On the Burden of Achieving Fairness in Conformal Prediction

Ziang Gao[*,†]
McGill University
Montreal, Canada
ziang.gao@mail.mcgill.ca

Pengqi Liu[*]
McGill University
Montreal, Canada
pengqi.liu@mail.mcgill.ca

Archer Yi Yang
McGill University
Montreal, Canada
archer.yang@mcgill.ca

Mouloud Belbahri
TD Insurance
Montreal, Canada
mouloud.belbahri@td.com

Jesse C. Cresswell
Layer 6 AI
Toronto, Canada
jesse@layer6.ai

Masoud Asgharian
McGill University
Montreal, Canada
masoud.asgharian2@mcgill.ca

Abstract

Conformal prediction is often calibrated with a single pooled threshold, but this can hide cross-group heterogeneity in score distributions and distort group-wise coverage. We study this phenomenon through the population score distributions underlying split conformal calibration. First, we derive a conservation law and lower bound showing that pooled calibration incurs irreducible group-wise coverage distortion at a scale set by cross-group quantile heterogeneity. Second, we demonstrate that the two leading fairness definitions for conformal prediction, Equalized Coverage and Equalized Set Size, are fundamentally in tension. Third, we quantify the cost of moving between policies which treat groups separately or pool them. Experiments on synthetic and real data confirm the same bidirectional trade-off after finite-sample calibration. Our results show that, for the policy families studied here, calibration choice does not remove cross-group heterogeneity; it determines whether the resulting distortion appears in the coverage or size dimension, providing a principled lens for analyzing fairness-oriented calibration choices in practice.

1 Introduction

Conformal prediction (CP) [27, 22] provides a model-agnostic solution for quantifying uncertainty over machine learning predictions. We focus throughout on classification with a finite label space $\mathcal{Y}$. Let Y be the outcome variable and $X \in \mathbb{R}^p$ be a vector of features. Given a calibration dataset $\{(X_i, Y_i)\}_{i=1}^n$, CP generates *prediction sets* $\hat{C}(X_{n+1})$ for test label Y_{n+1}, with a *coverage* guarantee that the true label is in the set with user-specified probability $1-\alpha$ for $\alpha \in (0,1)$, i.e., $\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1-\alpha$. Smaller sets indicate less uncertainty about the prediction for Y_{n+1}, so average set size is commonly used as a metric given a fixed coverage level $1-\alpha$. CP is extremely versatile in applications since the coverage guarantee is valid in finite samples and is distribution-free, assuming only that test data is exchangeable with the calibration dataset.

CP is often implemented with a single threshold calibrated on the pooled calibration set. This delivers marginal coverage over the entire distribution, but still allows some groups within the data to be under-covered if others are over-covered to compensate. Exact distribution-free conditional coverage is impossible in general [8]. We therefore study group-conditional CP as an intermediate target. Group-wise calibration resolves the potential disparity in coverage relative to pooled calibration [26],

[*]Equal contribution. Alphabetically ordered by last name.
[†]Corresponding author. Email: ziang.gao@mail.mcgill.ca.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

but the same heterogeneity reappears as disparity in expected set size. Thus, a calibration policy does not remove heterogeneity; it determines whether the disparity appears in coverage or in set size.

In this work, we present a theoretical and empirical study of the fundamental tension between coverage disparity and set size disparity when applying CP over data containing heterogeneous groups. Our theoretical results give practitioners a principled way to understand what is gained and what is sacrificed when one fairness-oriented calibration objective is chosen over another. We study this question through the population score distributions underlying split conformal prediction. This viewpoint separates the structural effect of group heterogeneity from finite-sample calibration noise.

Our contributions are as follows:

- We characterize pooled conformal calibration through a conservation law and lower bound for group-wise coverage, showing that pooled calibration can hide nontrivial group-level distortion.
- We establish a bidirectional impossibility result showing that *exact group-wise coverage and equalized expected set size cannot, in general, be achieved simultaneously.* This identifies a structural limitation of fairness-oriented conformal calibration policies.
- We quantify the costs of switching from pooled calibration to group-wise calibration, and from the coverage-calibrated setting to equalized-size calibration.

The rest of the paper is organized as follows. Section 2 reviews related work. Sections 3 and 4 develop our theoretical results. Section 5 translates the theory into empirical results on synthetic and real datasets. Section 6 concludes with limitations and future directions. Technical discussions, supplemental results, proofs, and additional experimental details are deferred to the appendices.

2 Related Work

A central line of work studies fairness notions for CP. **Equalized Coverage** [20] identifies unequal group-wise empirical coverage under pooled calibration as a fairness concern, and later work extends this perspective to adaptively selected groups and more general frameworks [30, 25]. Other works extend additional fairness notions to conformal prediction, including demographic parity [17], equal opportunity [28], and counterfactual fairness criteria [12]. A related literature studies fairness in downstream decision-making, where conformal sets are provided as a decision aid: conformal sets can improve human decisions [4], but human-subject experiments show that enforcing Equalized Coverage can worsen downstream fairness whereas **Equalized Set Size** can improve it [5]; Liu et al. [18] extend this perspective with an LLM-in-the-loop evaluator. Tasar [24] studies a trade-off between coverage parity and deferral parity in a binary human-in-the-loop setting. Our work is complementary to this literature: we formalize a structural incompatibility between Equalized Coverage and Equalized Set Size, and provide practitioners with a clear lens for weighing the costs of these disparities. The closest conceptual precedent to our results is the classical algorithmic-fairness literature on scores and classifiers: Hardt et al. [14] introduce equalized odds and equal opportunity; Chouldechova [3] and Kleinberg et al. [15] establish impossibility results for competing fairness criteria under unequal base rates; Lazar Reich and Vijaykumar [16] identify settings where partial reconciliation is possible. Our work identifies analogous impossibility results in CP, where the structural driver comes from cross-group heterogeneity in conformal quantiles. More broadly, our work is also related to limits of conditional coverage [8, 11] and efficiency-oriented work on volume optimality for structured prediction sets [10] and learned size-coverage trade-offs [2], but differs in isolating how cross-group heterogeneity constrains the simultaneous attainment of two group-level calibration objectives: exact group-wise coverage and equalized expected set size.

Our focus on Equalized Coverage and Equalized Expected Set Size is deliberate. Both are fundamental procedural criteria proposed and studied in previous CP work. Importantly, we do not regard either criterion as universally fair: which criterion is substantively appropriate depends on how prediction sets enter the downstream decision process. Rather than advocating one criterion over the other, we discover a precise bidirectional characterization of the tension between Equalized Coverage and Equalized Expected Set Size and quantify the disparity induced when either criterion is prioritized.

3 Pooled Calibration and Groupwise Coverage Distortion

This section is organized around two questions. What does a single pooled threshold guarantee? How large is the resulting group-wise coverage distortion when group-specific quantiles differ? We

first recall the finite-sample CP constructions to fix notation, then introduce the population-level objects that help answer these questions.

3.1 Conformal Prediction

We work in the standard setting of split conformal prediction [22]. Let $X \in \mathbb{R}^p$ be a covariate vector, Y be a label, and $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be any measurable real-valued nonconformity score, with larger values indicating greater nonconformity. Given a calibration sample $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$, let $S_i = S(X_i, Y_i)$ for $i = 1, \ldots, n$ denote the calibration scores. Pooled split CP computes an empirical quantile at level $1 - \alpha$ over the entire calibration sample as

$$\hat{q} := \text{Quantile}(1 - \alpha; \{S_i\}_{i=1}^n \cup \{\infty\}) \tag{1}$$

where the (pooled) quantile is equivalently the $\lceil (n+1)(1-\alpha) \rceil$-th order statistic of the augmented multiset. For any threshold t, we define the associated prediction set as the collection of labels whose scores fall below the threshold

$$\hat{C}_t(X_{n+1}) := \{y : S(X_{n+1}, y) \leq t\}. \tag{2}$$

The CP set for a new test point is thus given by $\hat{C}_{\hat{q}}(X_{n+1})$, which, under exchangeability of the calibration and test examples, has the marginal coverage guarantee $\mathbb{P}(Y_{n+1} \in \hat{C}_{\hat{q}}(X_{n+1})) \geq 1 - \alpha$.

On the other hand, to construct a prediction set with a group-wise coverage guarantee [26], we define $G : \mathcal{X} \to \mathcal{G} = \{1, \ldots, K\}$ to be a prespecified discrete partition of the covariate space, with $\mathbb{P}\{G(X) = g\} > 0$ for every $g \in \mathcal{G}$, and group label $G(X) = g$, and choose threshold t in Equation (2) to be the $1 - \alpha$ quantile (group-wise) of calibration scores within group g, i.e. $\hat{q}_g := \text{Quantile}(1-\alpha; \{S_i\}_{i:G(X_i)=g} \cup \{\infty\})$. The prediction set $\hat{C}_{\hat{q}_g}(X_{n+1})$, under exchangeability within each prespecified group, satisfies $\mathbb{P}\{Y_{n+1} \in \hat{C}_{\hat{q}_g}(X_{n+1}) \mid G(X_{n+1}) = g\} \geq 1 - \alpha$ for a test point in group g. Appendix A Table 1, summarizes the main population quantities used below.

3.2 Conservation Law for the Pooled Threshold

We first investigate the group-wise miscoverage of a marginal prediction set. The standard split CP set $\hat{C}_{\hat{q}}(X)$ with the pooled empirical quantile $\hat{q}$ only has a marginal coverage guarantee rather than a group-wise one: some groups may be under-covered while others are over-covered. To study the group-wise miscoverage of $\hat{C}_{\hat{q}}(X)$, we define the *signed group-wise coverage distortion* for a test point in group g by

$$\varepsilon_g(\hat{q}) := \mathbb{P}\{Y \in \hat{C}_{\hat{q}}(X) \mid G(X) = g\} - (1 - \alpha). \tag{3}$$

A positive $\varepsilon_g(\hat{q})$ corresponds to over-coverage in group g, while a negative value corresponds to under-coverage. Under standard quantile-consistency conditions, the empirical pooled quantile $\hat{q}$ concentrates around the pooled population quantile q of the nonconformity score S:

$$q := \inf\{t \in \mathbb{R} : F_S(t) \geq 1 - \alpha\} \tag{4}$$

with $F_S(t) = \mathbb{P}(S \leq t)$. For any fixed threshold t, we can see that the group-wise coverage is the conditional CDF of the nonconformity score S:

$$\mathbb{P}\{Y \in \hat{C}_t(X) \mid G(X) = g\} = \mathbb{P}\{S \leq t \mid G(X) = g\} =: F_{S|g}(t). \tag{5}$$

In the large calibration sample case, $\varepsilon_g(\hat{q})$ can be approximated by the *population signed group-wise coverage distortion* $\varepsilon_g(q)$, defined as:

$$\varepsilon_g(q) := F_{S|g}(q) - (1 - \alpha). \tag{6}$$

We begin by characterizing what the pooled threshold q guarantees at the aggregate level for $\varepsilon_g(q)$. By definition of q (Equation (4)), Theorem 1 below gives the aggregate identity for group-wise coverage distortion; when F_S is continuous at q, this identity becomes a zero-sum conservation law. (All proofs are presented in Appendix B.)

Theorem 1. *(Conservation law for pooled calibration) Let $p_g := \mathbb{P}\{G(X) = g\}$ be the group mass. Group-wise coverage distortion $\varepsilon_g(q)$ under the pooled threshold satisfies the aggregate identity*

$$\sum_{g \in \mathcal{G}} p_g \, \varepsilon_g(q) = F_S(q) - (1 - \alpha) =: \delta(q). \tag{7}$$

If F_S is continuous at q, then $\delta(q) = 0$, and hence $\sum_{g \in \mathcal{G}} p_g \, \varepsilon_g(q) = 0$.

Remark 1. If F_S has a jump at q, randomized tie handling via the probability integral transform restores the zero-sum identity. Let $U \sim \mathrm{Uniform}(0,1)$ be independent of S and G, and define $Z := F_S(S^-) + U \cdot (F_S(S) - F_S(S^-))$. Then $Z \sim \mathrm{Uniform}(0,1)$, which means that the event $\{Z \leq 1-\alpha\}$ implements randomized tie handling at q. Next, we define $\tilde{\varepsilon}_g := \mathbb{P}(Z \leq 1-\alpha \mid G = g) - (1-\alpha)$ to obtain $\sum_{g\in\mathcal{G}} p_g \tilde{\varepsilon}_g = 0$.

Take-away. A pooled threshold q can satisfy marginal coverage while miscovering individual groups, which can be a fairness concern. Under continuity, or under randomized tie handling in the atomic case, Theorem 1 shows that weighted over-coverage in some groups is exactly balanced by weighted under-coverage in others.

Theorem 1 is an additive form. A product-type form of the conservation law is given in Appendix A.1.

3.3 The Pooled Threshold Uncertainty Relation

We now move from balance to magnitude. For each group, let $q_g := \inf\{t \in \mathbb{R} : F_{S|g}(t) \geq 1-\alpha\}$ denote the population group quantile. We quantify the irreducible root mean square (RMS) group-wise miscoverage induced by using a pooled threshold in the presence of group-quantile heterogeneity.

Assumption 1. There exists a common $\eta > 0$ such that, for each $g \in \mathcal{G}$, the conditional CDF $F_{S|g}$ is absolutely continuous on the interval $I_g := [\min(q, q_g) - \eta, \max(q, q_g) + \eta]$, with density $f_{S|g}$ satisfying $f_{S|g}(t) > 0$ for almost every $t \in I_g$. Moreover, whenever $q \neq q_g$, $\operatorname{ess\,inf}_{t\in J_g} f_{S|g}(t) > 0$, where $J_g := [\min(q, q_g), \max(q, q_g)]$. In particular,

$$m_g(q) := \operatorname*{ess\,inf}_{t\in J_g} f_{S|g}(t), \tag{8}$$

with the convention $m_g(q) = 0$ when $q = q_g$.

Let q_G and $\varepsilon_G(q)$ denote the random variable versions of q_g and $\varepsilon_g(q)$, respectively, and

$$F_{S|g}(q) - F_{S|g}(q_g) = \int_{q_g}^{q} f_{S|g}(t)\, dt. \tag{9}$$

Definition 1. The *effective stiffness* is defined as

$$m_{\mathrm{eff}}(q) := \sqrt{\frac{\mathbb{E}[m_G(q)^2\,(q-q_G)^2]}{\mathbb{E}[(q-q_G)^2]}} \quad \text{whenever } \mathbb{E}[(q-q_G)^2] > 0. \tag{10}$$

Let

$$\sigma_\Delta^2 := \mathrm{Var}(q_G) = \mathbb{E}\big[(q_G - \mathbb{E}[q_G])^2\big] = \sum_{g\in\mathcal{G}} p_g (q_g - \sum_{g'\in\mathcal{G}} p_{g'} q_{g'})^2. \tag{11}$$

The quantity σ_Δ measures the cross-group dispersion of the target quantiles $\{q_g\}$. When $\sigma_\Delta = 0$, pooled and group-conditional calibration coincide at the population level. As a property of the population and choice of nonconformity score, σ_Δ, which we refer to as *intrinsic heterogeneity*, is a central object in our study of the differences between pooled and group-wise conformal calibration.

Theorem 2. *(Pooled-threshold uncertainty relation) Suppose Assumption 1 holds. Then*

$$\mathrm{Var}(\varepsilon_G(q)) \geq m_{\mathrm{eff}}(q)^2 \mathrm{Var}(q_G). \tag{12}$$

Take-away. Under pooled calibration, the cross-group variation in the target quantiles induces a non-zero cross-group disparity in group-wise coverage distortions at the scale $m_{\mathrm{eff}}(q)\sigma_\Delta$.

Thus, the fact that a pooled threshold causes group-wise coverage distortion is not an artifact of finite-sample calibration noise, but is already present at the population level.

4 Calibration Policies and Coverage–Size Trade-offs

4.1 Equalized Coverage versus Equalized Expected Set Size

The results above quantify the irreducible coverage error incurred by a single pooled threshold q. We now ask what happens when one changes the calibration policy to mitigate pooled threshold distortion. We formalize two directions of the resulting trade-off between two group-level fairness criteria: *Equalized Coverage* [20], meaning $\mathbb{P}(S \leq t_g \mid G = g) = 1-\alpha$ for all g, and *Equalized Expected Set Size* [4], meaning the expected size of the CP set under a group-specific threshold t_g is equal to a constant λ, i.e. $\mathbb{E}[|\hat{C}_{t_g}(X)| \mid G = g] = \lambda$ for all g.

Recall that, for each group $g \in \mathcal{G}$, the threshold $q_g := F^{-1}_{S|g}(1-\alpha)$ achieves exact group-wise coverage level $1-\alpha$ whenever $F_{S|g}$ is continuous (Equation (5)). We first present one sufficient condition, going from equalized coverage to expected-size disparity. To do so, fix a reference group $r \in \mathcal{G}$ and define $\mathcal{H}_r := \{g \in \mathcal{G} : q_g \geq q_r\}$, the set of groups whose coverage-calibrated thresholds are at least as large as that of the chosen reference group. Let $\ell_g(t) := \mathbb{E}[|\hat{C}_t(X)| \mid G = g]$ be the group-wise expected size of the CP set at threshold t. For a reference group r, define the restricted mean squared size gap

$$D_r^2 := \mathbb{E}\big[(\ell_G(q_G) - \ell_r(q_r))^2 \mathbf{1}\{G \in H_r \setminus \{r\}\}\big]. \tag{13}$$

Lemma 1. *(Monotonicity of the expected set size) For any $t_1 \leq t_2$, $\hat{C}_{t_1}(x) \subseteq \hat{C}_{t_2}(x)$ for every x,*

$$\ell_g(t_1) \leq \ell_g(t_2) \quad \textit{for every } g \in \mathcal{G}. \tag{14}$$

Indeed, $S(x,y) \leq t_1 \leq t_2$ implies that every label included at t_1 is also included at t_2; taking conditional expectations gives the second statement. While Lemma 1 establishes within-group monotonicity of expected set size, it does not guarantee a disparity in expected set size across groups under exact group-wise coverage. To see this, for any $q_g \geq q_r$, the coverage-calibrated size difference decomposes as

$$\ell_g(q_g) - \ell_r(q_r) = \underbrace{\ell_g(q_g) - \ell_g(q_r)}_{\text{within-group expansion}} + \underbrace{\ell_g(q_r) - \ell_r(q_r)}_{\text{cross-group size difference}}. \tag{15}$$

By Lemma 1, the first term is nonnegative, whereas the second term can have either sign and may offset the first term. To rule out such compensation, we impose the following sufficient condition.

Assumption 2. There exists a reference group $r \in \mathcal{G}$ with $\mathcal{H}_r \setminus \{r\} \neq \varnothing$, and positive constants $\{c_g\}_{g \in \mathcal{H}_r \setminus \{r\}}$, such that $\ell_g(q_r) - \ell_r(q_r) \geq c_g$ for every $g \in \mathcal{H}_r \setminus \{r\}$.

Assumption 2 provides a sufficient no-compensation condition rather than a necessary characterization, and it does not assume that the final coverage-calibrated expected set sizes are already unequal. A natural regime captured by this assumption occurs when greater true-label calibration difficulty is aligned with broader candidate-label ambiguity: a harder-to-predict group may require a larger threshold to attain the target coverage, while less separated candidate-label scores cause more competing labels to enter its prediction set at the same reference threshold. Under Assumption 2, the following theorem establishes that exact group-wise coverage necessarily induces cross-group disparity in expected set size.

Theorem 3. *(Equalized coverage induces cross-group expected set size disparity) Assume the score CDF $F_{S|g}$ is continuous $\forall g \in \mathcal{G}$. Suppose Assumption 2 holds for some reference group $r \in \mathcal{G}$. Then,*

1. *The group-wise thresholds $\{q_g\}_{g \in \mathcal{G}}$, which achieve an exact group-wise coverage level $1-\alpha$, necessarily induce a nonzero cross-group disparity in expected set size. More precisely, for every $g \in \mathcal{H}_r \setminus \{r\}$, $\ell_g(q_g) - \ell_r(q_r) \geq c_g > 0$. Consequently,*

$$\max_{g,g' \in \mathcal{G}} |\ell_g(q_g) - \ell_{g'}(q_{g'})| \geq \max_{g \in \mathcal{H}_r \setminus \{r\}} c_g > 0. \tag{16}$$

   Therefore, exact group-wise coverage cannot simultaneously satisfy equalized expected set size across groups.
2. *The restricted mean squared cross-group size disparity relative to the reference group r satisfies*

$$D_r^2 = \sum_{g \in \mathcal{H}_r \setminus \{r\}} p_g (\ell_g(q_g) - \ell_r(q_r))^2 \geq \sum_{g \in \mathcal{H}_r \setminus \{r\}} p_g c_g^2 > 0. \tag{17}$$

Take-away. Exact group-wise coverage generally entails unequal expected set sizes across groups.

We now turn to the reverse direction and consider an equalized expected set size policy at target level λ. We define $\lambda_g := \ell_g(q_g),\ g \in \mathcal{G}$, the coverage-calibrated expected set size of group g, and we denote λ_G as the random-group version of λ_g. We ask how an equalized-size policy alters group-wise coverage. Let τ_g denote a group-specific threshold satisfying $\ell_g(\tau_g) = \lambda$. That is, τ_g is the threshold required for group g to attain the equalized-size target. To quantify the coverage effect of moving from q_g to τ_g, we impose local regularity only on the segment between q_g and τ_g.

Assumption 3. (Local regularity for equalized-size perturbations) For each group g, on the segment between q_g and τ_g both $F_{S|g}$ and ℓ_g are absolutely continuous, with derivatives satisfying $f_{S|g}(t) \geq m_g > 0$, $|\ell_g'(t)| \leq V_g$ for almost every t on the segment, where $0 < V_g < \infty$.

Theorem 4 below establishes the reverse direction of the trade-off compared to Theorem 3. Under an equalized-size policy $\{\tau_g\}_{g\in\mathcal{G}}$, one may eliminate cross-group disparity in expected set size, but only at the expense of introducing cross-group disparity in coverage whenever the common target level λ lies strictly between two distinct coverage-calibrated set sizes. We utilize the local coverage–size conversion factor $\kappa_g := m_g/V_g$ for each group g.

Theorem 4. *(Equalized expected set size induces cross-group coverage disparity) Suppose Assumption 3 holds, and suppose for some common target level λ the thresholds $\{\tau_g\}_{g\in\mathcal{G}}$ satisfy $\ell_g(\tau_g) = \lambda,\ g \in \mathcal{G}$. Then the following results hold.*

(i) *For every group g with $\lambda_g < \lambda$, $F_{S|g}(\tau_g) - (1-\alpha) \ \geq\ \kappa_g\,(\lambda - \lambda_g)$.*

(ii) *For every group g' with $\lambda_{g'} > \lambda$, $(1-\alpha) - F_{S|g'}(\tau_{g'}) \ \geq\ \kappa_{g'}\,(\lambda_{g'} - \lambda)$.*

Consequently, for any pair $g, g' \in \mathcal{G}$ such that $\lambda_g < \lambda < \lambda_{g'}$, we have the pairwise lower bound

$$F_{S|g}(\tau_g) - F_{S|g'}(\tau_{g'}) \ \geq\ \kappa_g(\lambda - \lambda_g) + \kappa_{g'}(\lambda_{g'} - \lambda). \tag{18}$$

In particular, if there exist groups g, g' with $\lambda_g < \lambda_{g'}$ and $\lambda \in (\lambda_g, \lambda_{g'})$, then

$$\max_{a,b\in\mathcal{G}} |F_{S|a}(\tau_a) - F_{S|b}(\tau_b)| \ \geq\ \kappa_g(\lambda - \lambda_g) + \kappa_{g'}(\lambda_{g'} - \lambda) \ >\ 0. \tag{19}$$

Take-away. Whenever the equalized-size level λ lies strictly between two distinct coverage-calibrated set sizes, the equalized set size policy induces nonzero coverage disparity.

Together, Theorem 3 and Theorem 4 show that **the two leading fairness notions in CP, namely Equalized Coverage and Equalized Set Size, cannot in general be achieved simultaneously**. To the best of our knowledge, this is the first formal result in CP establishing this structural incompatibility. This places our result alongside the classical impossibility theorems of Kleinberg et al. [15] and Chouldechova [3], but in the distinct setting of CP, where the trade-off is driven by cross-group heterogeneity in conformal quantiles and its effect on coverage and set size. We next quantify the scale of policy-conversion distortions.

4.2 Quantitative Bounds for Policy-Conversion Distortions

We now turn from directional trade-off results to quantitative policy-conversion bounds. The goal is to measure the scale of the distortions induced when one moves between pooled calibration, exact group-wise coverage, and equalized-size calibration. We first quantify the cost of moving from pooled calibration to exact group-wise coverage. To do so, we require a local responsiveness condition on the group-wise size curves along the segment between q and q_g.

Assumption 4. (Local set size responsiveness for group-wise calibration) For each group g, the mapping $t \mapsto \ell_g(t)$ is absolutely continuous on the segment between q and q_g, and is non-decreasing on the segment. Moreover, its derivative $\ell_g'(t)$ is defined almost everywhere and satisfies $\ell_g'(t) \geq v_g > 0$ for almost every t on that segment.

Next, we define the *effective set-size responsiveness* as

$$v_{\text{eff}}(q) := \sqrt{\frac{\mathbb{E}[v_G^2\,(q - q_G)^2]}{\mathbb{E}[(q - q_G)^2]}} \quad \text{whenever } \mathbb{E}[(q - q_G)^2] > 0. \tag{20}$$

The next corollary converts the intrinsic heterogeneity σ_Δ into a lower bound on set size distortion.

Corollary 1. *(Exact group-wise coverage induces set size distortion) Suppose Assumption 4 holds. Let the pooled threshold q be replaced by the group-wise thresholds $\{q_g\}$. Then the RMS set size lower bound is*

$$\sqrt{\mathbb{E}[(\ell_G(q_G) - \ell_G(q))^2]} \ \geq\ v_{\text{eff}}(q)\,\sigma_\Delta. \tag{21}$$

Take-away. Exact group-wise coverage generally incurs a measurable expected set size distortion.

Corollary 1 turns the directional statement into a scale statement. It shows that moving from pooled calibration to exact group-wise coverage produces a set size change with magnitude controlled by the same underlying cross-group quantile distortion.

We next quantify the reverse conversion cost. Theorem 4 establishes the reverse direction of the trade-off by showing that an equalized-size policy induces coverage disparity. The following corollary converts the same argument into an aggregate RMS lower bound by measuring how far the equalized-size thresholds $\{\tau_g\}$ shift coverage from the exact group-wise benchmark $\{q_g\}$. From Theorem 4,

we obtain $\mathbb{E}\big[(F_{S|G}(\tau_G) - F_{S|G}(q_G))^2\big] \geq \mathbb{E}\big[\kappa_G^2(\lambda_G - \lambda)^2\big]$, $\kappa_G := m_G/V_G$. Before we formally state this result, we define the *effective coverage responsiveness*

$$\kappa_{\text{eff}}(\lambda) := \sqrt{\frac{\mathbb{E}[(m_G/V_G)^2(\lambda - \lambda_G)^2]}{\mathbb{E}[(\lambda - \lambda_G)^2]}} \quad \text{whenever} \quad \mathbb{E}[(\lambda - \lambda_G)^2] > 0, \tag{22}$$

and the set-size heterogeneity $\sigma_\lambda := \text{sd}(\lambda_G) = \sqrt{\text{Var}(\lambda_G)}$.

Corollary 2. *(Equalized expected set size induces coverage distortion) Suppose Assumption 3 holds. Let the group-wise thresholds $\{\tau_g\}_{g\in G}$ satisfy $\ell_g(\tau_g) = \lambda$ for all $g \in G$, so that they enforce equalized expected set size at a common level λ. Then the RMS coverage distortion relative to the exact group-wise coverage satisfies*

$$\sqrt{\mathbb{E}[(F_{S|G}(\tau_G) - F_{S|G}(q_G))^2]} \;\geq\; \kappa_{\text{eff}}(\lambda)\sigma_\lambda. \tag{23}$$

Take-away. Equalizing expected set size generally incurs a measurable coverage cost.

Corollary 2 provides the corresponding quantitative statement for the reverse conversion compared to Corollary 1. The induced coverage distortion under equalized size calibration is governed by the dispersion of the coverage-calibrated set sizes and the local coverage–size sensitivity. Finally, the coverage distortion induced by changing from q to τ_g can also be characterized with local regularity conditions. Let $\ell_g(\tau_g) = \lambda$ for every group g, and define the direct adjustment interval

$$I_g^\circ := [\min\{q, \tau_g\}, \max\{q, \tau_g\}], \tag{24}$$

where $\circ$ denotes the direct move from q to τ_g.

Corollary 3. *(Pooled to Equalized Set Size coverage change) Suppose that, on I_g°, both $F_{S|g}$ and ℓ_g are absolutely continuous and, for almost every $t \in I_g^\circ$, $f_{S|g}(t) \geq m_g^\circ > 0, |\ell_g'(t)| \leq V_g^\circ < \infty$, with $V_g^\circ > 0$. Define $\kappa_g^\circ := m_g^\circ/V_g^\circ$. Then,*

$$\big|F_{S|g}(\tau_g) - F_{S|g}(q)\big| \geq \kappa_g^\circ \,|\lambda - \ell_g(q)| \;\; \textit{for every group } g \tag{25}$$

and

$$\sqrt{\mathbb{E}\Big[\big(F_{S|G}(\tau_G) - F_{S|G}(q)\big)^2\Big]} \geq \sqrt{\mathbb{E}\Big[(\kappa_G^\circ)^2\,(\lambda - \ell_G(q))^2\Big]}. \tag{26}$$

Take-away. Switching from pooled calibration to Equalized Set Size calibration induces a group-wise coverage change whose magnitude is lower-bounded by the required expected set size adjustment.

Taken together, Corollaries 1–3 quantify the cost of moving from pooled calibration to the two fairness policies studied in Section 4.1, as well as between them. Corollary 1 measures how intrinsic heterogeneity reappears as set size distortion when exact group-wise coverage is enforced, while Corollary 2 measures how it reappears as coverage distortion when expected set size is equalized across groups. Corollary 3 additionally quantifies the direct group-wise coverage change from pooled calibration to Equalized Set Size calibration.

5 Experiments

The results in Sections 3 and 4 were derived using population analysis. Our experiments in this section are not designed as a benchmark comparison against fairness-aware conformal methods; rather, they are designed to test whether the distortion transfer pattern remains visible under finite-sample calibration, matching how CP is used in practice. We focus on three finite-sample consequences of Sections 3 and 4: the pooled threshold lower-bound behavior (Theorem 2) with the effective lower-bound scale $m_{\text{eff}}(q)\sigma_\Delta$, the RMS set-size distortion after switching from q to q_g (Corollary 1), and the RMS coverage distortion under equalized expected set size (Corollary 2). Our code is available at `https://github.com/GreenPenguin001/group-cp-tradeoffs`.

5.1 Synthetic Simulations

We generate calibration and test nonconformity scores directly from known group-conditional score distributions, and then apply the same empirical quantile rule used throughout the paper. This lets us compare finite-sample distortions with the corresponding population lower-bound scales from Sections 3 and 4. Figure 1 studies the bidirectional policy conversion with eight equally weighted groups $g_0, \ldots, g_7$, and with Gaussian mixture scores for the coverage-to-size direction, and Student-t/Gaussian mixture scores for size-to-coverage.

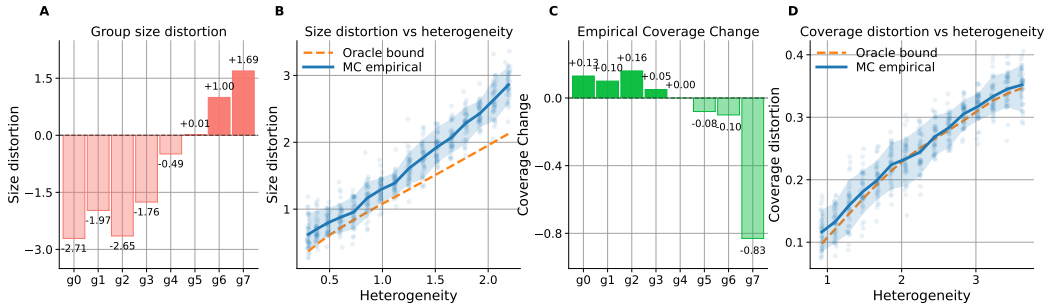


Figure 1: Bidirectional policy conversion in the synthetic study. Panels A–B illustrate the coverage-to-size direction in Theorem 3 and Corollary 1. Panel A shows the signed change in expected set size for each of eight equally weighted groups, and Panel B shows that the aggregate size distortion grows with cross-group score heterogeneity and tracks the oracle lower-bound scale. Panels C–D correspond to the size-to-coverage direction in Theorem 4 and Corollary 2. Panel C reports the signed finite-sample change $\hat{F}_{S|g}(\hat{\tau}_g) - \hat{F}_{S|g}(\hat{q}_g)$ after replacing empirical group-wise thresholds by equalized-size thresholds; hence the displayed values are changes relative to the empirical group-wise benchmark. Panel D shows that the aggregate coverage distortion grows with heterogeneity.

For the policy conversion panels, we use monotone proxy size curves $\ell_g(t) = a_g + b_g t$ to map thresholds to expected set sizes. In Figure 1, Panels A–B show the coverage-to-size direction in Corollary 1, with oracle scale $v_{\text{eff}}(q)\sigma_\Delta$; Panels C–D show the size-to-coverage direction in Corollary 2, with oracle scale $\kappa_{\text{eff}}(\lambda)\sigma_\lambda$, where coverage change is measured using the empirical quantile $\hat{q}_g$. We find that mean distortions in the finite sample case are still bounded by the population analysis results.

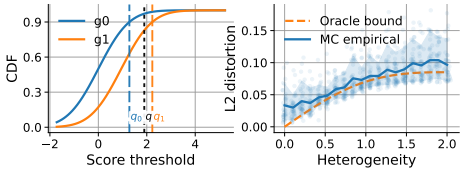


Figure 2: Two-group Gaussian pooled-threshold picture with q lying between q_0 and q_1; across the heterogeneity sweep, the empirical RMS miscoverage remains above the effective lower-bound scale, consistent with Theorem 2.

Figure 2 gives a two-group Gaussian example. To achieve group-wise conditional coverage, one group requires a smaller quantile q_0, whereas the other requires a larger quantile q_1. The two groups have different target thresholds, and the pooled threshold lies between them, so under pooled calibration one group is over-covered and the other is under-covered (Theorem 1). As the separation between q_0 and q_1 increases (increasing intrinsic heterogeneity σ_Δ), the finite-sample mean distortion remains above the oracle effective lower bound (Equation (12)). Together, the synthetic studies support the bidirectional trade-off. All empirical curves use calibration thresholds constructed at target level $\alpha = 0.1$, and the reported distortions are evaluated on independent test samples and averaged over 40 Monte Carlo seeds with equal group weights. Detailed score families, heterogeneity ranges, additional diagnostics, and empirical–oracle ratio results under imbalance are deferred to Appendix C.

5.2 `Bias in Bios` Experiments

We use `Bias in Bios` [6] as a real-data illustration of the policy-conversion mechanism from Sections 3 and 4. We restrict to the ten most frequent professions, two demographic groups (Male and Female), and use a `DistilBERT` classifier [21]. Unless otherwise noted, we use the simple nonconformity score $s(x, y) = 1 - \hat{p}_y(x)$. In Figure 3, Panel A shows empirical score CDFs with the pooled and distinct group-specific thresholds. Panel B shows that equalizing coverage by moving from q to q_g enlarges the cross-group size disparity. Panel C shows the reverse move: equalizing set size requires shifting the thresholds away from q_g, reintroducing coverage distortion. Panel D summarizes the three distortions. At $\alpha = 0.1$, the corresponding RMS quantities, associated with Equations (12), (21) and (23) are 0.0015, 0.0051, and 0.0017, respectively. Thus, `Bias in Bios` supports the pooled-threshold distortion mechanism in Theorem 2 and the two policy-conversion effects quantified in Corollaries 1 and 2. Detailed per-group quantities are deferred to Table 8. Additional robustness experiments, including alternative score comparisons and finite-calibration diagnostics, are reported in Appendix D.

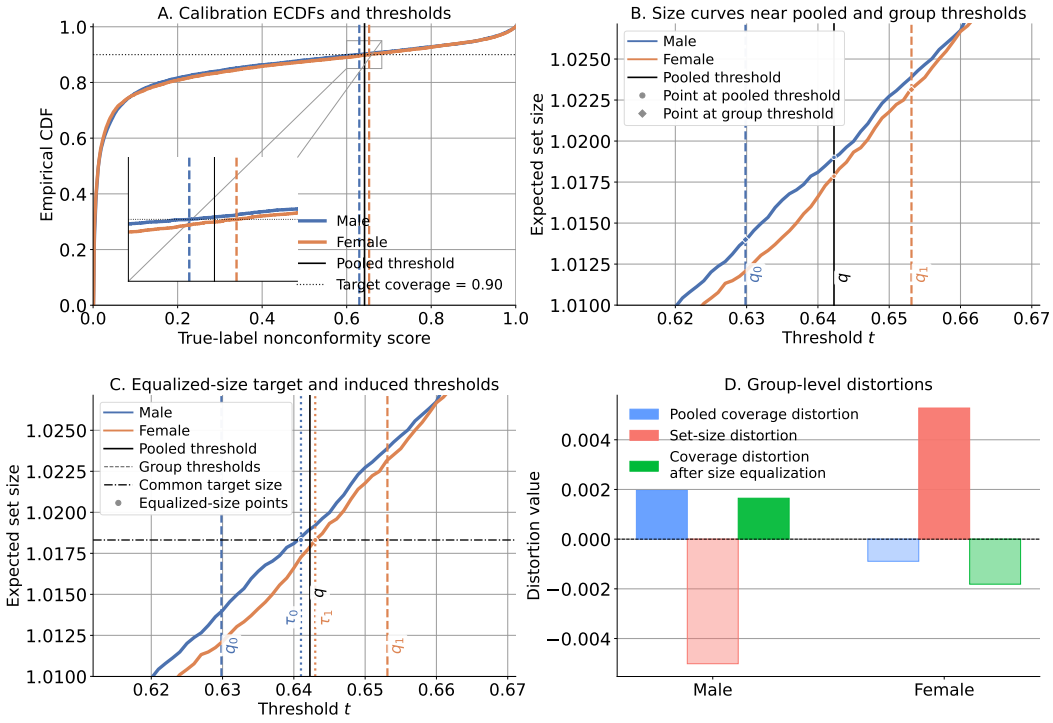


Figure 3: `Bias in Bios` mechanism view at $\alpha = 0.1$ for the simple score. Panel A illustrates the pooled-threshold mechanism in Theorem 1; Panels B–C illustrate Theorems 3–4 and Corollaries 1–2; Panel D summarizes the three distortions (Theorem 2, Corollaries 1–2) for male and female groups.

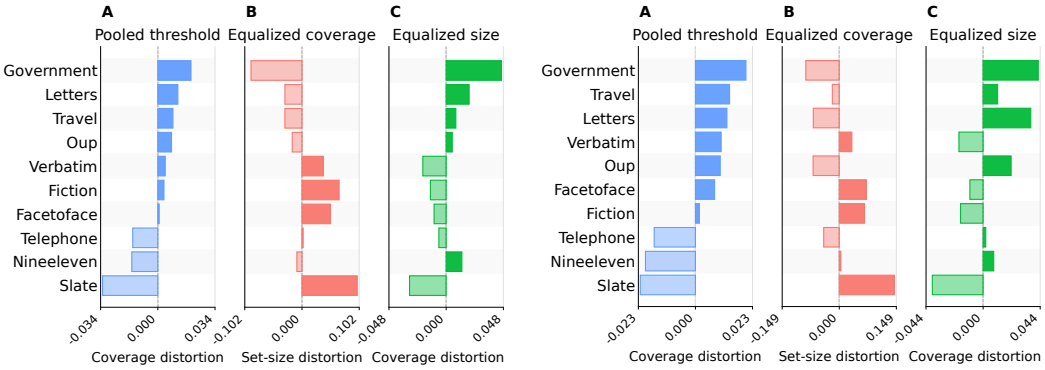


Figure 4: `MultiNLI` at $\alpha = 0.1$ with simple (left) and RAPS (right) scores. For each score, Panel A shows signed coverage distortion $\hat{F}_{S|g}(\hat{q}) - (1 - \alpha)$ under pooled threshold (Theorem 1): positive bars indicate over-coverage and negative bars indicate under-coverage. Panel B shows the signed change in expected set size $\hat{\ell}_g(\hat{q}_g) - \hat{\ell}_g(\hat{q})$ after switching to group-wise thresholds that equalize coverage (Corollary 1). Panel C shows the signed coverage distortion $\hat{F}_{S|g}(\hat{\tau}_g) - \hat{F}_{S|g}(\hat{q}_g)$ after enforcing a common expected set size (Corollary 2).

5.3 `MultiNLI` Experiments

Using the same post-hoc protocol as in Section 5.2, we treat the ten `MultiNLI` [29] genres as groups. Figure 4 shows the same qualitative transfer pattern (Theorems 3 and 4) for both the simple score, $s(x, y) = 1 - \hat{p}_y(x)$, and the RAPS nonconformity score [1]. For the simple score at $\alpha = 0.1$, the pooled RMS genre-wise coverage distortion is 0.0150. The corresponding RMS set-size distortion after moving from q to q_g is 0.0532. In addition, the RMS coverage distortion under equalized expected set size is 0.0209. Thus, `MultiNLI` exhibits the signed genre-wise pattern and is consistent with the

bidirectional trade-off described in Sections 3 and 4. Per-genre summaries, additional robustness experiments, alternative-score results, and finite-calibration diagnostics are reported in Appendix E.

5.4 FACET Experiments

We next show the same mechanisms on FACET [13] using the RAPS score on the age group split (Younger, Middle, Older, Unknown) with a zero-shot CLIP ViT-L/14 classifier [19]. Figure 5 shows the same transfer pattern on a more group imbalanced computer-vision dataset: under pooled calibration, the Younger group is over-covered while the others are under-covered. Switching from q to q_g removes the pooled-regime distortion but induces set-size distortion. Equalizing expected set size then reintroduces coverage distortion. Specifically, Panel A shows pooled-threshold coverage distortion, Panel B the set-size shift after equalized coverage, and Panel C the coverage shift after equalized expected size. At $\alpha = 0.1$, the empirical pooled RMS coverage distortion is 0.0083. The RMS set-size distortion after changing from q to q_g is 0.1717. The RMS coverage distortion under equalized expected set size is 0.0199. Thus, FACET supports the behavior in Theorem 1 together with the two policy-conversion effects quantified in Corollaries 1 and 2. Per-group summaries, additional robustness experiments, and calibration-resampling stability results are reported in Appendix F.

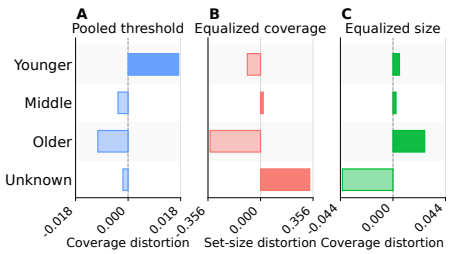


Figure 5: FACET at $\alpha = 0.1$ with the RAPS score; Panel A illustrates Theorem 1, and Panels B–C illustrate Corollaries 1–2.

Taken together, the experiments in this section examine three finite-sample consequences of Sections 3 and 4: the pooled-threshold effective lower-bound behavior, the RMS set-size distortion after switching from q to q_g, and the RMS coverage distortion under equalized expected set size. We use the effective constants only as oracle/proxy diagnostic scales, not as finite-sample estimators for deployment. Appendix A.5 describes the empirical plug-in diagnostics used to compute the heterogeneity scales and lower-bound proxies reported in our experiments. Additional detectability experiments on how large calibration splits should be for the structural floor to become empirically resolvable are deferred to Appendices D.5, E.5, and F.3.

6 Conclusion

We studied group-conditional conformal prediction through the population score distributions underlying split conformal calibration. Our main result is structural: when group-wise conformal quantiles differ, the two group-level objectives studied here, exact group-wise coverage and equalized expected set size, cannot in general be achieved simultaneously. Under a pooled threshold, this heterogeneity appears as a group-wise coverage disparity; enforcing exact group-wise coverage shifts the heterogeneity to expected set size disparity across groups, while equalizing expected set size reintroduces coverage disparity. These results show that the two leading fairness notions in CP, Equalized Coverage and Equalized Set Size, exhibit a structural trade-off, thereby identifying the CP counterpart of classical impossibility results in algorithmic fairness.

Our goal is not to solve full conditional coverage, nor to prescribe a normative fairness criterion. Rather, we identify a population-level constraint induced by heterogeneous group quantiles under conformal calibration. The quantitative bounds in Section 4.2 rely on local regularity assumptions, including continuity of score distributions and nondegenerate local sensitivity of the set size curves. In addition, the analysis is developed at the level of prespecified discrete groups and population score distributions, while finite-sample behavior is studied empirically. These limitations are deliberate: they separate the structural effect of heterogeneity from broader questions about adaptive subgroup validity, end-to-end training, or fully conditional guarantees.

In this sense, a calibration policy should be understood not as removing group heterogeneity, but as determining whether it manifests as cross-group disparity in coverage or set size, and how large the resulting policy-conversion distortions will be. A natural next step is to extend the analysis beyond prespecified discrete groups and population score distributions, and to develop finite-sample theory for the policy-conversion distortions. Another extension is to connect these structural trade-offs to adaptive subgroup validity and to learned calibration policies in practice.

Acknowledgments and Disclosure of Funding









We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) through Discovery Grants held by Masoud Asgharian [RGPIN-2024-05640] and Archer Yi Yang [RGPIN-2024-06780]. This work was also supported in part by research funding held by Archer Yi Yang from the Canadian Statistical Sciences Institute (CANSSI) through its Collaborative Research Teams program and from the Fonds de recherche du Québec (FRQ), Nature et technologies sector (Team Research Project Grant No. 327788). Ziang Gao acknowledges scholarship support from the Institut des sciences mathématiques (ISM).

Competing interests.

The authors declare no additional competing interests beyond their disclosed affiliations.

References

[1] 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.

[2] Francis Bach. A Convex Loss Function for Set Prediction with Optimal Trade-offs Between Size and Conditional Coverage. *arXiv:2512.19142*, 2025.

[3] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017. doi: 10.1089/big.2016.0047.

[4] 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.

[5] Jesse C. Cresswell, Bhargava Kumar, Yi Sui, and Mouloud Belbahri. Conformal prediction sets can cause disparate impact. In *The Thirteenth International Conference on Learning Representations*, 2025.

[6] 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*, page 120–128, 2019. ISBN 9781450361255. doi: 10.1145/3287560.3287572.

[7] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

[8] Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

[9] Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. De finetti's theorem and related results for infinite weighted exchangeable sequences. *Bernoulli*, 30(4): 3004–3028, 2024.

[10] Chao Gao, Liren Shan, Vaidehi Srinivas, and Aravindan Vijayaraghavan. Volume optimality in conformal prediction with structured prediction sets. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 18495–18527, 2025.

[11] 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, 03 2025. ISSN 1369-7412. doi: 10.1093/jrsssb/qkaf008.

[12] Ozgur Guldogan, Neeraj Sarna, Yuanyuan Li, and Michael Berger. Counterfactually fair conformal prediction. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

[13] Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. FACET: Fairness in computer vision evaluation benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20370–20382, 2023.

[14] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29*, pages 3315–3323, 2016.

[15] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference*, volume 67, pages 43:1–43:23, 2017. doi: 10.4230/LIPIcs.ITCS.2017.43.

[16] Claire Lazar Reich and Suhas Vijaykumar. A Possibility in Algorithmic Fairness: Can Calibration and Equal Error Rates Be Reconciled? In *2nd Symposium on Foundations of Responsible Computing*, volume 192, pages 4:1–4:21, 2021. doi: 10.4230/LIPIcs.FORC.2021.4.

[17] 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.

[18] Pengqi Liu, Zijun Yu, Mouloud Belbahri, Arthur Charpentier, Masoud Asgharian, and Jesse C Cresswell. Beyond procedure: Substantive fairness in conformal prediction. *arXiv:2602.16794*, 2026.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763, 2021.

[20] 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):4, 2020.

[21] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108*, 2019.

[22] Glenn Shafer and Vladimir Vovk. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008.

[23] Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[24] Davut Emre Tasar. The coverage-deferral trade-off: Fairness implications of conformal prediction in human-in-the-loop decision systems. *Preprints*, 2025. doi: 10.20944/preprints202512.2631.v1.

[25] Aditya T. Vadlamani, Anutam Srinivasan, Pranav Maneriker, Ali Payani, and Srinivasan Parthasarathy. A generic framework for conformal fairness. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

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

[28] 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.

[29] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, 2018. doi: 10.18653/v1/N18-1101.

[30] 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: Technical Discussion
- Appendix B: Proofs of Theoretical Results
- Appendix C: Additional Experimental Details
- Appendix D: `Bias in Bios` Experiments
- Appendix E: `MultiNLI` Experiments
- Appendix F: `FACET` Experiments
- Appendix G: Additional Robustness and Scale Analyses
- Appendix H: Computational Resources

A Technical Discussion

A.1 Two-group Specialization of the Conservation Law and Uncertainty Relations

In finite-sample split conformal prediction, the empirical quantile $\hat{q}$ satisfies the identity (7) approximately and converges to the population identity asymptotically. We first formalize a two-group version of the conservation law. Let $K = 2$, with groups $g \in \{0, 1\}$ and $p := \mathbb{P}(G = 1) \in (0, 1)$. Without loss of generality, we assume the two quantiles satisfy $q_0 < q_1$. The distance between them is denoted by $\Delta := q_1 - q_0$. For notational convenience, in the two-group case we write $F_0 := F_{S|0}$ and $F_1 := F_{S|1}$. Whenever densities exist, we likewise write $f_0 := f_{S|0}$ and $f_1 := f_{S|1}$.

Lemma 2. *Assume $q_0 < q_1$, $p \in (0, 1)$ and that F_0 and F_1 are continuous. Then the pooled quantile $q = F_S^{-1}(1 - \alpha)$ satisfies $q \in [q_0, q_1]$.*

Assumption 5 (Local density lower bounds for two-group case)**.** For each group $g \in \{0, 1\}$, the conditional distribution of scores admits a density $f_{S|g}$ on the interval $[q_0, q_1]$ and $\operatorname{ess\,inf}_{t\in[q_0,q_1]} f_0(t) \geq m_0 > 0$, and $\operatorname{ess\,inf}_{t\in[q_0,q_1]} f_1(t) \geq m_1 > 0$.

We define the over- and under-coverage magnitudes $\omega_o := \varepsilon_0(q) = F_0(q) - (1 - \alpha)$, $\omega_u := -\varepsilon_1(q) = (1 - \alpha) - F_1(q)$. By Theorem 1, the two-group conservation identity gives

$$(1 - p)\,\omega_o = p\,\omega_u \qquad \Longleftrightarrow \qquad \omega_o = \frac{p}{1 - p}\omega_u. \tag{27}$$

Finally, we write

$$\rho := \frac{p}{1 - p}, \qquad B_{01} := \frac{\Delta}{\frac{1}{m_1} + \rho\frac{1}{m_0}}. \tag{28}$$

Next, we present the two-group uncertainty relation in the following theorem.

Theorem 5. *(Two-group product-type uncertainty relation) Suppose Assumption 5 holds. Let $q_0 < q_1$. Then*

$$\omega_u \geq B_{01}, \tag{29}$$

$$\omega_o \geq \rho B_{01}, \tag{30}$$

$$\omega_u\,\omega_o \geq \rho B_{01}^2. \tag{31}$$

We note that the constant Δ is an intrinsic heterogeneity disparity at level $1 - \alpha$. The ratio $p/(1 - p)$ quantifies how much the pooled threshold is pulled toward group 1. The terms m_0 and m_1 encode local score sensitivities, implying that smaller local densities amplify the threshold displacement required to compensate for a given level of group-wise miscoverage. We next extend this product-type statement to the multi-group setting.

A.2 Generalization of a Product-type Conservation Law and Uncertainty Relation

The two-group statement above admits the following aggregate multi-group extension.

Table 1: Notation summary

Symbol	Meaning
$\mathcal{G} = \{1, \dots, K\}$	Prespecified discrete group set.
$G(X), \quad p_g$	Group label of X and group proportion $p_g = \mathbb{P}(G(X) = g)$.
S	Nonconformity score.
$F_{S\|g}(t)$	Conditional score CDF for group g : $\mathbb{P}(S \le t \mid G = g)$.
q_g	Group-specific $1-\alpha$ quantile of $F_{S\|g}$.
$F_S(t)$	Mixture score CDF: $\sum_{g\in\mathcal{G}} p_g F_{S\|g}(t)$.
q	Pooled $1-\alpha$ quantile of F_S.
$\varepsilon_g(q)$	Group-wise miscoverage under pooled q: $F_{S\|g}(q) - (1-\alpha)$.
$q_G, \quad \varepsilon_G(q)$	Random-group versions of q_g and $\varepsilon_g(q)$ when $G \sim p$.
σ_Δ	Cross-group quantile heterogeneity: $\mathrm{sd}(q_G)$.
$\ell_g(t)$	Expected set size for group g at threshold t.
λ_g	Coverage-calibrated expected set size for group g: $\ell_g(q_g)$.
τ_g	Group-wise threshold satisfying $\ell_g(\tau_g) = \lambda$ under an equalized size.
$m_g(q), \quad m_{\mathrm{eff}}(q)$	Local score density stiffness and its effective aggregate version.
$v_g, \quad v_{\mathrm{eff}}(q)$	Local responsiveness of ℓ_g and its aggregate version.
$\kappa_g, \quad \kappa_{\mathrm{eff}}(\lambda)$	Local coverage–size conversion factor and its aggregate version.
σ_λ	Dispersion of coverage-calibrated set size: $\sigma_\lambda = \mathrm{sd}(\lambda_G)$.

Assumption 6 (Local density lower bounds for multi-group case)**.** Let $q_{\min} := \min_{g\in\mathcal{G}} q_g$ and $q_{\max} := \max_{g\in\mathcal{G}} q_g$, with $q_{\min} < q_{\max}$, and let $q \in (q_{\min}, q_{\max})$ be the threshold under consideration. For each group $g \in \mathcal{G}$, $F_{S|g}$ is absolutely continuous on $[q_{\min}, q_{\max}]$, satisfies $F_{S|g}(q_g) = 1-\alpha$, and has density $f_{S|g}$ satisfying $f_{S|g}(t) \ge m_g > 0$ for almost every $t \in [q_{\min}, q_{\max}]$.

We define the over- and under-coverage magnitudes for K groups as

$$\Omega_o(q) := \sum_{\varepsilon_g(q)>0} p_g \varepsilon_g(q), \qquad \Omega_u(q) := \sum_{\varepsilon_g(q)<0} p_g[-\varepsilon_g(q)], \tag{32}$$

and

$$w_g := p_g m_g, \qquad \bar{q}_m := \frac{\sum_{g\in\mathcal{G}} w_g q_g}{\sum_{g\in\mathcal{G}} w_g}, \qquad B_K := \frac{1}{2}\sum_{g\in\mathcal{G}} w_g |q_g - \bar{q}_m|. \tag{33}$$

Theorem 6. *(Generalization of a product-type uncertainty relation) Suppose Assumption 6 holds. Then*

1. $\max\{\Omega_o(q), \Omega_u(q)\} \ge B_K$ *and* $\min\{\Omega_o(q), \Omega_u(q)\} \ge \big(B_K - |\delta(q)|\big)_+$, *where* $(x)_+ = \max\{x, 0\}$.
2. $\Omega_o(q)\,\Omega_u(q) \ge B_K\big(B_K - |\delta(q)|\big)_+$.
3. *Under exact conservation,* $\delta(q) = 0$, *we have* $\Omega_o(q)\,\Omega_u(q) \ge B_K^2$.

When q is the pooled population quantile and the mixture CDF F_S is continuous at q, Theorem 1 gives $\delta(q) = 0$, so the exact-conservation form in part 3 is the relevant pooled-calibration case.

Theorem 6 refines Theorem 1 from a signed additive conservation law to a magnitude lower bound. The product-type form echoes the fact that pooled calibration does not remove group heterogeneity but redistributes it into over- and under-coverage aggregates.

A.3 Coverage Disparity and Quantile Variance

In this section, we give an elaboration of Theorem 2 when $\mathbb{E}[\varepsilon_G(q)] = 0$. The resulting inequality exhibits a Cramér–Rao-type structure.

Definition 2. For each group, define the segment average density

$$\bar{f}_{S|g}(q) := \begin{cases} \dfrac{F_{S|g}(q) - F_{S|g}(q_g)}{q - q_g} = \dfrac{\varepsilon_g(q)}{q - q_g}, & q \ne q_g, \\ 1, & q = q_g\ . \end{cases} \tag{34}$$

Assume $\bar{f}_{S|g}(q) > 0$ for all groups with $q \neq q_g$; when $q = q_g$, the corresponding term in (35) is zero. We define the apparatus cost term

$$\mathrm{L}(q) := \sqrt{\mathbb{E}\Big[\Big(\frac{q - q_G}{\bar{f}_{S|G}(q)}\Big)^2\Big]}. \tag{35}$$

Theorem 7. *Suppose Assumption 1 holds. With Definition 2,*

$$\sqrt{\mathbb{E}\big[\varepsilon_G(q)^2\big]} \cdot \mathrm{L}(q) \;\geq\; \mathbb{E}\big[(q - q_G)^2\big] \;\geq\; \sigma_\Delta^2. \tag{36}$$

The term $\mathrm{L}(q)$ measures the inevitable cost of imposing a pooled q. The term is large for groups whose quantile displacement is large relative to the induced coverage distortion. In particular, the term quantifies how much the device must move to compensate for heterogeneity. The inequality (36) states that one cannot simultaneously make RMS miscoverage small and keep the apparatus cost small when σ_Δ is non-negligible.

The Cauchy–Schwarz step in Theorem 7 also has a Hölder version. Let $r, s \in [1, \infty]$ be conjugate exponents, $1/r + 1/s = 1$. Under Definition 2,

$$\|\varepsilon_G(q)\|_{L^r(p)} \left\| \frac{q - q_G}{\bar{f}_{S|G}(q)} \right\|_{L^s(p)} \geq \mathbb{E}[(q - q_G)^2] \geq \sigma_\Delta^2. \tag{37}$$

Indeed, writing $Q = q - q_G$, Definition 2 gives $|\varepsilon_G(q)| = \bar{f}_{S|G}(q)|Q|$, and hence

$$\mathbb{E}[Q^2] = \mathbb{E}\left[|\varepsilon_G(q)| \left| \frac{Q}{\bar{f}_{S|G}(q)} \right| \right]. \tag{38}$$

Applying Hölder's inequality gives the first inequality. The second follows from

$$\mathbb{E}[(q - q_G)^2] \geq \mathrm{Var}(q_G) = \sigma_\Delta^2. \tag{39}$$

Taking $r = s = 2$ recovers Theorem 7.

A.4 Connection to Classical Fairness Impossibility Results

A closely related precursor appears in the algorithmic fairness literature. Specifically, Chouldechova [3] shows that in binary settings, when two groups have different base rates, $\pi_g = \mathbb{P}(Y = 1 \mid G = g)$, predictive parity, i.e., equal $\mathrm{PPV}_g = \mathbb{P}(Y = 1 \mid \hat{Y} = 1, G = g)$ across groups, cannot generally hold simultaneously with equalized error profiles matching the false positive rates $\mathrm{FPR}_g = \mathbb{P}(\hat{Y} = 1 \mid Y = 0, G = g)$ and the true positive rates $\mathrm{TPR}_g = \mathbb{P}(\hat{Y} = 1 \mid Y = 1, G = g)$. The incompatibility follows from a Bayes coupling identity relating predictive values, base rates, and the likelihood ratio $\mathrm{TPR}_g/\mathrm{FPR}_g$. Expressed in log-odds form, one may rewrite the incompatibility in the following multiplicative form

$$\exp\big(|\Delta \operatorname{logit}(\mathrm{PPV})|\big) \exp\big(\big|\Delta \log(\mathrm{TPR}/\mathrm{FPR})\big|\big) \geq \exp\big(|\Delta \operatorname{logit}(\pi)|\big), \tag{40}$$

where $\operatorname{logit}(x) = \log(x/(1-x))$ and $\Delta f := f_1 - f_2$. The two exponential terms quantify deviations from predictive parity and from equalized error rate structure, while the right-hand side depends only on the base rate gap. Thus, in that setting, unequal base rates act as a structural lower bound that prevents fairness deviations from simultaneously vanishing.

A.5 Practical Plug-in Diagnostics

Our experiments use empirical plug-in diagnostics for the heterogeneity scales and effective lower-bound proxies. For these diagnostics, $\hat{q}$ and $\hat{q}_g$ are computed from the calibration split, whereas $\hat{p}_g$, $\hat{F}_{S|g}$, $\hat{\ell}_g$, and the reported RMS distortions are computed on the corresponding test sample, namely the independent test sample in the synthetic simulations and the test split in the real-data experiments. The quantile heterogeneity can be estimated by

$$\hat{\sigma}_\Delta^2 = \sum_g \hat{p}_g \left(\hat{q}_g - \sum_h \hat{p}_h \hat{q}_h \right)^2.$$

For any threshold t, let

$$\hat{F}_{S|g}(t) = \frac{1}{n_g} \sum_{i:G_i=g} \mathbf{1}\{S_i \le t\}, \qquad \hat{\ell}_g(t) = \frac{1}{n_g} \sum_{i:G_i=g} |\hat{C}_t(X_i)|$$

be the empirical group score CDF and empirical set-size curve on this test sample, where n_g is the number of test examples in group g. Rather than estimating essential infima, segment-average proxies along the same policy-conversion paths are:

$$\hat{m}_g^{\text{seg}} = \frac{|\hat{F}_{S|g}(\hat{q}) - \hat{F}_{S|g}(\hat{q}_g)|}{|\hat{q} - \hat{q}_g|}, \qquad \hat{v}_g^{\text{seg}} = \frac{|\hat{\ell}_g(\hat{q}_g) - \hat{\ell}_g(\hat{q})|}{|\hat{q}_g - \hat{q}|}.$$

Next, define

$$\hat{\lambda}_g := \hat{\ell}_g(\hat{q}_g), \qquad \hat{\lambda} := \sum_g \hat{p}_g \hat{\lambda}_g, \qquad \hat{\sigma}_\lambda^2 = \sum_g \hat{p}_g (\hat{\lambda}_g - \hat{\lambda})^2.$$

If $\hat{\tau}_g$ is the empirical threshold used to attain the common size target $\hat{\lambda}$, then

$$\hat{\kappa}_g^{\text{seg}} = \frac{|\hat{F}_{S|g}(\hat{\tau}_g) - \hat{F}_{S|g}(\hat{q}_g)|}{|\hat{\lambda} - \hat{\lambda}_g|}.$$

When a denominator is zero, the corresponding segment proxy is set to zero.

Finally, for group-wise coefficients a_g and gaps d_g, define

$$\mathcal{E}(a, d) = \left(\frac{\sum_g \hat{p}_g a_g^2 d_g^2}{\sum_g \hat{p}_g d_g^2} \right)^{1/2},$$

which can be used to compute

$$\widehat{m}_{\text{eff}}^{\text{seg}} = \mathcal{E}(\hat{m}^{\text{seg}}, \hat{q} - \hat{q}_g), \quad \widehat{v}_{\text{eff}}^{\text{seg}} = \mathcal{E}(\hat{v}^{\text{seg}}, \hat{q} - \hat{q}_g), \quad \widehat{\kappa}_{\text{eff}}^{\text{seg}} = \mathcal{E}(\hat{\kappa}^{\text{seg}}, \hat{\lambda}_g - \hat{\lambda}),$$

yielding the diagnostic scales

$$\widehat{m}_{\text{eff}}^{\text{seg}} \hat{\sigma}_\Delta, \qquad \widehat{v}_{\text{eff}}^{\text{seg}} \hat{\sigma}_\Delta, \qquad \widehat{\kappa}_{\text{eff}}^{\text{seg}} \hat{\sigma}_\lambda.$$

These quantities are empirical diagnostic proxy scales for assessing whether group heterogeneity is large enough for coverage or set-size distortions to be visible.

B Proofs of Theoretical Results

Proof of Theorem 1

Theorem 1. *(Conservation law for pooled calibration) Let $p_g := \mathbb{P}\{G(X) = g\}$ be the group mass. Group-wise coverage distortion $\varepsilon_g(q)$ under the pooled threshold satisfies the aggregate identity*

$$\sum_{g \in \mathcal{G}} p_g \, \varepsilon_g(q) = F_S(q) - (1 - \alpha) =: \delta(q). \tag{7}$$

If F_S is continuous at q, then $\delta(q) = 0$, and hence $\sum_{g \in \mathcal{G}} p_g \, \varepsilon_g(q) = 0$.

Proof. By definition of F_S,

$$\sum_{g \in \mathcal{G}} p_g \, \varepsilon_g(q) = \sum_{g \in \mathcal{G}} p_g F_{S|g}(q) - (1 - \alpha) = F_S(q) - (1 - \alpha) = \delta(q) \ge 0,$$

which yields the generalized conservation law. If F_S is continuous at q, then by definition of the quantile $F_S(t) < 1 - \alpha$ for all $t < q$; hence, we have $F_S(q^-) \le 1 - \alpha$. The continuity at q gives

$$F_S(q) = F_S(q^-) \le 1 - \alpha,$$

while definition of q gives $F_S(q) \ge 1 - \alpha$. Thus, we have $F_S(q) = 1 - \alpha$, i.e., $\delta(q) = 0$.

□

Proof of Theorem 2

Theorem 2. *(Pooled-threshold uncertainty relation) Suppose Assumption 1 holds. Then*

$$\mathrm{Var}(\varepsilon_G(q)) \geq m_{\mathrm{eff}}(q)^2 \mathrm{Var}(q_G). \tag{12}$$

Proof. If $\mathbb{E}[(q - q_G)^2] = 0$, then $\mathrm{Var}(q_G) = 0$, so the claim is trivial. We therefore assume $\mathbb{E}[(q - q_G)^2] > 0$, so that $m_{\mathrm{eff}}(q)$ is well defined.

Fix a group $g \in \mathcal{G}$. Since q_g is the group-specific $1 - \alpha$ quantile and $F_{S|g}$ is continuous on the relevant segment, $F_{S|g}(q_g) = 1 - \alpha$. Hence, by Assumption 1,

$$\varepsilon_g(q) = F_{S|g}(q) - F_{S|g}(q_g) = \int_{q_g}^{q} f_{S|g}(t)\, dt.$$

By the definition of $m_g(q)$,

$$|\varepsilon_g(q)| = \left| \int_{q_g}^{q} f_{S|g}(t)\, dt \right| \geq m_g(q)|q - q_g|.$$

Squaring and averaging over $G \sim p$ gives

$$\mathbb{E}[\varepsilon_G(q)^2] \geq \mathbb{E}\left[m_G(q)^2 (q - q_G)^2\right].$$

By the definition of $m_{\mathrm{eff}}(q)$,

$$\mathbb{E}\left[m_G(q)^2 (q - q_G)^2\right] = m_{\mathrm{eff}}(q)^2 \mathbb{E}[(q - q_G)^2].$$

Moreover,

$$\mathbb{E}[(q - q_G)^2] = \mathrm{Var}(q_G) + (q - \mathbb{E}[q_G])^2 \geq \mathrm{Var}(q_G).$$

Finally, Assumption 1 implies that the mixture CDF F_S is continuous at q, so Theorem 1 gives

$$\mathbb{E}[\varepsilon_G(q)] = \sum_{g \in \mathcal{G}} p_g \varepsilon_g(q) = 0.$$

Therefore,

$$\mathrm{Var}(\varepsilon_G(q)) = \mathbb{E}[\varepsilon_G(q)^2] \geq m_{\mathrm{eff}}(q)^2 \mathrm{Var}(q_G).$$

□

Proof of Lemma 1

Lemma 1. *(Monotonicity of the expected set size) For any $t_1 \leq t_2$, $\hat{C}_{t_1}(x) \subseteq \hat{C}_{t_2}(x)$ for every x,*

$$\ell_g(t_1) \leq \ell_g(t_2) \quad \textit{for every } g \in \mathcal{G}. \tag{14}$$

Proof. Fix $t_1 \leq t_2$ and $x \in \mathcal{X}$. Let $y \in \hat{C}_{t_1}(x)$. By definition,

$$S(x, y) \leq t_1.$$

Since $t_1 \leq t_2$, we have $S(x, y) \leq t_2$, and hence $y \in \hat{C}_{t_2}(x)$. Therefore,

$$\hat{C}_{t_1}(x) \subseteq \hat{C}_{t_2}(x),$$

which implies

$$|\hat{C}_{t_1}(x)| \leq |\hat{C}_{t_2}(x)|.$$

Applying this inequality to X and taking conditional expectations given $G = g$ yields

$$\ell_g(t_1) = \mathbb{E}\Big[|\hat{C}_{t_1}(X)| \mid G = g\Big] \leq \mathbb{E}\Big[|\hat{C}_{t_2}(X)| \mid G = g\Big] = \ell_g(t_2).$$

□

Proof of Theorem 3

Theorem 3. *(Equalized coverage induces cross-group expected set size disparity) Assume the score CDF $F_{S|g}$ is continuous $\forall g \in \mathcal{G}$. Suppose Assumption 2 holds for some reference group $r \in \mathcal{G}$. Then,*

1. *The group-wise thresholds $\{q_g\}_{g\in\mathcal{G}}$, which achieve an exact group-wise coverage level $1-\alpha$, necessarily induce a nonzero cross-group disparity in expected set size. More precisely, for every $g \in \mathcal{H}_r \setminus \{r\}$, $\ell_g(q_g) - \ell_r(q_r) \geq c_g > 0$. Consequently,*

$$\max_{g,g'\in\mathcal{G}} |\ell_g(q_g) - \ell_{g'}(q_{g'})| \geq \max_{g\in\mathcal{H}_r\setminus\{r\}} c_g > 0. \tag{16}$$

   Therefore, exact group-wise coverage cannot simultaneously satisfy equalized expected set size across groups.
2. *The restricted mean squared cross-group size disparity relative to the reference group r satisfies*

$$D_r^2 = \sum_{g\in\mathcal{H}_r\setminus\{r\}} p_g(\ell_g(q_g) - \ell_r(q_r))^2 \geq \sum_{g\in\mathcal{H}_r\setminus\{r\}} p_g c_g^2 > 0. \tag{17}$$

Proof. The group-wise thresholds $\{q_g\}_{g\in\mathcal{G}}$ achieve equalized coverage at level $1-\alpha$ across groups. Now fix any $g \in \mathcal{H}_r \setminus \{r\}$. By definition of $\mathcal{H}_r$, we have $q_g \geq q_r$. By Lemma 1,

$$\ell_g(q_g) \geq \ell_g(q_r).$$

Subtracting $\ell_r(q_r)$ from both sides and using Assumption 2 gives

$$\ell_g(q_g) - \ell_r(q_r) \geq \ell_g(q_r) - \ell_r(q_r) \geq c_g > 0.$$

This proves the first displayed claim.

Taking the maximum over $g \in \mathcal{H}_r \setminus \{r\}$ yields

$$\max_{g,g'\in\mathcal{G}} |\ell_g(q_g) - \ell_{g'}(q_{g'})| \geq \max_{g\in\mathcal{H}_r\setminus\{r\}} |\ell_g(q_g) - \ell_r(q_r)| \geq \max_{g\in\mathcal{H}_r\setminus\{r\}} c_g > 0.$$

Thus, equalized expected set size across all groups is impossible under the equalized group-wise coverage policy.

For the mean square claim, using the previously established bound for each $g \in \mathcal{H}_r \setminus \{r\}$,

$$(\ell_g(q_g) - \ell_r(q_r))^2 \geq c_g^2.$$

Multiplying by p_g and summing over $g \in \mathcal{H}_r \setminus \{r\}$ gives

$$D_r^2 = \sum_{g\in\mathcal{H}_r\setminus\{r\}} p_g(\ell_g(q_g) - \ell_r(q_r))^2 \geq \sum_{g\in\mathcal{H}_r\setminus\{r\}} p_g c_g^2 > 0.$$

This proves the claim. □

Proof of Theorem 4

Theorem 4. *(Equalized expected set size induces cross-group coverage disparity) Suppose Assumption 3 holds, and suppose for some common target level λ the thresholds $\{\tau_g\}_{g\in\mathcal{G}}$ satisfy $\ell_g(\tau_g) = \lambda,\ g \in \mathcal{G}$. Then the following results hold.*

(i) *For every group g with $\lambda_g < \lambda$, $F_{S|g}(\tau_g) - (1-\alpha) \ \geq \ \kappa_g\,(\lambda - \lambda_g)$.*

(ii) *For every group g' with $\lambda_{g'} > \lambda$, $(1-\alpha) - F_{S|g'}(\tau_{g'}) \ \geq \ \kappa_{g'}\,(\lambda_{g'} - \lambda)$.*

Consequently, for any pair $g, g' \in \mathcal{G}$ such that $\lambda_g < \lambda < \lambda_{g'}$, we have the pairwise lower bound

$$F_{S|g}(\tau_g) - F_{S|g'}(\tau_{g'}) \ \geq \ \kappa_g(\lambda - \lambda_g) + \kappa_{g'}(\lambda_{g'} - \lambda). \tag{18}$$

In particular, if there exist groups g, g' with $\lambda_g < \lambda_{g'}$ and $\lambda \in (\lambda_g, \lambda_{g'})$, then

$$\max_{a,b\in\mathcal{G}} |F_{S|a}(\tau_a) - F_{S|b}(\tau_b)| \ \geq \ \kappa_g(\lambda - \lambda_g) + \kappa_{g'}(\lambda_{g'} - \lambda) \ > \ 0. \tag{19}$$

Proof. Fix a group $g \in \mathcal{G}$. If $\lambda_g < \lambda$, then, since ℓ_g is monotone on the segment between q_g and τ_g and $\ell_g(q_g) = \lambda_g$, the identity $\ell_g(\tau_g) = \lambda$ implies

$$\tau_g > q_g.$$

By absolute continuity and Assumption 3,

$$\lambda - \lambda_g = \ell_g(\tau_g) - \ell_g(q_g) = \int_{q_g}^{\tau_g} \ell_g'(t)\, dt.$$

Hence

$$\lambda - \lambda_g \le \int_{q_g}^{\tau_g} |\ell_g'(t)|\, dt \le V_g(\tau_g - q_g),$$

so that

$$\tau_g - q_g \ge \frac{\lambda - \lambda_g}{V_g}.$$

Again by Assumption 3,

$$F_{S|g}(\tau_g) - F_{S|g}(q_g) = \int_{q_g}^{\tau_g} f_{S|g}(t)\, dt \ge m_g(\tau_g - q_g) \ge \frac{m_g}{V_g}(\lambda - \lambda_g) = \kappa_g(\lambda - \lambda_g).$$

Since $F_{S|g}(q_g) = 1 - \alpha$, this proves

$$F_{S|g}(\tau_g) - (1 - \alpha) \ge \kappa_g(\lambda - \lambda_g).$$

Now suppose $\lambda_{g'} > \lambda$. Since $\ell_{g'}(q_{g'}) = \lambda_{g'}$ and $\ell_{g'}(\tau_{g'}) = \lambda$, we have

$$\ell_{g'}(q_{g'}) > \ell_{g'}(\tau_{g'}).$$

Because $\ell_{g'}$ is non-decreasing on the segment between $q_{g'}$ and $\tau_{g'}$, this implies

$$\tau_{g'} < q_{g'}.$$

Indeed, if $\tau_{g'} \ge q_{g'}$, monotonicity would imply

$$\ell_{g'}(\tau_{g'}) \ge \ell_{g'}(q_{g'}),$$

contradicting $\ell_{g'}(\tau_{g'}) < \ell_{g'}(q_{g'})$. Then

$$\lambda_{g'} - \lambda = \ell_{g'}(q_{g'}) - \ell_{g'}(\tau_{g'}) = \int_{\tau_{g'}}^{q_{g'}} \ell_{g'}'(t)\, dt \le \int_{\tau_{g'}}^{q_{g'}} |\ell_{g'}'(t)|\, dt \le V_{g'}(q_{g'} - \tau_{g'}),$$

hence

$$q_{g'} - \tau_{g'} \ge \frac{\lambda_{g'} - \lambda}{V_{g'}}.$$

Using Assumption 3 again,

$$F_{S|g'}(q_{g'}) - F_{S|g'}(\tau_{g'}) = \int_{\tau_{g'}}^{q_{g'}} f_{S|g'}(t)\, dt \ge m_{g'}(q_{g'} - \tau_{g'}) \ge \frac{m_{g'}}{V_{g'}}(\lambda_{g'} - \lambda) = \kappa_{g'}(\lambda_{g'} - \lambda).$$

Since $F_{S|g'}(q_{g'}) = 1 - \alpha$, we obtain

$$(1 - \alpha) - F_{S|g'}(\tau_{g'}) \ge \kappa_{g'}(\lambda_{g'} - \lambda).$$

Finally, if $\lambda_g < \lambda < \lambda_{g'}$, then combining the two inequalities gives

$$F_{S|g}(\tau_g) - F_{S|g'}(\tau_{g'}) = \big(F_{S|g}(\tau_g) - (1-\alpha)\big) + \big((1-\alpha) - F_{S|g'}(\tau_{g'})\big) \ge \kappa_g(\lambda - \lambda_g) + \kappa_{g'}(\lambda_{g'} - \lambda).$$

Because $\lambda_g < \lambda < \lambda_{g'}$, both terms on the right-hand side are strictly positive, and hence the right-hand side is > 0. Therefore

$$\max_{a,b \in \mathcal{G}} |F_{S|a}(\tau_a) - F_{S|b}(\tau_b)| \ge F_{S|g}(\tau_g) - F_{S|g'}(\tau_{g'}) > 0,$$

which proves that coverage cannot be equalized across groups under this equalized-size policy. □

Proof of Corollary 1

Corollary 1. *(Exact group-wise coverage induces set size distortion) Suppose Assumption 4 holds. Let the pooled threshold q be replaced by the group-wise thresholds $\{q_g\}$. Then the RMS set size lower bound is*

$$\sqrt{\mathbb{E}[(\ell_G(q_G) - \ell_G(q))^2]} \;\geq\; v_{\text{eff}}(q)\,\sigma_\Delta. \tag{21}$$

Proof. Let g be fixed. By Assumption 4 and the fundamental theorem of calculus, we have

$$\ell_g(q_g) - \ell_g(q) = \int_q^{q_g} \ell_g'(t)\,dt.$$

Under Assumption 4, $|\ell_g'(t)| \geq v_g$ along the segment between q and q_g, hence $|\ell_g(q_g) - \ell_g(q)| \geq v_g|q_g - q|$. Squaring and averaging gives $\mathbb{E}[(\ell_G(q_G) - \ell_G(q))^2] \geq \mathbb{E}[v_G^2(q_G - q)^2]$. Taking square roots and using the definition of $v_{\text{eff}}(q)$ gives $\sqrt{E[(\ell_G(q_G) - \ell_G(q))^2]} \geq v_{\text{eff}}(q)\sqrt{E[(q_G - q)^2]}$. Since $E[(q_G - q)^2] \geq \text{Var}(q_G) = \sigma_\Delta^2$, the desired bound follows. □

Proof of Corollary 2

Corollary 2. *(Equalized expected set size induces coverage distortion) Suppose Assumption 3 holds. Let the group-wise thresholds $\{\tau_g\}_{g\in G}$ satisfy $\ell_g(\tau_g) = \lambda$ for all $g \in G$, so that they enforce equalized expected set size at a common level λ. Then the RMS coverage distortion relative to the exact group-wise coverage satisfies*

$$\sqrt{\mathbb{E}[(F_{S|G}(\tau_G) - F_{S|G}(q_G))^2]} \;\geq\; \kappa_{\text{eff}}(\lambda)\sigma_\lambda. \tag{23}$$

Proof. We fix group g. Since $\ell_g(q_g) = \lambda_g$ by definition, we have

$$\lambda - \lambda_g = \ell_g(\tau_g) - \ell_g(q_g) = \int_{q_g}^{\tau_g} \ell_g'(t)dt\,.$$

Therefore, $|\lambda - \lambda_g| \leq V_g|\tau_g - q_g|$, which implies

$$|\tau_g - q_g| \geq \frac{|\lambda - \lambda_g|}{V_g}\,.$$

Moreover,

$$|F_{S|g}(\tau_g) - F_{S|g}(q_g)| = \left|\int_{q_g}^{\tau_g} f_{S|g}(t)dt\right| \geq m_g|\tau_g - q_g| \geq m_g\frac{|\lambda - \lambda_g|}{V_g}.$$

Squaring, averaging, taking square roots, and using the definition of $\kappa_{\text{eff}}(\lambda)$ gives

$$\sqrt{E[(F_{S|G}(\tau_G) - F_{S|G}(q_G))^2]} \geq \kappa_{\text{eff}}(\lambda)\sqrt{E[(\lambda - \lambda_G)^2]}. \tag{41}$$

Since $E[(\lambda - \lambda_G)^2] \geq \text{Var}(\lambda_G) = \sigma_\lambda^2$, the desired bound follows. □

Proof of Corollary 3

Corollary 3. *(Pooled to Equalized Set Size coverage change) Suppose that, on I_g°, both $F_{S|g}$ and ℓ_g are absolutely continuous and, for almost every $t \in I_g^\circ$, $f_{S|g}(t) \geq m_g^\circ > 0, |\ell_g'(t)| \leq V_g^\circ < \infty$, with $V_g^\circ > 0$. Define $\kappa_g^\circ := m_g^\circ/V_g^\circ$. Then,*

$$\left|F_{S|g}(\tau_g) - F_{S|g}(q)\right| \geq \kappa_g^\circ\,|\lambda - \ell_g(q)| \;\; \textit{for every group } g \tag{25}$$

and

$$\sqrt{\mathbb{E}\Big[\big(F_{S|G}(\tau_G) - F_{S|G}(q)\big)^2\Big]} \geq \sqrt{\mathbb{E}\Big[(\kappa_G^\circ)^2\,(\lambda - \ell_G(q))^2\Big]}. \tag{26}$$

Proof. For each group g,

$$|\lambda - \ell_g(q)| = \left| \int_q^{\tau_g} \ell_g'(t)\, dt \right| \le V_g^\circ |\tau_g - q|. \tag{42}$$

Also,

$$\left|F_{S|g}(\tau_g) - F_{S|g}(q)\right| = \left| \int_q^{\tau_g} f_{S|g}(t)\, dt \right| \ge m_g^\circ |\tau_g - q|. \tag{43}$$

Combining the two inequalities gives

$$\left|F_{S|g}(\tau_g) - F_{S|g}(q)\right| \ge \kappa_g^\circ \left|\lambda - \ell_g(q)\right|. \tag{44}$$

Squaring and averaging over G gives the RMS bound. □

Proof of Lemma 2

Lemma 2. *Assume $q_0 < q_1$, $p \in (0,1)$ and that F_0 and F_1 are continuous. Then the pooled quantile $q = F_S^{-1}(1-\alpha)$ satisfies $q \in [q_0, q_1]$.*

Proof. Assume $q < q_0$ and prove by contradiction: Given $q < q_0$, we have $q < q_1$ since $q_0 < q_1$. According to the definition of q_0 and q_1, we must have $F_0(q) < 1-\alpha$ and $F_1(q) < 1-\alpha$, so

$$F_S(q) = (1-p)F_0(q) + pF_1(q) < (1-p)(1-\alpha) + p(1-\alpha) = 1-\alpha.$$

However, the above derivation of $F_S(q) < 1-\alpha$ contradicts the definition of q, which requires $F_S(q) \ge 1-\alpha$. Therefore, we must have $q \ge q_0$. Regarding q_1, we have

$$F_S(q_1) = (1-p)F_0(q_1) + pF_1(q_1) \ge (1-p)(1-\alpha) + p(1-\alpha) = 1-\alpha.$$

Thus, $q_1 \in \{t \in \mathbb{R} : F_S(t) \ge 1-\alpha\}$. Because $q = \inf\{t \in \mathbb{R} : F_S(t) \ge 1-\alpha\}$, we have $q \le q_1$. Together with $q \ge q_0$, we have $q \in [q_0,\, q_1]$. □

Proof of Theorem 5

Theorem 5. *(Two-group product-type uncertainty relation) Suppose Assumption 5 holds. Let $q_0 < q_1$. Then*

$$\omega_u \ge B_{01}, \tag{29}$$

$$\omega_o \ge \rho B_{01}, \tag{30}$$

$$\omega_u\, \omega_o \ge \rho B_{01}^2. \tag{31}$$

Proof. Let $d_0 := q - q_0 \ge 0$ and $d_1 := q_1 - q \ge 0$, so $d_0 + d_1 = \Delta$. By the fundamental theorem of calculus and Assumption 5,

$$\omega_o = F_0(q) - F_0(q_0) = \int_{q_0}^{q} f_0(t)\, dt \ge m_0 d_0, \qquad \omega_u = F_1(q_1) - F_1(q) = \int_q^{q_1} f_1(t)\, dt \ge m_1 d_1.$$

The relationship (27) gives $(1-p)m_0 d_0 \le (1-p)\omega_o = p\omega_u$, while $\omega_u \ge m_1 d_1$. A convenient way to combine the constraints is to express d_0 in terms of ω_u: $(1-p)\omega_o = p\omega_u$ and $\omega_o \ge m_0 d_0$ imply $(1-p)m_0 d_0 \le p\omega_u$, so $d_0 \le \frac{p}{(1-p)m_0}\omega_u$. Similarly, $\omega_u \ge m_1 d_1$ implies $d_1 \le \frac{1}{m_1}\omega_u$. Since $\Delta = d_0 + d_1$, we obtain

$$\Delta \le \Big(\frac{1}{m_1} + \frac{p}{1-p}\frac{1}{m_0}\Big)\omega_u,$$

which yields (29). Then (30) follows from (27). Multiplying (29) and (30) gives (31). □

Proof of Theorem 6

Theorem 6. *(Generalization of a product-type uncertainty relation) Suppose Assumption 6 holds. Then*

1. $\max\{\Omega_o(q), \Omega_u(q)\} \geq B_K$ *and* $\min\{\Omega_o(q), \Omega_u(q)\} \geq \big(B_K - |\delta(q)|\big)_+$, *where* $(x)_+ = \max\{x, 0\}$.
2. $\Omega_o(q)\ \Omega_u(q) \geq B_K\big(B_K - |\delta(q)|\big)_+$.
3. *Under exact conservation,* $\delta(q) = 0$, *we have* $\Omega_o(q)\ \Omega_u(q) \geq B_K^2$.

Proof. For any group with $q \geq q_g$, we have $\varepsilon_g(q) \geq m_g(q - q_g)$. Similarly, for any group with $q \leq q_g$, we have $\varepsilon_g(q) \leq m_g(q - q_g)$. Therefore, we have

$$\Omega_o(q) \geq \sum_g w_g(q - q_g)_+ =: A_+(q) \qquad \Omega_u(q) \geq \sum_g w_g(q_g - q)_+ =: A_-(q). \tag{45}$$

At the crossing point $\bar{q}_m$,

$$A_+(\bar{q}_m) = A_-(\bar{q}_m) = \frac{1}{2}\sum_g w_g|q_g - \bar{q}_m| = B_K. \tag{46}$$

Since $A_+(q) - A_-(q) = \sum_{g\in\mathcal{G}} w_g(q - q_g)$, the two functions cross at $\bar{q}_m$. Moreover, $A_+(q)$ is non-decreasing in q and $A_-(q)$ is non-increasing in q. Hence, the minimum of $\max\{A_+(q), A_-(q)\}$ is attained at the crossing point, which equals B_K. Thus, we obtain $\max\{\Omega_o(q), \Omega_u(q)\} \geq B_K$. Furthermore, Theorem 1 gives $\Omega_o(q) - \Omega_u(q) = \delta(q)$. Since $\max\{\Omega_o(q), \Omega_u(q)\} - \min\{\Omega_o(q), \Omega_u(q)\} = |\delta(q)|$, we also have $\min\{\Omega_o(q), \Omega_u(q)\} \geq \big(B_K - |\delta(q)|\big)_+$. Therefore,

$$\Omega_o(q)\Omega_u(q) \geq B_K\big(B_K - |\delta(q)|\big)_+. \tag{47}$$

Under exact conservation, $\delta(q) = 0$, and hence

$$\Omega_o(q)\ \Omega_u(q) \geq B_K^2. \tag{48}$$

□

Proof of Theorem 7

Theorem 7. *Suppose Assumption 1 holds. With Definition 2,*

$$\sqrt{\mathbb{E}\big[\varepsilon_G(q)^2\big]} \cdot \mathrm{L}(q) \ \geq\ \mathbb{E}\big[(q - q_G)^2\big] \ \geq\ \sigma_\Delta^2. \tag{36}$$

Proof. Let $Q := q - q_G$. For groups with $q \neq q_g$, (34) implies $\varepsilon_G(q) = \bar{f}_{S|G}(q) \cdot Q$ almost surely. Then $\mathbb{E}[Q^2] = \mathbb{E}[\varepsilon_G(q) \cdot \frac{Q}{\bar{f}_{S|G}(q)}]$. Applying Cauchy–Schwarz yields $\mathbb{E}[Q^2]^2 \leq \mathbb{E}[\varepsilon_G(q)^2] \cdot \mathbb{E}[(\frac{Q}{\bar{f}_{S|G}(q)})^2]$. Taking square roots gives the first inequality in (36). The second inequality is identical to the argument in Theorem 2. □

C Additional Synthetic Experimental Details

C.1 Synthetic Simulations

In this experiment, we directly synthesize nonconformity scores, rather than computing them from data and a model. For each group, we estimate the split-conformal threshold by the empirical $1 - \alpha$ quantile of the calibration scores. The pooled threshold $\hat{q}$ is obtained by concatenating all calibration scores across subgroups and applying the same quantile rule. Empirical coverage quantities are then evaluated on an independent test sample, so the observed distortion reflects finite-sample threshold estimation.

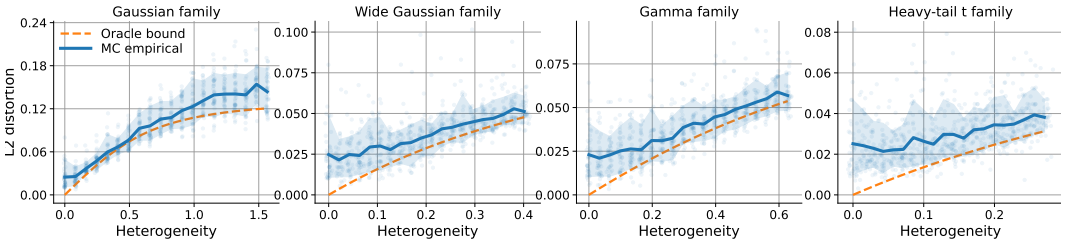


Figure 6: Four multi-group families. Across all four score families, the empirical RMS miscoverage increases with heterogeneity and remains above the lower-bound scale from Theorem 2.

Table 2: Summary of empirical-oracle comparison at the largest heterogeneity value in each setting (means over 40 seeds). Ratios are computed from unrounded means.

Settings	Empirical	Oracle	Ratio
Two-group Gaussian	0.097	0.085	1.141
Four-group Gaussian	0.144	0.121	1.190
Four-group wide Gaussian	0.051	0.048	1.063
Four-group Gamma	0.057	0.054	1.056
Four-group heavy-tail t	0.038	0.031	1.226
Coverage→size	2.860	2.127	1.345
Size→coverage	0.352	0.347	1.014

Figure 6 provides additional pooled-threshold validation. Across the two-group sweep and the four multi-group families, the empirical RMS miscoverage increases with heterogeneity and remains above the oracle scale $m_{\text{eff}}(q)\sigma_\Delta$. Table 2 reports the empirical-oracle comparison at the largest heterogeneity value in each setting. Table 3 provides details about the distributions of the simulations.

For the set-size based experiments, we use the monotone linear proxy $\ell_g(t) = a_g + b_g t$, with $a_g = 2 + 0.15(g-1)$ and $b_g = 0.8 + 0.1(g-1)$ after labeling the group index as $g = 1, \ldots, K$. In the size-to-coverage experiment, the common target size is the weighted average of the coverage-calibrated group sizes, namely $\hat{\lambda} = \sum_g p_g \ell_g(\hat{q}_g)$, and the equalized-size threshold $\hat{\tau}_g$ solves $\ell_g(\hat{\tau}_g) = \hat{\lambda}$.

C.2 Imbalance Bottleneck and Ratio Diagnostics

As an additional diagnostic beyond the balanced synthetic experiments, we perform pooled-threshold simulations in four imbalanced settings: Gaussian, wide Gaussian, gamma, and Student-t distributions. Each setting has four distributions of the same type but with imbalanced group masses $p = (0.60, 0.25, 0.10, 0.05)$. We use a total calibration budget of $n_{\text{cal}} = 400$, 4000 test points per group, target level $\alpha = 0.1$, and 40 Monte Carlo seeds. Figure 7 shows that the RMS pooled distortion remains above the oracle scale of all four families, whereas the minority-group absolute gap grows more rapidly with heterogeneity. To verify that the bottleneck is indeed concentrated on the rarest group, Table 5 reports how often the minority group is the largest-gap group for each heterogeneity level. The entries report counts out of 40 seeds for which the minority group attains the largest absolute gap. The frequency rises rapidly with heterogeneity and reaches $40/40$ at the largest heterogeneity of 2.0. This pattern indicates a clear minority bottleneck under pooled calibration.

Tables 6 and 7 record empirical and oracle ratios for the two policy-conversion diagnostics at selected heterogeneity levels. The coverage-to-size ratio remains above one throughout, with the finite-sample slack shrinking as n_{cal} grows. The size-to-coverage ratio is tighter and concentrates near one much more quickly.

Table 3: Simulation setups for the pooled-threshold experiments in Figure 2 and Figure 6. All groups have equal weight $p_g = 1/K$, all experiments use $\alpha = 0.1$, and the empirical miscoverage is compared against the theoretical scale $m_{\text{eff}}(q)\sigma_\Delta$.

Setting	Score family	Cal/Test
Two-group Gaussian shift	$S \mid G = g \sim \mathcal{N}(\mu_g, 1)$, with $(\mu_0, \mu_1) = (0, \delta)$ and $\delta \in [0, 2]$	50/500
Four-group Gaussian shift	$S \mid G = g \sim \mathcal{N}(\mu_g, 1)$, with $\mu_g = s(-1.5, -0.5, 0.5, 1.5)$ and $s \in [0, 1.4]$	50/500
Four-group wide Gaussian	$S \mid G = g \sim \mathcal{N}(0, \sigma_g^2)$, with $\sigma_g = 1 + s(0, 0.2, 0.4, 0.6)$ and $s \in [0, 1.4]$	50/500
Four-group Gamma	$S \mid G = g \sim \text{Gamma}(4, \theta_g)$, with $\theta_g = 0.55 + s(0, 0.06, 0.12, 0.18)$ and $s \in [0, 1.4]$	50/500
Four-group heavy-tail t	$S \mid G = g \sim t_6(0, \sigma_g)$, with $\sigma_g = 0.9 + s(0, 0.12, 0.24, 0.36)$ and $s \in [0, 1.4]$	50/500

Table 4: Simulation setups for the group-adjusted trade-off experiments in Figure 1. All groups have equal weight $p_g = 1/K$ and all experiments use $\alpha = 0.1$.

Setting	Score family	Cal/Test	Compared quantity
Coverage vs. size	Two-component asymmetric Gaussian mixtures with group-specific offsets, scales, and weights; $s \in [0, 1.8]$ scales group centers and increases quantile separation	25/100	size change vs. $v_{\text{eff}}(q)\sigma_\Delta$
Size vs. coverage	Two-component mixtures of one Student-t component and one Gaussian component, with group-specific offsets, scales, and weights; $s \in [0, 2.0]$ scales group centers	25/100	coverage change vs. $\kappa_{\text{eff}}(\lambda)\sigma_\lambda$

D `Bias in Bios` Experiments

D.1 Experimental Settings

We perform conformal analysis on outputs from a `DistilBERT`[3] classifier trained on the ten most frequent professions from `Bias in Bios`[4]. Training uses cleaned, preprocessed text with a maximum sequence length of 160, a learning rate of 2×10^{-5}, weight decay of 0.01, batch sizes of $64/32$ for training/validation, warm-up ratio 0.05, and two training epochs. The validation accuracy and macro-F1 score are 0.8886 and 0.8569, respectively. The dataset is partitioned into training, validation, calibration, and test splits of sizes 185758, 20640, 31764, and 79397, with stable group balance across splits. Details are provided in Table 9. Throughout, we use the simple score in the main presentation, while SAPS and RAPS are included to show that the same conclusions are not

[3]We use the Hugging Face checkpoint `distilbert-base-uncased`; model card: https://huggingface.co/distilbert/distilbert-base-uncased. The model card lists the license as Apache-2.0.

[4]We use the Hugging Face dataset `LabHC/bias_in_bios`; dataset card: https://huggingface.co/datasets/LabHC/bias_in_bios. The dataset card lists the license as MIT. For the original dataset source, see [6].

Table 5: Minority frequency across heterogeneity levels.

	Heterogeneity level				
Family	0.0	0.5	1.0	1.5	2.0
Gaussian	13/40	40/40	40/40	40/40	40/40
Wide Gaussian	13/40	36/40	39/40	40/40	40/40
Gamma	13/40	37/40	40/40	40/40	40/40
Heavy-tail t	9/40	30/40	39/40	37/40	40/40

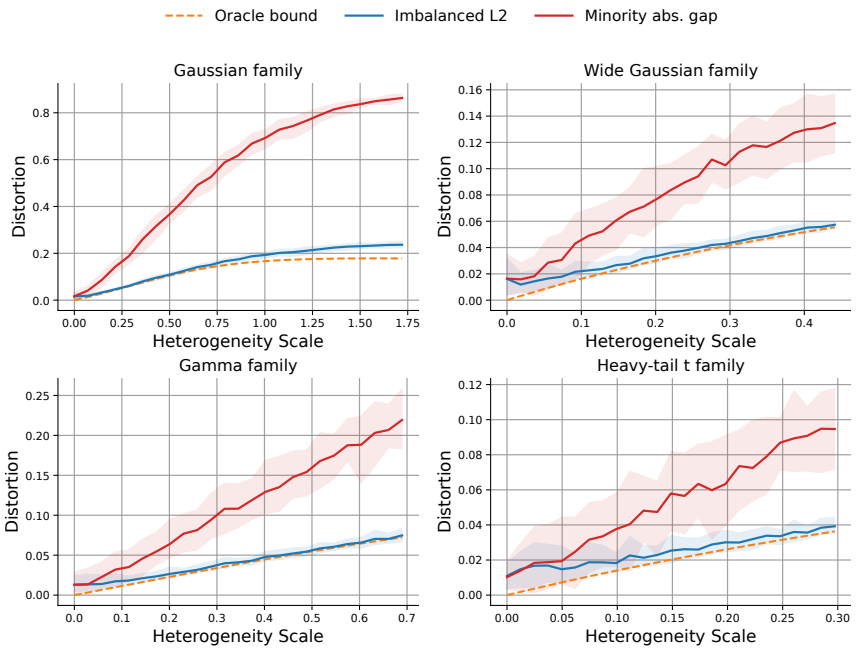


Figure 7: Imbalanced four-group pooled-threshold diagnostics. Across all four families, the weighted RMS distortion remains above the lower-bound scale from Theorem 2. At the same time, the minority-group absolute gap grows more rapidly with heterogeneity, showing that pooled distortion can concentrate on rare groups.

Table 6: Coverage-to-size empirical/oracle ratio at heterogeneity levels σ_Δ.

	Selected heterogeneity level σ_Δ				
n_{cal}	0.378	0.855	1.250	1.652	2.192
12	1.837	1.417	1.327	1.334	1.392
25	1.443	1.174	1.243	1.244	1.296
50	1.284	1.174	1.178	1.232	1.287
100	1.226	1.087	1.172	1.227	1.305
200	1.086	1.085	1.161	1.210	1.298
400	1.034	1.068	1.139	1.218	1.293

specific to a single score construction. Detailed per-group values for the simple score are shown in Table 8.

Table 7: Size-to-coverage empirical/oracle ratio at heterogeneity levels σ_Δ.

	Selected heterogeneity level σ_Δ				
n_{cal}	0.437	0.741	1.201	1.862	2.369
12	1.313	1.181	1.028	1.087	1.038
25	1.095	1.025	1.009	0.990	1.012
50	0.981	0.994	0.997	1.001	0.994
100	1.002	0.997	0.987	1.003	1.002
200	1.019	1.009	1.003	0.999	1.004
400	1.017	1.002	0.996	1.004	1.004

Table 8: Summary for `Bias in Bios` at $\alpha = 0.10$ for the simple score.

Quantity	Group (Male)	Group (Female)
p_g	0.5275	0.4725
q_g	0.6298	0.6531
$\varepsilon_g(q)$	0.0020	-0.0009
$\lambda_g = \ell_g(q_g)$	1.0140	1.0231
$\ell_g(q_g) - \ell_g(q)$	-0.0050	0.0053
τ_g	0.6410	0.6430
$\hat{F}_{S\|g}(\tau_g) - \hat{F}_{S\|g}(q_g)$	0.0016	-0.0018

Table 9: Split sizes and group composition for `Bias in Bios`.

Split	Total n	Group 0 (Male)	Group 1 (Female)
Model train	185,758	97,985 (52.75%)	87,773 (47.25%)
Model validation	20,640	10,886 (52.74%)	9,754 (47.26%)
Calibration	31,764	16,755 (52.75%)	15,009 (47.25%)
Test	79,397	41,881 (52.75%)	37,516 (47.25%)

D.2 Robustness across Target Coverage

We vary the target miscoverage level over $\alpha \in \{0.05, 0.07, 0.085, 0.10\}$ and keep the rest of the pipeline unchanged across three score families. The simple score gives the baseline pattern, while the same trade-off is preserved under SAPS and RAPS. Tables 10, 11 and 12 report the α-robustness summaries for simple, SAPS and RAPS.

D.3 Controlled Group Temperature Sweep

In this section, we construct a controlled heterogeneity view at the fixed target level $\alpha = 0.1$ by temperature scaling the class probability vector of group 1 through a specific temperature $T \in \{1.00, 1.10, 1.25, 1.50, 1.75, 2.00\}$ and renormalizing $p^{1/T}$. Adjusting the temperature parameter amplifies the same mechanism shown at the baseline. Parallel to the previous section, we present the simple, SAPS, and RAPS scores under temperature sweep, showing a score-robust conversion pattern in Tables 13, 14, and 15.

D.4 Robustness to Alternative Scores

The same conversion mechanism is also visible under the SAPS and RAPS scores. Table 16 shows that all three scores exhibit the same pattern: pooled threshold coverage discrepancy, nonzero set size distortion after applying the group-wise thresholds q_g, and nonzero coverage distortion after imposing an equalized set size.

D.5 Finite Calibration of the Pooled-Threshold Floor

This section complements the population uncertainty relations in Section 3, where the pooled-threshold floor obeys

$$\|\varepsilon(q)\|_{L^2(p)} \geq m_{\text{eff}}(q)\, \sigma_\Delta, \tag{49}$$

where q is the pooled threshold, q_g are the group-specific $1-\alpha$ quantiles, and $\sigma_\Delta = \mathrm{sd}(q_G)$. The question here is finite-sample rather than population-level: how large must the calibration split be before the structural floor in Equation (49) becomes empirically resolvable under split conformal calibration?

We fix the trained model outputs and vary only the calibration size. For each score choice (simple, SAPS, RAPS), we repeatedly form group-stratified subsamples of the original calibration split, recompute the pooled conformal threshold $\hat{q}_n$, and evaluate the resulting RMS group miscoverage on a fixed test set. Throughout, we set the target level $\alpha = 0.1$ and the number of groups $K = 2$, and we use 300 subsamples for each

$$n \in \{200, 500, 1000, 2000, 4000, 7000, 15000, 20000, 25000, 28000\}.$$

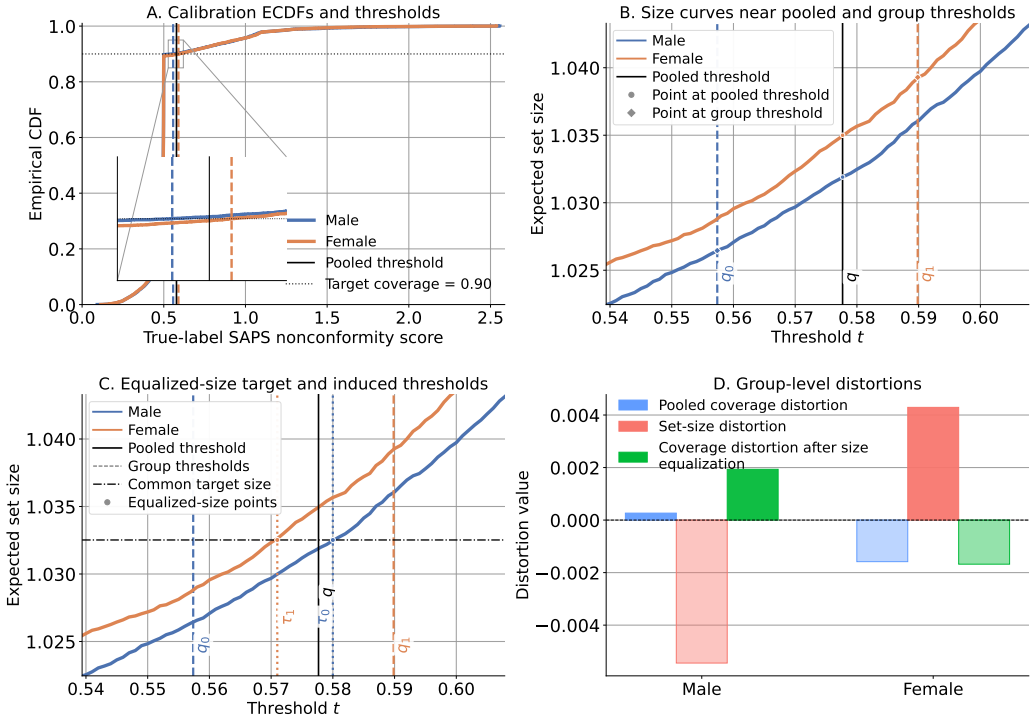


Figure 8: `Bias in Bios` mechanism view at $\alpha = 0.10$ for SAPS score. Panel A illustrates the pooled-threshold mechanism in Theorem 1; Panels B–C illustrate Theorems 3– 4 and Corollaries 1–2; Panel D summarizes the three distortions for male and female groups.

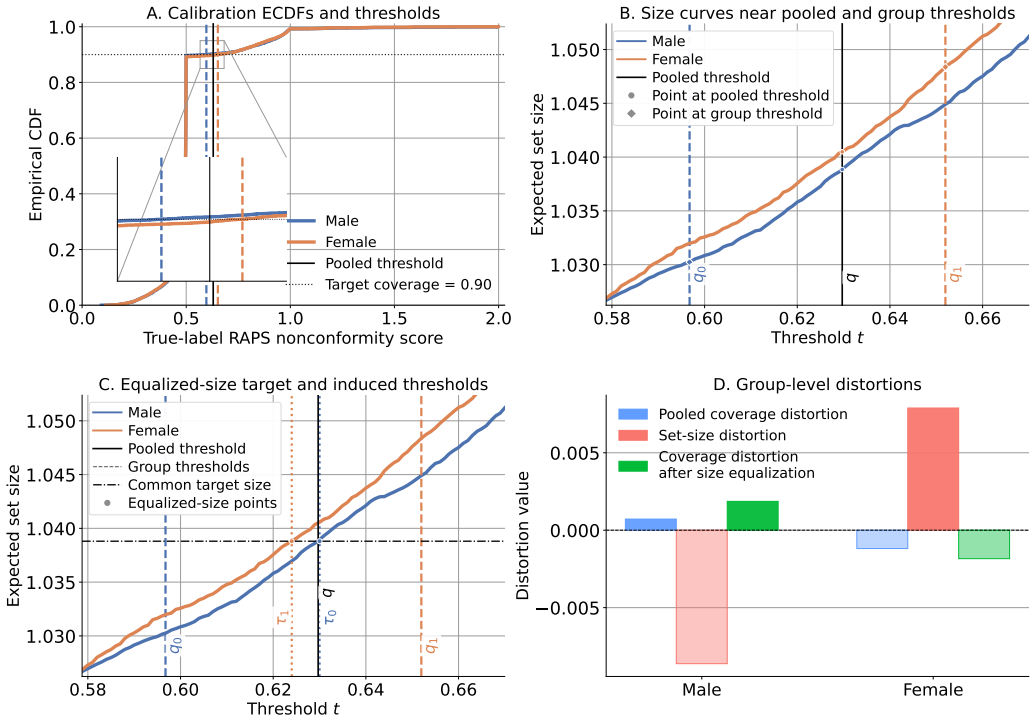


Figure 9: `Bias in Bios` mechanism view at $\alpha = 0.10$ for RAPS score. Panel A illustrates the pooled-threshold mechanism in Theorem 1; Panels B–C illustrate Theorems 3– 4 and Corollaries 1–2; Panel D summarizes the three distortions for male and female groups.

Table 10: α-robustness for the simple score.

α	σ_Δ	RMS pooled coverage	RMS size deviation from q_g	RMS coverage after equalized size
0.050	4.60×10^{-5}	0.0021	0.0002	0.0006
0.070	0.0019	0.0005	0.0017	0.0014
0.085	0.0054	0.0005	0.0035	0.0016
0.100	0.0116	0.0015	0.0051	0.0017

Table 11: α-robustness for SAPS.

α	σ_Δ	RMS pooled coverage	RMS size deviation from q_g	RMS coverage after equalized size
0.050	0.0039	0.0011	0.0047	0.0005
0.070	0.0022	0.0009	0.0019	0.0008
0.085	0.0015	0.0007	0.0009	0.0013
0.100	0.0162	0.0011	0.0049	0.0018

Let $q^\star$ denote the pooled threshold computed from the full test set, used as a population proxy. We then define

$$\mathrm{floor}^\star := \Big(\sum_{g=1}^{K} p_g^{\mathrm{test}}\, \varepsilon_g(q^\star)^2\Big)^{1/2}, \qquad \widehat{\mathrm{floor}}(n) := \Big(\sum_{g=1}^{K} p_g^{\mathrm{test}}\, \varepsilon_g(\hat{q}_n)^2\Big)^{1/2},$$

and summarize the signal-to-noise ratio by $\mathrm{SNR}(n) := \frac{\mathrm{floor}^\star}{\mathrm{sd}(\widehat{\mathrm{floor}}(n))}$. We interpret $\mathrm{SNR}(n) \geq 1$ as the point where the structural floor is at least as large as the calibration-induced standard deviation.

For the empirical estimation, the stiffness factor $m_{\mathrm{eff}}(q)$ is replaced by a local density proxy at $q^\star$, estimated from the pooled test scores by a window

$$\widehat{m_{\mathrm{eff}}}(q^\star) := \frac{1}{2hn_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \mathbf{1}\{|S_i - q^\star| \leq h\},$$

with a Silverman-type [23] bandwidth $h = 0.9\,\hat{\sigma}_{\mathrm{rob}}\, n_{\mathrm{test}}^{-1/5}$, $\hat{\sigma}_{\mathrm{rob}} = \min\big(\hat{\sigma},\, \mathrm{IQR}/1.34\big)$, where n_{test} is the test set size and $\hat{\sigma}$ is the standard deviation of test scores. Then, the estimated stiffness factor $\widehat{m_{\mathrm{eff}}}(q^\star)$ serves as an empirical detectability proxy for the local CDF slope entering Equation (49), not as an exact plug-in estimate of the group-specific quantity in Definition 1.

To motivate the detectability scale, we use a DKW-style empirical-CDF heuristic [7], treating the calibration scores within each group as IID. Under an infinitely exchangeable model, this IID step can also be read conditionally on the de Finetti directing measure [9]. This is stronger than the exchangeability assumption needed for CP and is used only for the sample-size diagnostic below. If n denotes the total calibration size and the groups are roughly balanced, each group has about n/K calibration scores. A union bound over the K group-wise empirical CDFs gives a uniform fluctuation scale of order

$$\sqrt{\frac{\log(K/\xi)}{n/K}} = \sqrt{\frac{K\log(K/\xi)}{n}}. \tag{50}$$

After quantile inversion, the corresponding threshold-noise scale is of order $\sqrt{\frac{K\log(K/\xi)}{n\, m_{\mathrm{eff}}(q)^2}}$. Requiring this noise scale to be no larger than the intrinsic quantile separation scale σ_Δ motivates

$$n \gtrsim \frac{K\log(K/\xi)}{m_{\mathrm{eff}}(q)^2 \sigma_\Delta^2}, \tag{51}$$

where K is the number of groups and ξ is a fixed confidence parameter. We choose $K = 2$ and $\xi = 0.05$ in the plots across the score choices.

Panels A–B of Figure 10 show a clear signal-resolution transition. The mean empirical floor decreases monotonically with n, while the detectability ratio rises steadily and crosses the $\mathrm{SNR}(n) = 1$

Table 12: α-robustness for RAPS.

α	σ_Δ	RMS pooled coverage	RMS size deviation from q_g	RMS coverage after equalized size
0.050	0.0055	0.0018	0.0121	0.0002
0.070	0.0005	0.0013	0.0008	0.0013
0.085	0.0007	0.0003	0.0006	0.0006
0.100	0.0276	0.0010	0.0083	0.0019

Table 13: `Bias in Bios` temperature sweep for the simple score at $\alpha = 0.10$.

T	σ_Δ	RMS pooled coverage	RMS size from q_g	RMS coverage after equalized size
1.00	0.0116	0.0015	0.0051	0.0017
1.10	0.0136	0.0018	0.0061	0.0015
1.25	0.0169	0.0027	0.0083	0.0015
1.50	0.0241	0.0044	0.0126	0.0015
1.75	0.0346	0.0067	0.0205	0.0022
2.00	0.0432	0.0088	0.0275	0.0028

threshold first for the simple score, and on the 1.5×10^4 scale for both SAPS and RAPS. The order matches Equation (51). In particular, the score with the steepest local CDF near q requires the fewest calibration points to resolve the floor.

Panels C–D of Figure 10 show that small calibration sets do not merely add variance. They also inflate both the observed lower bound and the empirical heterogeneity estimate $\hat{\sigma}_\Delta$, so limited calibration size may exaggerate the lower bound.

Finally, let

$$C_{\mathrm{det}} := \frac{n_{\mathrm{detect}}\, m_{\mathrm{eff}}(q)^2 \sigma_\Delta^2}{K \log(K/\xi)},\ \xi = 0.05.$$

Here, $m_{\mathrm{eff}}(q)$ is the local density proxy at the pooled test-proxy threshold, σ_Δ is the test-proxy heterogeneity of the group quantiles, $\mathrm{floor}^\star$ is the true-floor proxy, and n_{detect} is the smallest calibration size with $\mathrm{SNR}(n) \geq 1$. Table 17 shows that C_{det} is stable across all three scores up to a common multiplicative scale, which serves as a diagnostic of detectability.

In summary, Section 3 identifies a population lower bound driven by intrinsic heterogeneity σ_Δ and local stiffness $m_{\mathrm{eff}}(q)$. The present experiments show that observing the lower bound in finite-sample split conformal calibration is itself a sample-size-related phenomenon. Empirically, the relevant resolution scale is well summarized by $m_{\mathrm{eff}}(q)^2 \sigma_\Delta^2$. That is, once n is large enough relative to the inverse of the resolution scale, the lower bound becomes detectable and the empirical curves stabilize near the population proxy.

Table 14: `Bias in Bios` temperature sweep for SAPS at $\alpha = 0.10$.

T	σ_Δ	RMS pooled coverage	RMS size from q_g	RMS coverage after equalized size
1.00	0.0162	0.0011	0.0049	0.0018
1.10	0.0058	0.0009	0.0018	0.0020
1.25	0.0086	0.0015	0.0024	0.0020
1.50	0.0315	0.0028	0.0077	0.0067
1.75	0.0396	0.0056	0.0128	0.0063
2.00	0.0489	0.0107	0.0226	0.0107

Table 15: `Bias in Bios` temperature sweep for RAPS at $\alpha = 0.10$.

T	σ_Δ	RMS pooled coverage	RMS size from q_g	RMS coverage after equalized size
1.00	0.0276	0.0010	0.0083	0.0019
1.10	0.0121	0.0004	0.0034	0.0020
1.25	0.0080	0.0013	0.0024	0.0020
1.50	0.0355	0.0033	0.0098	0.0022
1.75	0.0584	0.0045	0.0150	0.0086
2.00	0.0683	0.0093	0.0248	0.0076

Table 16: `Bias in Bios` comparison across three scores at $\alpha = 0.10$.

Score	q	σ_Δ	RMS pooled coverage	RMS size from q_g	RMS coverage after equalized size
Simple	0.6423	0.0116	0.0015	0.0051	0.0017
SAPS	0.5777	0.0162	0.0011	0.0049	0.0018
RAPS	0.6297	0.0276	0.0010	0.0083	0.0019

Table 17: Detectability summary at $\alpha = 0.1$.

Score	$m_{\text{eff}}(q)$	σ_Δ	floor*	full-calibration floor	n_{detect}	C_{det}
RAPS	0.0564	0.0139	0.0010	0.0010	15000	0.0013
SAPS	0.0672	0.0094	0.0009	0.0011	15000	0.0008
Simple	0.1581	0.0094	0.0015	0.0016	7000	0.0021

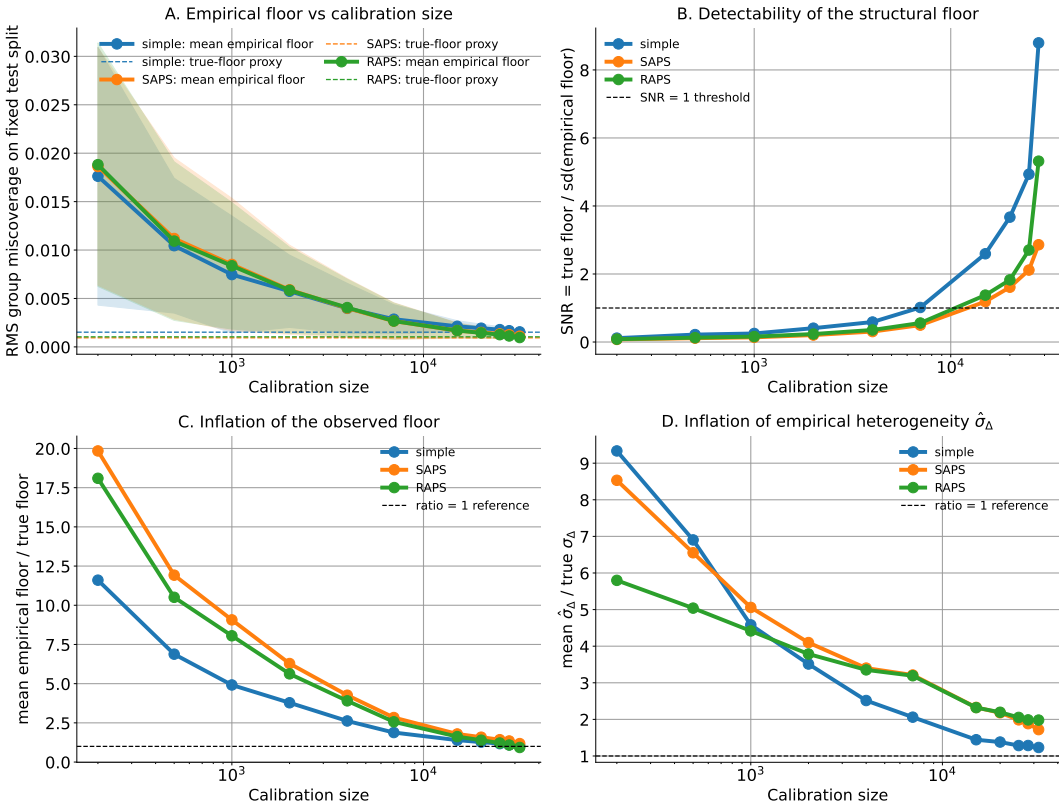


Figure 10: Finite-calibration detectability diagnostics of the pooled-threshold floor from Theorem 2 on `Bias in Bios`. Panels A–B show how the empirical floor and its detectability improve with the calibration size. The dashed line at $\mathrm{SNR}(n) = 1$ marks the empirical detectability benchmark. Panels C–D show that small calibration splits can also inflate the observed floor and estimated heterogeneity.

E `MultiNLI` Experiments

E.1 Experiment Settings

All post-hoc quantities are defined exactly as in Section 5.2 except for the dataset-specific details. We use the Hugging Face `MultiNLI` corpus[5], treat the ten genres as groups, and train a `DistilBERT` classifier for three-class NLI. The run uses a maximum sequence length of 256, a learning rate of 2×10^{-5}, a weight decay of 0.01, batch sizes of 64/32, a warm-up ratio of 0.05, and two epochs. The validation accuracy and macro-F1 are 0.8104 and 0.8100. The detailed split sizes and the summary of the key quantities are shown in Tables 18 and 19. In addition, Table 20 reports the full per-genre quantities at level $\alpha = 0.1$. The alternative score results for SAPS and RAPS are deferred to Section E.4.

E.2 Robustness across Target Coverage

Figure 11 shows that across $\alpha \in \{0.05, 0.07, 0.085, 0.10\}$, the empirical pooled-threshold distortion stays above the estimated lower-bound scale, while the induced size and coverage distortions remain nonzero throughout.

E.3 Controlled Genre Temperature Sweep

At $\alpha = 0.1$, we perturb only the `facetoface` genre via temperature scaling. Figure 12 shows that across the sweep, the empirical pooled-threshold distortion remains above the estimated lower bound, and the induced size and coverage distortions stay nonzero.

E.4 Alternative Scores

The same mechanism persists under SAPS and RAPS with the model, groups, and split protocol fixed. SAPS uses the default temperature $T = 1.0$ and $\lambda_{\text{SAPS}} = 0.3$; RAPS uses the default temperature $T = 0.6$, $k_{\text{reg}} = 1$, and $\lambda_{\text{RAPS}} = 0.02$. Both scores utilize the randomization level $u = 0.5$. Figure 13 demonstrates that at $\alpha = 0.1$, the SAPS score shows a pooled floor, a nonzero set size distortion after switching from pooled q to group-wise q_g, and a nonzero coverage distortion after imposing the equalized expected set size. Robustness checks are provided in Figures 14 and 15. We do not view these results as requiring pointwise lower-bound dominance at every α or temperature value. For SAPS and RAPS, Panel A is best interpreted as a finite-sample diagnostic based on an estimated lower-bound proxy, while the main evidence remains visible.

[5]Hugging Face dataset: `nyu-mll/multi_nli`; https://huggingface.co/datasets/nyu-mll/multi_nli. The dataset card states that most of the corpus is released under the OANC license, while the FICTION section includes Creative Commons Attribution 3.0, Creative Commons Share-Alike 3.0, and U.S. public-domain materials.

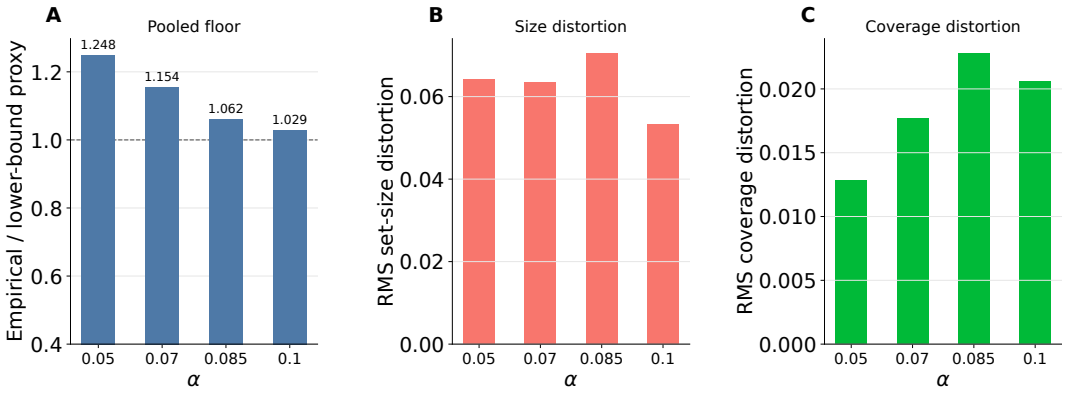


Figure 11: `MultiNLI` robustness across α for the simple score. Panel A displays Theorem 2: the empirical pooled-threshold distortion stays near or above the lower-bound scale across the tested α-grid. Panels B–C illustrate Corollaries 1–2: the induced set-size and coverage distortions remain nonzero throughout.

Table 18: Split sizes for `MultiNLI`.

Split	n
Model train	353,431
Model validation	39,271
Calibration	9,823
Test	9,824

Table 19: Summary for `MultiNLI` at $\alpha = 0.10$ with simple score.

σ_Δ	RMS pooled coverage	λ	RMS size distortion	σ_λ	RMS coverage distortion
0.0354	0.0150	1.2717	0.0532	0.0779	0.0209

E.5 Detectability under Finite Calibration

This subsection follows Appendix D.5 with the same finite calibration protocol. Because `MultiNLI` has fewer genre-stratified calibration examples than Bias in Bios, we use the grid

$$n \in \{50, 100, 150, 200, 250, 300, 350, 400, 450, 500\}.$$

We fix the trained `MultiNLI` outputs and test split, vary only the calibration size through genre-stratified subsampling, recompute the pooled threshold on each subsample, and evaluate the resulting RMS genre-wise miscoverage on the fixed test set. Figure 16a shows that the floor becomes visible around $n = 100$, and is clearly resolved for all three scores thereafter. Hence, on the `MultiNLI` dataset, the structural signal is empirically visible once the calibration split is moderately large.

Table 20: Per-genre summary for `MultiNLI` at $\alpha = 0.1$ with simple score.

Genre	p_g	q_g	$\varepsilon_g(q)$	λ_g	$\lambda_g - \ell_g(q)$	τ_g	$\hat{F}_{S\|g}(\tau_g) - \hat{F}_{S\|g}(q_g)$
Government	0.0989	0.7215	0.0198	1.1420	-0.0905	0.8200	0.0463
Letters	0.1006	0.7561	0.0119	1.1741	-0.0304	0.8360	0.0192
Travel	0.1007	0.7747	0.0090	1.2285	-0.0303	0.8030	0.0081
Oup	0.0998	0.7802	0.0082	1.2439	-0.0173	0.8050	0.0051
Verbatim	0.0990	0.8151	0.0044	1.3464	0.0380	0.7730	-0.0195
Fiction	0.1004	0.8371	0.0037	1.3398	0.0659	0.7910	-0.0132
Facetoface	0.1006	0.8223	0.0008	1.3168	0.0506	0.7960	-0.0101
Telephone	0.1001	0.7951	-0.0150	1.3001	0.0020	0.7810	-0.0061
Nineeleven	0.1005	0.7883	-0.0155	1.2249	-0.0091	0.8220	0.0132
Slate	0.0996	0.8410	-0.0329	1.4008	0.0982	0.7750	-0.0307

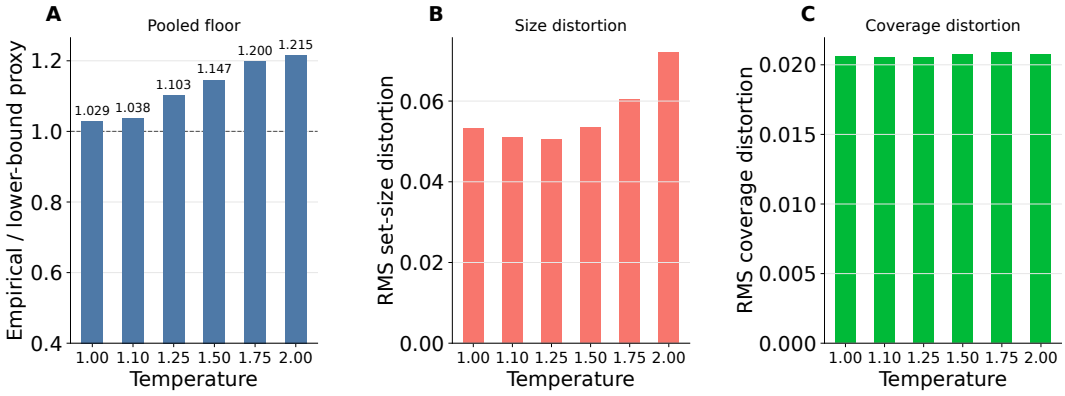


Figure 12: Controlled `MultiNLI` temperature sweep at $\alpha = 0.10$ for the simple score, perturbing the *facetoface* genre only. Panel A corresponds to Theorem 2. Panels B–C show Corollaries 1–2. The same trade-off mechanism remains visible across the temperature sweep.

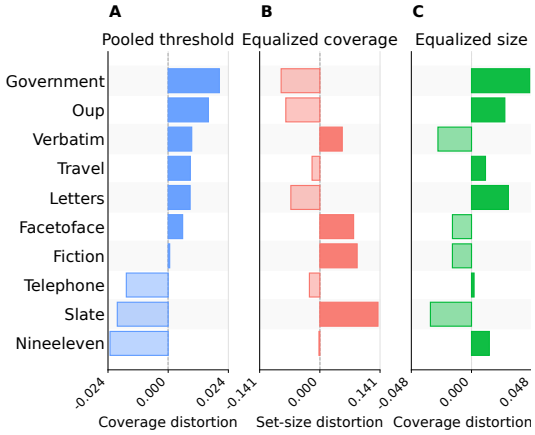


Figure 13: `MultiNLI` at $\alpha = 0.10$ using SAPS score. For this score, Panel A shows pooled quantile consequence of Theorem 1; Panels B and C illustrate the set size distortion in Corollary 1, and the coverage distortion in Corollary 2, respectively.

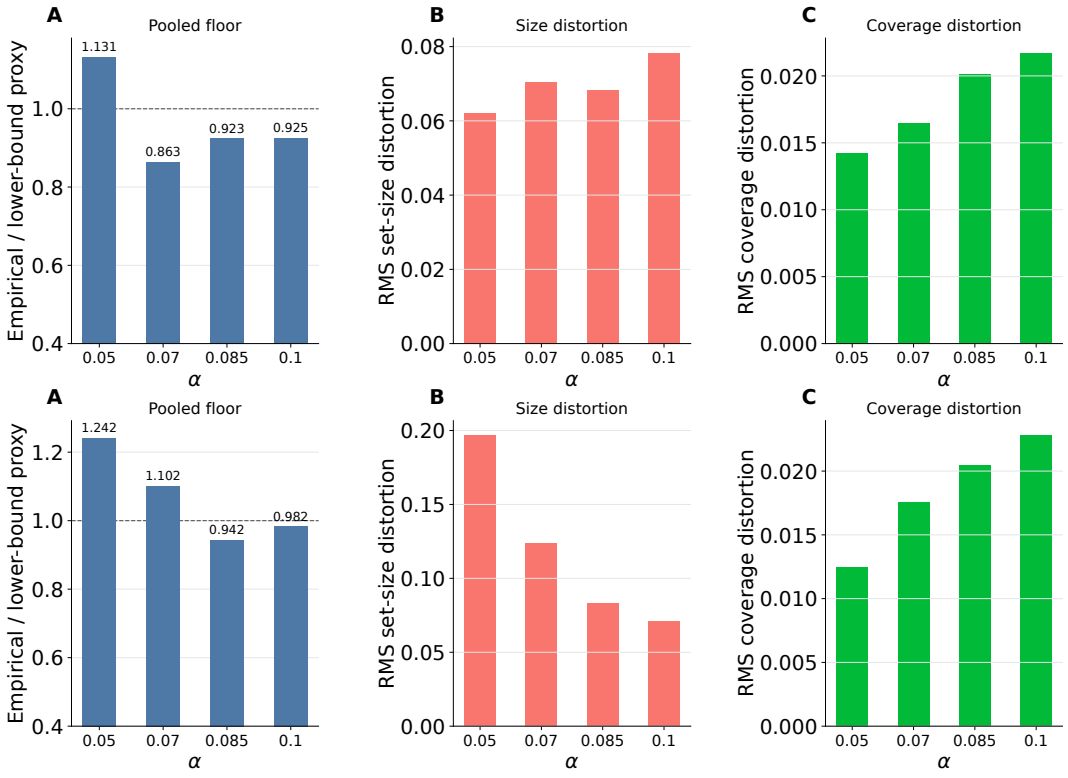


Figure 14: `MultiNLI` robustness across α for SAPS (top) and RAPS (bottom) scores. In each row, Panel A is best read as a finite-sample diagnostic for Theorem 2 based on an estimated lower-bound proxy, rather than a pointwise lower-bound verification. Panels B–C show Corollaries 1–2. The induced set-size and coverage distortions remain visible across the tested α-grid.

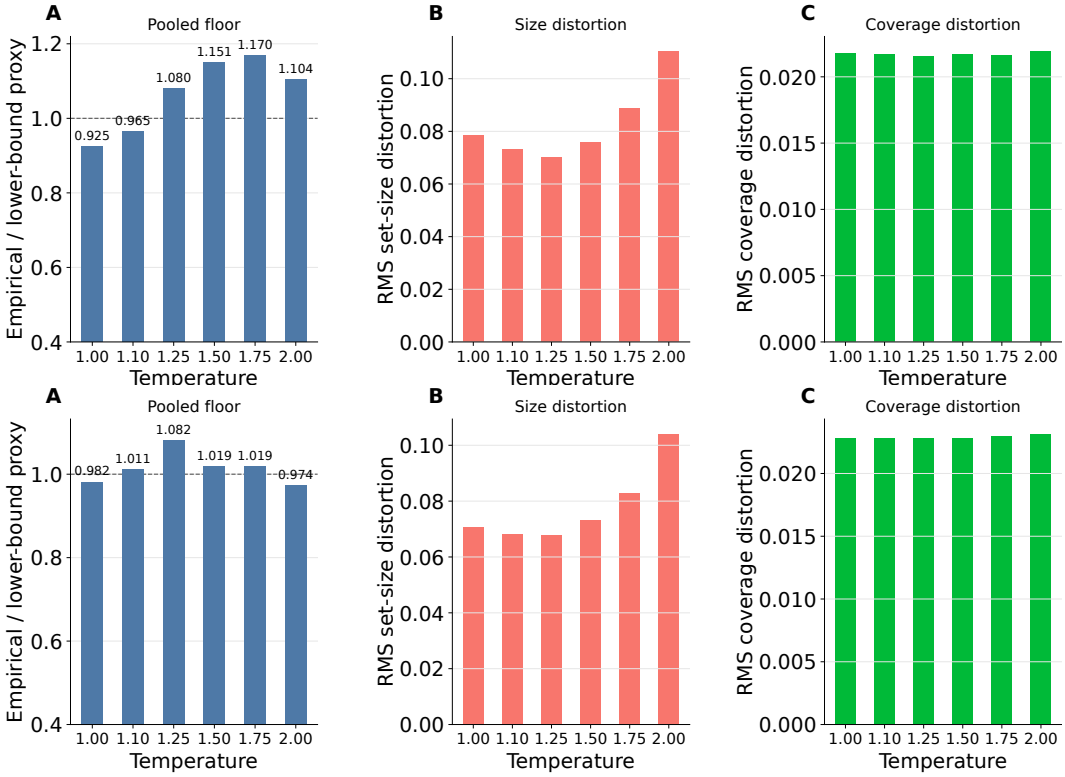
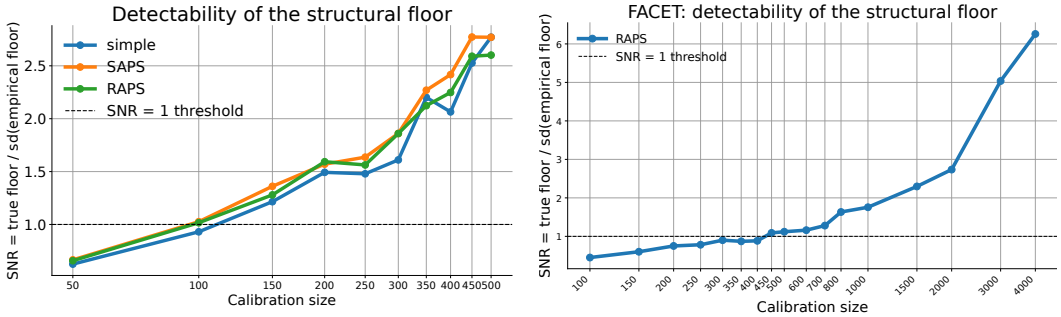


Figure 15: Controlled `MultiNLI` temperature sweep at $\alpha = 0.10$ for SAPS (top) and RAPS (bottom) scores, perturbing the *facetoface* genre only. Panel A is best read as a finite-sample diagnostic for Theorem 2. Panels B–C illustrate Corollaries 1–2. The same trade-off mechanism remains visible across the temperature sweep.



(a) (b)

Figure 16: Finite-calibration detectability of the pooled-threshold floor from Theorem 2. Left: `MultiNLI`, for simple, SAPS, and RAPS scores. Right: `FACET`, for RAPS score. The dashed line at $\mathrm{SNR}(n) = 1$ marks the empirical detectability benchmark, so crossing it indicates that the structural floor is becoming distinguishable from finite-sample fluctuation.

F FACET Experiments

F.1 Experiment Settings and Summary

For FACET,[6] we use CLIP ViT-L/14.[7] All post-hoc quantities are defined exactly as in Sections 5.2 and 5.3. Tables 21, 22 and 23 list the key quantities, the full per-group quantities, and the experiment configuration, respectively.

F.2 Robustness across Target Coverage and Temperature Perturbation

Figure 17 shows that for various target levels $\alpha \in \{0.05, 0.07, 0.085, 0.10\}$, the empirical pooled-threshold distortion stays above the empirical lower bound. The induced set size distortion remains nonzero throughout and the equalized expected set size policy continues to produce a nonzero RMS coverage distortion.

Next, we perturb only the Younger group through temperature scaling while keeping the rest of the groups fixed. Figure 18 shows that across perturbations, the pooled RMS gap remains at or above the lower bound proxy. The equalized expected set size policy continues to introduce a visible coverage distortion across the sweep.

[6] We use the FACET benchmark dataset released by Meta AI; official dataset page: https://ai.meta.com/datasets/facet/. The FACET data card in the ICCV 2023 supplemental material states that the license is "Custom license, see dataset download agreement," and that FACET is intended for evaluation only. For a description of the benchmark, see [13].

[7] CLIP ViT-L/14 from OpenAI repository: https://github.com/openai/CLIP. The repository LICENSE file states that the code is released under the MIT License.

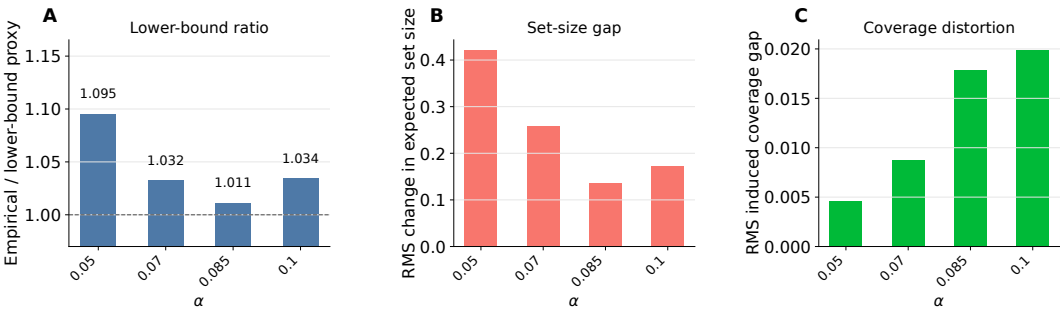


Figure 17: FACET robustness across α for the RAPS score. Panel A illustrates Theorem 2: the empirical pooled-threshold distortion stays near or above the lower-bound scale across the tested α-grid. Panels B–C show Corollaries 1–2: the induced set-size and coverage distortions remain nonzero throughout.

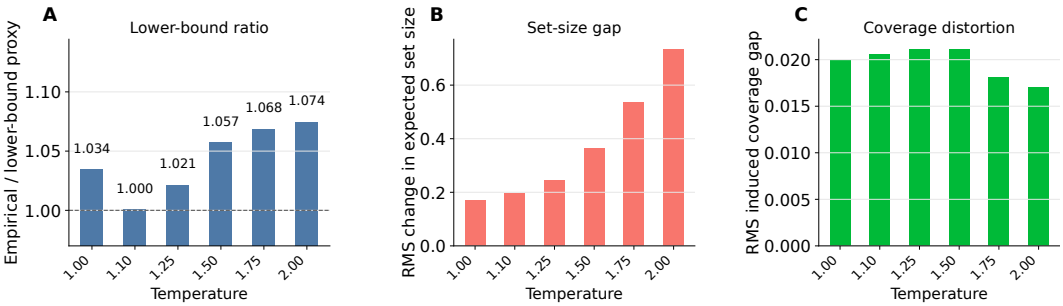


Figure 18: Controlled FACET temperature sweep at $\alpha = 0.10$ for the RAPS score, perturbing the Younger group only. Panel A evaluates the lower-bound behavior in Theorem 2. Panels B–C show Corollaries 1–2. The same trade-off mechanism remains visible across the temperature sweep.

Table 21: FACET RAPS summary at $\alpha = 0.10$.

Metric	Value
q	1.0194
σ_Δ	0.0040
Pooled RMS	0.0083
Lower bound	0.0080
Ratio	1.034
λ	2.1735
RMS set size	0.1717
σ_λ	0.2817
RMS coverage	0.0199

Table 22: Per-group quantities for FACET at $\alpha = 0.1$.

Age group	p_g	q_g	$\varepsilon_g(q)$	λ_g	$\lambda_g - \ell_g(q)$	τ_g	$\hat{F}_{S\|g}(\tau_g) - \hat{F}_{S\|g}(q_g)$
Younger	0.1795	1.0177	0.0176	1.8757	-0.0892	1.0200	0.0054
Middle	0.5718	1.0196	-0.0035	2.0921	0.0182	1.0200	0.0025
Older	0.0461	1.0111	-0.0105	1.9895	-0.3421	1.0171	0.0263
Unknown	0.2026	1.0276	-0.0018	2.7090	0.3329	1.0126	-0.0419

F.3 Detectability under Finite Calibration

Following the detectability analysis in Appendices D.5 and E.5, we fix the model outputs and test split, vary only the calibration sample size through age-stratified subsampling, recompute the pooled threshold on each subsample, and evaluate the resulting RMS group-wise coverage on the fixed test split. For FACET, we use the grid

$$n \in \{100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 1000, 1500, 2000, 3000, 4000\}.$$

Figure 16b illustrates the signal-to-noise ratio $\mathrm{SNR}(n) = \mathrm{floor}^\star / \mathrm{sd}(\widehat{\mathrm{floor}}_n)$. On FACET, the ratio crosses 1 at about $n = 450$ and then increases. Although FACET exhibits substantial group-size imbalance, the same detectability pattern is still observed. In particular, group imbalance affects finite-sample variability but does not eliminate the signal itself, which highlights that the trade-off mechanism is intrinsic.

F.4 Stability under Calibration Resampling

We verify that the trade-off mechanism observed on the FACET dataset is not an artifact of a single calibration split. We keep the model outputs and the test set fixed, and repeat age-stratified calibration subsampling at $\alpha = 0.1$. Concretely, we perform 20 repetitions using 70% of the full calibration size. For each repetition, we calculate the key quantities as in Section 5.4.

Table 24 demonstrates that the resampling means remain close to the values with full calibration size. The pooled RMS coverage gap stays near 0.008, the induced RMS set-size gap remains around 0.18, and the equalized set size policy continues to produce a nonzero RMS coverage gap of about 0.019. Therefore, the empirical trade-off pattern in Section 5.4 is stable under repeated calibration perturbations.

G Additional Robustness Analysis

G.1 Repeated-split and cross-fitted robustness

We evaluate the primary score settings—SIMPLE for Bias in Bios and MultiNLI, and RAPS for FACET—over 500 group-stratified calibration/test splits. Table 25 reports the split mean and the 2.5th–97.5th split percentiles.

The same bidirectional transfer persists across splits. For FACET, the repeated-split size-to-coverage mean is 0.944 pp, compared with 1.992 pp on the original fixed split; we therefore base the robustness

Table 23: Configuration for FACET with RAPS.

Setting	Value
Dataset	FACET
Groups	Younger, Middle, Older, Unknown
Labels	20 occupations
Vision-language model	CLIP ViT-L/14
Image preprocessing	bbox crop with 0.08 expansion
Validation samples	4,122
Calibration samples	11,776
Test samples	4,122
Primary target α	0.10
RAPS temperature	0.60
RAPS λ	0.02
RAPS k_{reg}	1
Temperature-perturbed group	Younger
Temperature values	1.0, 1.1, 1.25, 1.5, 1.75, 2.0
Test accuracy / macro-F1	0.6994 / 0.7369

Table 24: FACET stability under calibration subsampling at $\alpha = 0.10$.

Metric	Original	Resampling mean $\pm$ sd
Pooled RMS coverage gap	0.0083	0.0085 ± 0.0005
RMS set-size gap after $q \to q_g$	0.1717	0.1828 ± 0.0311
RMS coverage gap after equalized size	0.0199	0.0186 ± 0.0044

interpretation on the more conservative repeated-split distribution. Two-fold cross-fitting changes the size-to-coverage means by only 0.1%, 0.7%, and 2.5% on Bias in Bios, MultiNLI, and FACET, respectively. The repeated-split and cross-fitted analyses consistently reproduce the bidirectional coverage–size transfer, showing that it is robust to calibration/test resplits.

G.2 Practical scale of the observed disparities

Under the same score settings and 500 splits, the mean max-minus-min group ranges on Bias in Bios, MultiNLI, and FACET are, respectively: pooled-coverage gaps of 0.33, 5.13, and 3.30 pp; Equalized Coverage expected-set-size gaps of 0.009, 0.281, and 0.714 labels; and coverage-change gaps from moving to Equalized Expected Set Size of 0.33, 7.21, and 3.01 pp.

For context, Cresswell et al. [4] associate a 0.28-label increase in average set size with a 2.2 pp decrease in human accuracy. More directly, for FACET under marginal CP, Cresswell et al. [5] report a maximum cross-group set-size gap of 0.75 labels and a 3.23 pp maximum cross-group gap in the human accuracy improvement provided by prediction sets. Our FACET expected-set-size gap of 0.714 labels is therefore nearly identical in scale to the set-size disparity observed in that human study. Thus, the MultiNLI and FACET gaps occur at scales that prior human experiments have associated with measurable differences in downstream accuracy or in the accuracy benefit provided by prediction sets.

G.3 Independent population validation of the lower bounds

For the synthetic results, nonconformity scores are generated directly from the analytic data-generating models. We independently compute each population distortion and its theoretical lower bound: the bound uses analytic CDFs and densities, population quantiles, and expected-set-size slopes, whereas the distortion is evaluated separately. Any population left-hand-side-to-bound ratio below one would falsify the corresponding inequality.

All 131 independently evaluated population configurations satisfy their corresponding bounds, with the Equalized Coverage expected-set-size bound nearly attained in the tightest cases.

Table 25: Repeated-split robustness over 500 group-stratified calibration/test splits. Entries are mean [2.5th, 97.5th split percentiles]. Coverage quantities are reported in percentage points (pp). The intervals are descriptive split percentiles conditional on the fixed predictor and available held-out pool; they are not population confidence intervals.

Dataset	Pooled CP: groupwise coverage distortion (pp)	Equalized Coverage CP: groupwise expected set size disparity (labels)	Equalized Expected Set Size CP: groupwise coverage distortion (pp)
Bias in Bios	0.245 [0.106, 0.462]	0.0049 [0.0002, 0.0118]	0.166 [0.009, 0.428]
MultiNLI	1.521 [1.089, 2.011]	0.0591 [0.0400, 0.0782]	2.027 [1.481, 2.581]
FACET	1.000 [0.379, 1.680]	0.1490 [0.0675, 0.2403]	0.944 [0.483, 1.503]

Table 26: Independent population validation of the theoretical lower bounds.

Policy	Configurations	Minimum ratio	Median ratio
Pooled CP: coverage RMS	100	1.000	1.212
Equalized Coverage CP: size RMS	16	1.001	1.125
Equalized Expected Set Size CP: coverage RMS	15	1.546	1.686

H Computational Resources

All experiments were conducted on a Windows 11 computer with an Intel Core i9-12900H CPU, an NVIDIA GeForce RTX 3080 Ti Laptop GPU (16 GB VRAM), and 32 GB RAM. Each reported experiment was completed within one hour.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately state the paper's main claims, and these are supported by the theoretical results in Section 3 and the experiments in Section 5. The scope is also stated appropriately through the policy families studied here.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses its main limitations in the conclusion, including that it does not address full conditional coverage or prescribe a normative fairness criterion, that the quantitative results rely on local regularity assumptions, and that the analysis is developed for prespecified discrete groups and population score distributions, with finite-sample behavior studied empirically.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states and explicitly references the assumptions for each main theoretical result in Section 3, and complete proofs are provided in Appendix B. The supplemental technical discussion is also provided with assumptions and corresponding proofs in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient information to reproduce the main experiments including the datasets, group definitions, model choices, data splits, score selections, and calibration protocols in Section 5 and Appendices C, D, E, and F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplemental material includes a code package with script, dependency specifications, exact reproduction commands, and dataset preparation instructions for the main experiments. Due to dataset size, the raw datasets are not included.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the main experimental details needed to understand the results including the datasets, group definitions, model choices, data splits, score selections, and calibration protocols in Section 5 and Appendices C, D, E, and F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The synthetic experiments report averages over 40 Monte Carlo seeds, and the finite-calibration detectability experiments report variability across repeated calibration subsamples. `FACET` stability is additionally evaluated under repeated calibration resampling. However, the experiments are designed to demonstrate the fundamental trade-off mechanism under finite-sample calibration, rather than to provide benchmark-style performance comparisons with systematic variability reporting.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the hardware used for the experiments, including CPU, GPU, memory, and operating environment. The paper also states that all experiments were run on a single local machine. Each reported experiment completed within one hour. Details are provided in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research utilizes existing public datasets and standard pretrained models. It does not involve new human-subject experimentation or crowdsourcing. We are not aware of any aspect of the work that conflicts with the NeurIPS Code of Ethics.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper studies fairness-related calibration choices in conformal prediction. A positive impact is that it clarifies the structural cost of choosing between group-wise coverage and expected set-size policy. In addition, the paper is primarily foundational and does not prescribe a normative fairness policy or study a specific deployment setting.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new model, dataset, or other asset with elevated misuse risk. It studies calibration behavior using existing datasets and standard pretrained models, so special release safeguards are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original sources of the existing datasets and pretrained models it uses, and it explicitly states the corresponding asset identifiers, URLs, and license or terms-of-use information in the appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets, models, or other new assets. It develops theoretical results and validates them empirically using existing datasets and pretrained models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or new research with human subjects. Its experiments analyze existing datasets and pretrained models, including Bias in Bios, MultiNLI, and FACET.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or new research with human subjects. Its experiments analyze existing datasets and pretrained models, so IRB approval or equivalent review is not applicable to this work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology of the paper does not use an LLM as an important, original, or non-standard component. The paper develops theoretical results for conformal calibration trade-offs and validates them empirically using existing models and datasets.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.