage gaps d_z and d^θ_z, with the same tilt/calibration/test separation and group-coverage constraint. Harm is the largest set-size increase among cells that initially have the smaller mean set size; the harm budget is 0.02. For Folktables income, $K = 5$ and deterministic RAPS uses $k_{\text{reg}} = 5$. Its regularization term is therefore zero at every label rank, so RAPS and APS coincide in this experiment.

Table 24: Fixed-model robustness of set-size-targeted Tilted Mondrian. Each row summarizes 20 tilt/calibration/test resplits; continuous entries are mean ± SD. Set-size gaps and ΔCov. are evaluated on the test split. Harm is the largest tilt split set-size increase among initially smaller set cells. P/T/N counts positive, tied, or negative test worst-gap reductions.

Score	$L^0_{\max,S} \to L^\theta_{\max,S}$	$\Delta L_{\max,S}$	Harm	ΔCov. (pp)	P/T/N
Bias in Bios					
SIMPLE	$0.066 \pm 0.024 \to 0.065 \pm 0.025$	0.0013 ± 0.0039	0.0038 ± 0.0059	-0.001 ± 0.013	$5/12/3$
APS	$1.269 \pm 0.054 \to 0.514 \pm 0.058$	0.7551 ± 0.0790	0.0061 ± 0.0064	-0.019 ± 0.092	$20/0/0$
RAPS	$1.163 \pm 0.041 \to 0.515 \pm 0.050$	0.6480 ± 0.0573	0.0052 ± 0.0083	-0.028 ± 0.089	$20/0/0$
SAPS	$0.052 \pm 0.016 \to 0.054 \pm 0.032$	-0.0019 ± 0.0301	0.0018 ± 0.0022	-0.001 ± 0.021	$11/5/4$
Folktables occupation					
SIMPLE	$0.217 \pm 0.022 \to 0.168 \pm 0.022$	0.0490 ± 0.0111	0.0198 ± 0.0002	0.009 ± 0.022	$20/0/0$
APS	$2.238 \pm 0.022 \to 2.233 \pm 0.022$	0.0044 ± 0.0040	0.0004 ± 0.0006	0.000 ± 0.002	$17/2/1$
RAPS	$1.621 \pm 0.040 \to 1.580 \pm 0.044$	0.0411 ± 0.0313	0.0159 ± 0.0043	0.008 ± 0.019	$20/0/0$
SAPS	$0.273 \pm 0.027 \to 0.217 \pm 0.032$	0.0560 ± 0.0148	0.0193 ± 0.0009	0.006 ± 0.022	$20/0/0$
Folktables income					
SIMPLE	$0.183 \pm 0.041 \to 0.136 \pm 0.045$	0.0467 ± 0.0256	0.0150 ± 0.0070	0.000 ± 0.019	$20/0/0$
APS	$0.201 \pm 0.036 \to 0.188 \pm 0.042$	0.0132 ± 0.0229	0.0047 ± 0.0078	0.002 ± 0.016	$10/8/2$
RAPS	$0.201 \pm 0.036 \to 0.188 \pm 0.042$	0.0132 ± 0.0229	0.0047 ± 0.0078	0.002 ± 0.016	$10/8/2$
SAPS	$0.229 \pm 0.051 \to 0.190 \pm 0.049$	0.0394 ± 0.0198	0.0177 ± 0.0049	-0.007 ± 0.026	$20/0/0$
ATUS					
SIMPLE	$0.306 \pm 0.065 \to 0.303 \pm 0.061$	0.0030 ± 0.0116	0.0021 ± 0.0021	-0.001 ± 0.006	$7/11/2$
APS	$0.565 \pm 0.071 \to 0.552 \pm 0.069$	0.0128 ± 0.0123	0.0055 ± 0.0069	0.000 ± 0.005	$17/2/1$
RAPS	$0.478 \pm 0.051 \to 0.472 \pm 0.048$	0.0063 ± 0.0165	0.0056 ± 0.0053	0.000 ± 0.011	$9/7/4$
SAPS	$0.181 \pm 0.043 \to 0.179 \pm 0.043$	0.0024 ± 0.0068	0.0061 ± 0.0068	0.000 ± 0.009	$8/7/5$

Take-away. The set-size-targeted variant is feasible under the prespecified constraints. On the test split, the worst set-size gap improves, ties, or worsens in 234, 62, and 24 of the 320 resplits, respectively.

Table 25: Cross-metric changes under the two Tilted Mondrian targets. Entries are mean ± SD over 20 fixed-model resplits. Under coverage-targeted mitigation, we report the overall and maximum-cell set-size changes. Under set-size-targeted mitigation, we report the overall and maximum-cell coverage changes in percentage points.

	Coverage-targeted mitigation		Set-size-targeted mitigation	
Score	$\Delta\bar{M}$	$\max_{g,z} \lvert\Delta\bar{M}_{g,z}\rvert$	ΔCov. (pp)	$\max_{g,z} \lvert\Delta c_{g,z}\rvert$ (pp)
Bias in Bios				
SIMPLE	0.0038 ± 0.0034	0.104 ± 0.068	-0.001 ± 0.013	0.352 ± 0.551
APS	0.2955 ± 0.2220	0.887 ± 0.443	-0.019 ± 0.092	18.786 ± 10.999
RAPS	0.4799 ± 0.0959	1.255 ± 0.266	-0.028 ± 0.089	12.836 ± 9.650
SAPS	0.0071 ± 0.0036	0.137 ± 0.050	-0.001 ± 0.021	1.601 ± 3.629
Folktables occupation				
SIMPLE	0.0040 ± 0.0023	0.162 ± 0.033	0.009 ± 0.022	0.399 ± 0.101
APS	-0.0915 ± 0.0128	0.548 ± 0.059	0.000 ± 0.002	0.019 ± 0.045
RAPS	-0.0455 ± 0.0063	0.391 ± 0.040	0.008 ± 0.019	0.475 ± 0.115
SAPS	0.0057 ± 0.0033	0.177 ± 0.060	0.006 ± 0.022	0.485 ± 0.166
Folktables income				
SIMPLE	-0.0002 ± 0.0015	0.072 ± 0.031	0.000 ± 0.019	0.711 ± 0.318
APS	-0.0053 ± 0.0051	0.066 ± 0.047	0.002 ± 0.016	0.314 ± 0.556
RAPS	-0.0057 ± 0.0049	0.055 ± 0.028	0.002 ± 0.016	0.314 ± 0.556
SAPS	0.0008 ± 0.0011	0.054 ± 0.023	-0.007 ± 0.026	0.707 ± 0.307
ATUS				
SIMPLE	0.0015 ± 0.0024	0.149 ± 0.256	-0.001 ± 0.006	0.039 ± 0.080
APS	-0.0048 ± 0.0046	0.171 ± 0.139	0.000 ± 0.005	0.111 ± 0.211
RAPS	-0.0005 ± 0.0035	0.160 ± 0.139	0.000 ± 0.011	0.111 ± 0.206
SAPS	0.0009 ± 0.0016	0.163 ± 0.204	0.000 ± 0.009	0.073 ± 0.182

Take-away. Aggregate stability does not imply cellwise stability. Either mitigation target can redistribute the other quantity across finer strata, highlighting the fundamental trade-off between coverage and set sizes.

Table 26: Test set-size absolute change under both mitigation targets. For each resplit, we compute the median and maximum of $|\bar{M}^{\theta}_{g,z} - \bar{M}^{0}_{g,z}|$ over cells; entries are mean $\pm$ SD across 20 resplits.

	Coverage-targeted mitigation		Set-size-targeted mitigation	
Score	median cell $\lvert\Delta\bar{M}\rvert$	max cell $\lvert\Delta\bar{M}\rvert$	median cell $\lvert\Delta\bar{M}\rvert$	max cell $\lvert\Delta\bar{M}\rvert$
Bias in Bios				
SIMPLE	0.0048 ± 0.0026	0.104 ± 0.068	0.0000 ± 0.0001	0.005 ± 0.007
APS	0.1407 ± 0.1041	0.887 ± 0.443	0.0720 ± 0.0391	1.191 ± 0.116
RAPS	0.1920 ± 0.1535	1.255 ± 0.266	0.0535 ± 0.0253	0.939 ± 0.117
SAPS	0.0074 ± 0.0035	0.137 ± 0.050	0.0007 ± 0.0011	0.016 ± 0.035
Folktables occupation				
SIMPLE	0.0788 ± 0.0153	0.162 ± 0.033	0.0115 ± 0.0016	0.032 ± 0.006
APS	0.0625 ± 0.0121	0.548 ± 0.059	0.0001 ± 0.0002	0.005 ± 0.004
RAPS	0.0394 ± 0.0083	0.391 ± 0.040	0.0100 ± 0.0060	0.066 ± 0.018
SAPS	0.0752 ± 0.0190	0.177 ± 0.060	0.0124 ± 0.0030	0.042 ± 0.012
Folktables income				
SIMPLE	0.0199 ± 0.0100	0.072 ± 0.031	0.0093 ± 0.0053	0.041 ± 0.020
APS	0.0176 ± 0.0104	0.066 ± 0.047	0.0036 ± 0.0073	0.028 ± 0.047
RAPS	0.0170 ± 0.0064	0.055 ± 0.028	0.0036 ± 0.0073	0.028 ± 0.047
SAPS	0.0180 ± 0.0075	0.054 ± 0.023	0.0112 ± 0.0041	0.045 ± 0.017
ATUS				
SIMPLE	0.0033 ± 0.0033	0.149 ± 0.256	0.0002 ± 0.0004	0.006 ± 0.012
APS	0.0319 ± 0.0285	0.171 ± 0.139	0.0004 ± 0.0007	0.021 ± 0.019
RAPS	0.0201 ± 0.0197	0.160 ± 0.139	0.0007 ± 0.0011	0.015 ± 0.018
SAPS	0.0058 ± 0.0067	0.163 ± 0.204	0.0004 ± 0.0009	0.010 ± 0.009

Take-away. Tilted Mondrian for set sizes is typically more size-conservative, but coverage is less protected: the worst stratum coverage gap worsens in $130/320$ set-size-targeted mitigations versus $54/320$ coverage-targeted mitigations, with $25/320$ versus $0/320$ worsenings exceeding 5 percentage points.