JOURNAL OF SOCIAL SCIENCES, HUMANITIES AND INNOVATION

Vol 2 Issue 1 (2026)

http://jsshi.org/

Resource-Aware Fair Conformal Prediction for Stu-dent Early Warning Systems

Katherine Price[1], Daria Iyer[2], Julia Mitchell[3]

[1,2,3]College of Education, North Carolina State University, NC, United States

Corresponding author: Katherine Price. kp28301@ncsu.edu

Keywords:
Index Terms—early warning systems, conformal prediction, algo-rithmic fairness, resource allocation, learning to defer, educational data mining

Abstract:
Early-warning systems in schools and universities increasingly attach distribution-free guarantees to their decisions, bounding within every demographic group how often at-risk students are missed and how often students on track are flagged. Equalized conformal decision layers achieve this by referring ambiguous students to counselors, but they treat counselor workload as a by-product: at a 10% error level they refer 32% to 56% of a cohort, and because empty prediction sets are referred as well, relaxing the error level can even increase the workload. We propose Resource-Aware Fair Conformal Prediction (RAFCP), which takes the review capacity as an input and returns the strongest equalized guarantee that this capacity can support. RAFCP resolves empty prediction sets automatically, which pre-serves the group-level miss and false-flag bounds and makes re-ferrals nested in the error level; it certifies the level with a boot-strap capacity planner that uses only cross-fitted training scores, so finite-sample validity is retained; and it reconciles the plan with the screened cohort so that exactly the budgeted number of students is referred. We further prove that equal error levels across groups are minimax-optimal under any monotone capacity constraint. On three public U.S. cohorts (Tennessee STAR, High School and Beyond, and the LSAC Bar Passage Study) with 30 stratified splits, RAFCP lowers the worst-group automated error at a 10% referral budget from 0.43, 0.38, and 0.58 under confidence-based deferral to 0.19, 0.28, and 0.22, keeps between-group gaps below 0.02, and outperforms each deferral baseline in at least 27 of 30 paired splits.

1.INTRODUCTION

Early-warning systems (EWS) have become part of how schools and universities in the United States organize student support. Districts combine attendance, behavior, and course performance to identify students who may not graduate on time [1], [2], and colleges score first-year students for dropout risk so that advising can begin before a student disengages [3]. The value of such a system lies less in the accuracy of its ranking than in what follows a flag: a counselor reviews the case, a tutor is assigned, or a family is contacted. Each of these actions draws on capacity that is fixed for the term, and the number of students a counseling office can examine closely is usually far smaller than the number a model would like to discuss. The same systems operate on populations shaped by persistent gaps in achievement and access, and a model fitted to such data can turn those gaps into unequal error rates [4], [5]. When base rates differ between groups, a single thresholded score cannot be calibrated within groups and equalize error rates at the same time [6], [7], and post-processing that equalizes odds [8] does so by moving errors between students rather than removing them. A more promising route gives the decision an explicit option to defer. Conformal prediction turns any score into set-valued predictions with finite-sample coverage under exchangeability [9], [10], and its Mondrian, or category-conditional, variants provide coverage inside prespecified cells [11]–[13]. When the cells are indexed jointly by group and outcome, a singleton prediction set can be acted on automatically, an ambiguous set can be passed to a counselor, and the probability of missing an at-risk student or of flagging a student on track is bounded within every group. These guarantees are paid for with referrals, and the number of referrals is not controlled. In our reproduction of such an equalized layer (Section VI-A), an error level of 10% refers 31.7%, 55.6%, and 38.6% of the students in three public U.S. cohorts. No counseling office reviews half of a cohort in depth, so in practice some referred students receive a default action, and the way this default is chosen decides who bears the errors. The obvious remedies reintroduce the disparities that the guarantee was meant to remove: cutting the referral list by model confidence leaves more than half of the on-track Non-White law students flagged, and confidence-based deferral at the same workload, the setting in which selective classification is known to magnify disparities [14], does no better. A second problem is structural. Because empty prediction sets are referred as well, the workload is not monotone in the error level; relaxing the level first shortens the referral list and then lengthens it again, so that a counseling office able to review 5% of a cohort finds no feasible level at all on two of the three cohorts. This paper reverses the question. Instead of fixing the error level and accepting whatever workload results, we fix the review capacity and ask for the strongest equalized guarantee that it can support, delivered at exactly the budgeted workload. The resulting method, Resource-Aware Fair Conformal Prediction (RAFCP), builds on an existing equalized conformal early-warning framework (Section III), and its pipeline is shown in Fig. 1. The paper makes the following contributions. (1) We show that empty prediction sets can be resolved automatically without weakening any group-level bound on misses or false flags, and that the resolved layer has nested referral sets, so its workload decreases monotonically in the error level. From this monotonicity we prove that equal error levels across groups are minimax-optimal under any monotone capacity constraint: protecting one group never requires a weaker guarantee for another. (2) We design a capacity planner that bootstraps the workload of the deployed layer from cross-fitted training scores and selects the smallest level whose workload fits the budget with probability of about $1 - \delta$. Because the level never touches the calibration split, the finite-sample guarantees remain exact, and a reconciliation step fills or trims the referral list to exactly the budget. (3) On Tennessee STAR, High School and Beyond, and the LSAC Bar Passage Study, under budgets from 5% to 30% of the cohort, RAFCP attains the lowest worst-group automated error of all

compared methods in every setting, keeps between-group gaps below 0.02, and extends to intersectional groups and to priority weights that trade misses against false flags. (4) We also report the costs: a pooled error up to 0.023 higher than that of confidence-based deferral, overflow of the plan in 3% to 17% of the splits, and no measurable gain over a workload-proportional allocation when groups have similar base rates. The remainder of the paper is organized as follows. Section II reviews related work, Section III recalls the equalized conformal layer and states the capacity problem, Section IV presents RAFCP and its guarantees, Sections V and VI describe the experiments and results, Section VII discusses implications and limitations, and Section VIII concludes.

II. Related Work

A. Early Warning and Algorithmic Fairness in Education

Bowers et al. compared 110 dropout flags from 36 studies with receiver operating characteristic analysis and found wide differences in accuracy, with longitudinal growth models and grade-based indicators among the best [1]. Lakkaraju et al. developed a machine learning framework with U.S. school districts to identify students at risk of not graduating on time and emphasized ranking students by risk so that schools can prioritize interventions [2]. Yu et al. found that adding protected attributes to college dropout models leaves overall performance essentially unchanged and only marginally improves fairness [3]. Baker and Hawn reviewed the evidence of algorithmic bias across educational applications [4], and Gardner et al. introduced slicing analysis with the absolute between-ROC area to audit the differential accuracy of student models [5]. Mitigation usually acts on point predictions, for instance by post-processing group-specific thresholds to equalize odds [8], and it is limited by the impossibility of combining calibration within groups with balanced error rates [6], [7]. None of these works treats the downstream review capacity as part of the decision problem.

B. Conformal Prediction with Conditional Guarantees

Split conformal prediction needs only a held-out calibration set to turn a score into prediction sets with marginal coverage [9], [10]. Vovk analyzed conditional forms of validity, including label-conditional and category-conditional (Mondrian) predictors [11]; Sadinle et al. derived least ambiguous set-valued classifiers with class-specific error control [12]; and Romano et al. introduced equalized coverage across protected groups [13]. A related line controls risks other than miscoverage: risk-controlling prediction sets bound a monotone loss with high probability [15], conformal risk control bounds its expectation [16], and conformal selection controls the false discovery rate of a shortlist [17]. These methods tune one parameter to control a label-dependent risk. The constraint in our problem differs in kind: the workload of a decision rule does not depend on outcomes, so it can be evaluated for the cohort to be screened before any label is observed, while the label-dependent guarantees must stay exact as the workload is pinned to a budget. Group-conditional conformal prediction under sparse multi-horizon data provides a valuable statistical framework for auditing uncertainty reliability separately across underrepresented groups [31].

C. Deferral and Human–AI Decision Making

Learning to defer trains a model to hand cases to a downstream decision maker and can improve both accuracy and fairness [18], and Mozannar and Sontag gave consistent surrogate losses for this setting [19]. Jones et al. showed that selective classification, which abstains on low-confidence examples, can magnify accuracy disparities between groups [14]. Straitouri et al. used conformal prediction sets to restrict the label choices offered to human experts [20]. In these works the cost of deferral enters as a penalty or as a coverage target. In an EWS it is a hard

staffing limit, and the fairness requirement concerns the automated decisions made for every student whom the counselors cannot see.

III. Preliminaries and Problem Setting

A. Setting

Each student is described by features X, a group label A such as race or ethnicity, and a binary outcome $Y \in \{0,1\}$, where $Y = 1$ denotes the adverse outcome that support should prevent. Protected attributes are not model inputs. A decision rule $d(x,a)$ returns FLAG (proactive support), NONE (routine monitoring), or REFER (counselor review). For every group a we write $m_a = P(d = \text{NONE} \mid Y = 1, A = a)$ for the miss rate, $f_a = P(d = \text{FLAG} \mid Y = 0, A = a)$ for the false-flag rate, and $R = P(d = \text{REFER})$ for the referral rate. A cohort of N students with group sizes N_a is screened at once, and the counseling office can review at most $K = \lfloor BN \rfloor$ of them, where B is the referral budget.

B. Equalized Conformal Layer

We build on TRUE-EWS [21], an early-warning framework that separates the reliability of the communicated risk from the validity of the decisions, and adopt three of its components unchanged. First, a gradient-boosted decision tree f is trained on a training split $D_{\#}up\{tr\}$ without protected attributes, and five-fold cross-fitted scores $s_i^{\#}up\{oof\}$ are computed on $D_{\#}up\{tr\}$. Second, the communicated risk is $\hat{p}(x,a) = \mu_a(f(x))$ with the group-wise Beta calibration map [22]

$$\mu_a(s) = \sigma(c_a + a_a \ln s - b_a \ln(1 - s)), \qquad (1)$$

fitted by logistic regression on the cross-fitted scores of group a under the constraint $a_a, b_a \geq 0$, which keeps it strictly increasing. Third, the equalized conformal layer (ECC) partitions a calibration split $D_{\#}up\{cal\}$ into cells $I_{a,y} = \{i : A_i = a, Y_i = y\}$ of size $n_{a,y}$ and scores each example with the least ambiguous nonconformity score [12]

$$V(x,a,y) = 1 - \hat{p}_y(x,a), \qquad (2)$$

where $\hat{p}_1 = \hat{p}$ and $\hat{p}_0 = 1 - \hat{p}$. The threshold of every cell is the k-th smallest of its scores,

$$\hat{q}_{a,y} = V_{(k)}, \quad k = \lceil (n_{a,y} + 1)(1 - \alpha_y) \rceil, \qquad (3)$$

with $\hat{q}_{a,y} = +\infty$ if $k > n_{a,y}$. The prediction set is $C(x,a) = \{y : V(x,a,y) \leq \hat{q}_{a,y}\}$; ECC flags a student when $C = \{1\}$, takes no action when $C = \{0\}$, and refers every other case, including the empty set. If calibration and test examples are exchangeable within every cell, then $P(y \in C(X,A) \mid A = a, Y = y) \geq 1 - \alpha_y$, and therefore $m_a \leq \alpha_1$ and $f_a \leq \alpha_0$ in every group, in finite samples and for any score model. Because the sets depend on scores only through ranks within cells, they are also unchanged by any strictly increasing recalibration.

C. Two Thresholds and a Gap

Write $\tau_{a,1} = 1 - \hat{q}_{a,1}$ and $\tau_{a,0} = \hat{q}_{a,0}$. Label 1 enters $C(x,a)$ exactly when $\hat{p} \geq \tau_{a,1}$, and label 0 exactly when $\hat{p} \leq \tau_{a,0}$. Each group is therefore in one of two regimes. If $\tau_{a,1} \leq \tau_{a,0}$, every set is nonempty and the ambiguous sets $\{0,1\}$ occupy the band $[\tau_{a,1}, \tau_{a,0}]$. If $\tau_{a,1} > \tau_{a,0}$, which happens when the levels are loose or the score separates the outcomes well, no set is ambiguous and the sets are empty on the gap $(\tau_{a,0}, \tau_{a,1})$. Raising the levels moves the two thresholds toward each other and then past each other, so the band shrinks while the gap grows. Since ECC refers both, its workload first falls and then rises again (Fig. 2), and a budget below the minimum of this U-shaped curve cannot be met at any level.

D. Problem Statement

Given a budget B, we seek a decision rule that refers exactly K students, guarantees $m_a \leq \hat{\alpha}_1$ and $f_a \leq \hat{\alpha}_0$ in every group for certified levels that are as small as possible, and uses no outcome of the screened cohort.

IV. Resource-Aware Fair Conformal Prediction

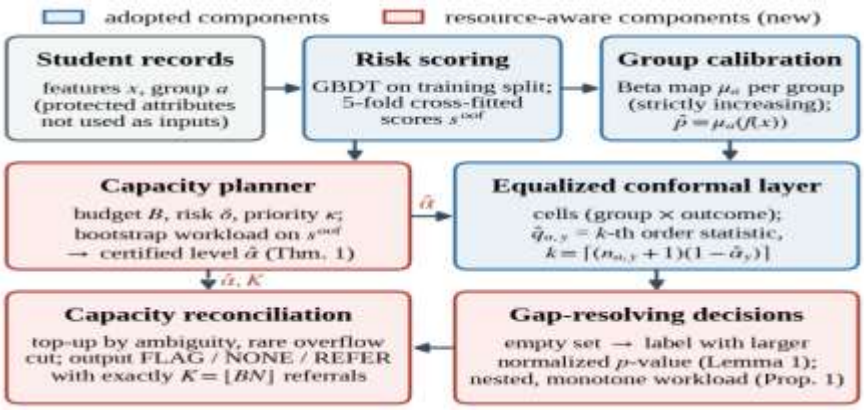


Figure 1shows the pipeline. Scoring, calibration, and the equalized cells are those of Section III-B; the three resource-aware components are described below and summarized in Algorithm 1.

Fig. 1. Overview of RAFCP. Blue boxes are adopted components of the equalized conformal early-warning framework; red boxes are the resource-aware components introduced here. The planner fixes the certified level before the calibration split is used, and reconciliation adjusts the referral list of the actual cohort to exactly $K = \lfloor BN \rfloor$ students.

A. Gap-Resolving Decisions

For a student (x,a), define the conformal p-value of label y within its cell as

$$u_y(x,a) = (1 + |\{i \in I_{a,y} : V_i \geq V(x,a,y)\}|) / (n_{a,y} + 1). \quad (4)$$

Excluding y from $C(x,a)$ corresponds, up to ties, to $u_y(x,a) \leq \alpha_y$. RAFCP keeps the actions of ECC for singleton and ambiguous sets and resolves an empty set automatically: it flags the student if $u_1/\alpha_1 \geq u_0/\alpha_0$ and takes no action otherwise. The ratio measures how far each label falls short of its own admission level, so the rule treats the two labels symmetrically when $\alpha_1 = \alpha_0$ and respects their priority otherwise. Because u_y depends on scores only through ranks within cells, the resolved decisions keep the invariance of the layer to strictly increasing recalibration.

Lemma 1 (gap resolution preserves the bounds): Under the conditions of the ECC guarantee, every rule that assigns FLAG or NONE to empty prediction sets, and in particular the gap-resolving rule, satisfies $m_a \leq \alpha_1$ *and* $f_a \leq \alpha_0$ *for every group* a.

Proof: A student receives NONE only if $1 \notin C$ and FLAG only if $0 \notin C$, and an empty set satisfies both conditions. Hence $\{d = \text{NONE}, Y = 1\} \subseteq \{Y \notin C\}$ and $\{d = \text{FLAG}, Y = 0\} \subseteq \{Y \notin C\}$, and the claim follows from cell-wise coverage. *Proposition 1 (nested referrals): Fix the calibration data. Under gap-resolving decisions a student of group* a *is referred if and only if* $\tau_{a,1}(\alpha_1) \leq \hat{p} \leq \tau_{a,0}(\alpha_0)$. *If* $\alpha_1 \leq \alpha'_1$ *and* $\alpha_0 \leq \alpha'_0$, *every student referred at* (α'_1, α'_0) *is also referred at* (α_1, α_0); *consequently the referral rate*

$$R(\alpha) = \sum_a w_a \, P(\tau_{a,1}(\alpha_1) \leq \hat{p} \leq \tau_{a,0}(\alpha_0) \mid A = a), \quad (5)$$

with group shares w_a, *is non-increasing in each level.*

Proof: A student is referred exactly when both labels belong to the set. Since k in (3) is non-increasing in α_y, so is $\hat{q}_{a,y}$; hence $\tau_{a,1}$ is non-decreasing in α_1, $\tau_{a,0}$ is non-increasing in α_0, and the band can only shrink. Under the original rule the gap probability $P(\tau_{a,0} < \hat{p} < \tau_{a,1} \mid A = a)$ is added to (5). It is zero while the band exists and grows with the levels afterwards, which produces the U-shape described in Section III-C. Proposition 2 (equal levels are minimax-optimal): Let group- and label-specific levels $\alpha_{a,y}$ be chosen subject to $W(\{\alpha_{a,y}\}) \leq B$, where W is any workload functional that is non-increasing in every level, such as (5) or a vector of per-group caps. Among

all feasible assignments, the worst level $\max_{a,y} \alpha_{a,y}$ is minimized by an assignment in which all levels are equal. The same holds for the weighted worst level $\max_a \max(\alpha_{a,1}, \alpha_{a,0}/\kappa)$ with $\kappa > 0$.
Proof: Let α^* be the optimal worst level, attained by a feasible assignment. Raising every level to α^* leaves the worst level unchanged and, by monotonicity, does not increase W, so the equal assignment is feasible and optimal. The weighted case follows from the same argument along the ray $\alpha_{a,0} = \kappa\alpha_{a,1}$. Proposition 2 has a direct policy reading. Giving every group the same share of counselor time, a rule that looks neutral, amounts to a particular unequal assignment of levels and cannot certify a better worst-case guarantee than equal levels. In Section VI-C it is significantly worse on the two cohorts whose groups differ in base rates.

B. Capacity Planner

The levels follow a priority path $\alpha_1 = \alpha$ and $\alpha_0 = \min(\kappa\alpha, 0.95)$, where $\kappa \geq 1$ states how many times more false flags than misses the institution is willing to tolerate; $\kappa = 1$ gives symmetric guarantees. The planner fixes α before the calibration split is used. Its inputs are the cross-fitted calibrated scores of the training split, which play the role of scores of the deployed model on new students, the sizes $n_{a,y}$ of the calibration cells, and the group sizes N_a of the cohort to be screened, which are known before any outcome. In replicate $j = 1, \ldots, L$ it draws $n_{a,y}$ scores with replacement from each training cell, computes the thresholds (3) for every level on a grid G, draws N_a scores from each group, and evaluates the referral fraction $R^{(j)}(\alpha)$ by (5) with sorting and binary search. The certified level is

$$\hat{\alpha} = \min \{\alpha \in G : Q_{1-\delta}(R^{(1)}(\alpha), \ldots, R^{(L)}(\alpha)) \leq B\}, \quad (6)$$

where $Q_{1-\delta}$ denotes the empirical $(1 - \delta)$-quantile. By Proposition 1 every replicate is non-increasing in α, and so is their pointwise quantile; (6) is therefore the first crossing of a monotone curve. The bootstrap [23] reproduces the two sources of workload variation at deployment, the sampling of calibration cells and the composition of the cohort. With L = 200 replicates and 97 grid levels, planning adds less than one second per split on a single CPU core.
Theorem 1 (certified guarantee): Suppose that calibration and test examples are exchangeable within every cell and that $\hat{\alpha}$ in (6) depends on the training split only. Then, conditionally on the training split, gap-resolving decisions at the levels $(\hat{\alpha}, \min(\kappa\hat{\alpha}, 0.95))$ satisfy $m_a \leq \hat{\alpha}$ and $f_a \leq \min(\kappa\hat{\alpha}, 0.95)$ for every group a. Proof: Conditionally on the training split the levels are fixed and independent of the calibration data and of the test student, so the cell-wise coverage of ECC holds at these levels, and Lemma 1 transfers it to the decisions. Selecting the level on the calibration split itself, or on the cohort with the calibration thresholds, would make the level depend on the calibration scores and void this argument; Section VI-E evaluates such an adaptive variant.

C. Capacity Reconciliation

The planned rule refers $\hat{R}$ students of the actual cohort, a number that differs from K through sampling variation. If $\hat{R} < K$, RAFCP additionally refers the $K - \hat{R}$ automated students whose excluded label is closest to admission, ranking FLAG decisions by $u_0/\hat{\alpha}_0$ and NONE decisions by $u_1/\hat{\alpha}_1$. If $\hat{R} > K$, an overflow event O, it resolves the $\hat{R} - K$ referred students with the smallest $\min(u_1/\hat{\alpha}_1, u_0/\hat{\alpha}_0)$ by the gap rule. Because the p-value of the true label is uniform within each cell, both rankings treat groups symmetrically and act like a common shift of the effective level in all cells.

Corollary 1*:* Under the conditions of Theorem 1, the reconciled rule refers exactly K students and satisfies $m_a \leq \hat{\alpha}_1 + P(O \mid Y = 1, A = a)$ *and* $f_a \leq \hat{\alpha}_0 + P(O \mid Y = 0, A = a)$ *for every group a.*
Proof: Without overflow, reconciliation only turns automated actions into referrals and cannot create a miss or a false flag. With overflow a student can additionally be resolved wrongly, which adds at most the indicator of O to the error indicator of the planned rule. The planner targets

P(O) ≈ δ. An institution that can absorb an occasional overflow may skip the trimming step, in which case the bounds of Theorem 1 hold exactly. Algorithm 1 summarizes the procedure.

Algorithm 1: Resource-aware fair conformal prediction

Input: training and calibration splits, cohort features and groups, budget B, risk δ, priority κ, grid G.
1: Train f on the training split, compute cross-fitted scores, and fit μ_a for every group.
2: For $j = 1, \ldots, L$, bootstrap the workload $R^{(j)}(\alpha)$, $\alpha \in G$, with the deployment sizes $n_{a,y}$ and N_a.
3: Set $\hat{\alpha}$ by (6), $\hat{\alpha}_1 = \hat{\alpha}$, and $\hat{\alpha}_0 = \min(\kappa\hat{\alpha}, 0.95)$.
4: On the calibration split, compute $\hat{q}_{a,y}$ at $(\hat{\alpha}_1, \hat{\alpha}_0)$ by (3).
5: For each student form $C(x,a)$ and act FLAG, NONE, or REFER; resolve empty sets by $u_1/\hat{\alpha}_1 \geq u_0/\hat{\alpha}_0$.
6: Reconcile the referral list to $K = \lfloor BN \rfloor$ students by normalized ambiguity.
Output: actions, certified levels $(\hat{\alpha}_1, \hat{\alpha}_0)$, and per-group referral rates.

V. Experimental Setup

A. Datasets

Following the protocol of TRUE-EWS [21], we use three public U.S. cohorts that span the education pipeline, with the same task definitions, protected groups, and analytic samples (Table I). Tennessee STAR randomly assigned students to class types from kindergarten through grade 3 [24]. The warning point is the end of grade 1; the inputs are kindergarten and grade-1 test scores, birth quarter, class types, school urbanicity, and grade-1 teacher characteristics, and a student is at risk if the sum of standardized grade-3 reading and mathematics scores lies in the bottom quartile. High School and Beyond (HS&B) is the 1980 senior cohort with a 1986 follow-up assembled by Rouse [25]; the inputs are a base-year achievement composite, parental college education, home ownership, urbanicity, local labor-market and tuition variables, distance to college, and region, and a student is at risk if no schooling beyond high school is reported. Both extracts come from the AER package [26], obtained through a public CSV mirror. The LSAC National Longitudinal Bar Passage Study followed 1991 law-school entrants through the bar examination [27]; we use the cleaned release of 20,798 students distributed with [28], the first-year inputs (first-year GPA, class-rank decile, LSAT score, undergraduate GPA, family income, full-time status, and school tier), and failure of the bar examination as the outcome. The primary groups are Black versus other students, African American or Hispanic (URM) versus other students, and Non-White versus White students; the secondary attributes used for intersectional cells are free-lunch status, family income below USD 25,000, and sex. Sample sizes, group shares, and group-wise prevalences coincide with the published task definitions.

Table 1DATASETS AND EARLY-WARNING TASKS

	STAR	HS&B	LSAC
Stage	Grades K–3	College access	Licensure
Students *n*	3,762	4,739	20,798
Inputs (encoded)	31	10	7
At-risk outcome	Low grade-3 score	No college	Bar not passed
***P*(*Y* = 1)**	0.250	0.387	0.110
Group *A* = 1	Black	URM	Non-White
Share of *A* = 1	0.285	0.356	0.159
***P*(*Y* = 1 | *A* = 0/1)**	0.184/0.415	0.370/0.416	0.079/0.277
Secondary attribute	Free lunch	Low income	Female
Smallest calib. cell	89	140	183
Test AUC (mean ± s.d.)	0.862 ± 0.012	0.741 ± 0.014	0.841 ± 0.007

Calibration cell sizes refer to the 20% calibration split; AUC of the GBDT on the test split over 30 splits.

B. Protocol and Implementation

Each experiment uses 30 random splits (seeds 0–29) stratified by (A, Y) into 60% training, 20% calibration, and 20% test data; the test split plays the role of the cohort to be screened. The GBDT is LightGBM [29] with 300 trees, learning rate 0.03, 15 leaves, at least 30 samples per leaf, row and column subsampling of 0.8, and an L2 penalty of 1, as in the original framework. Budgets are 5%, 10%, 15%, 20%, 25%, and 30% of the cohort, the overflow risk is $\delta = 0.1$, the planner uses $L = 200$ replicates on the grid $G = \{0.020, 0.025, \ldots, 0.500\}$, and $\kappa = 1$ unless stated otherwise. The implementation uses scikit-learn, LightGBM, and fairlearn [30], and the full pipeline runs in about three minutes on a single CPU core.

C. Baselines

All referral-based methods use the same group-wise calibrated scores and refer exactly K students; the point rule flags a student when $\hat{p} \geq \pi$, the training prevalence. *GBDT (point)* and *ThreshOpt-EO* refer no one; ThreshOpt-EO post-processes the GBDT with randomized group-specific thresholds that maximize balanced accuracy under equalized odds [8] and is fitted on the calibration split. *Capacity-cut ECC* runs the equalized layer at $\alpha = 0.1$; counselors review the K referred students closest to the boundary in $|\text{logit}\,\hat{p} - \text{logit}\,\pi|$, and the other referred students receive the point decision. *Confidence deferral* refers the K students with the smallest $|\text{logit}\,\hat{p} - \text{logit}\,\pi|$ and applies the point rule to the rest, the selective-classification setting studied in [14]. *Group-quota deferral* does the same within each group with quota $\lfloor BN_a \rfloor$, which enforces workload parity. *Label-conditional CP with reconciliation* indexes cells by outcome only [11], [12] and uses the planner, gap rule, and reconciliation of RAFCP, which isolates the role of group conditioning. *Workload-proportional ECC* plans a separate level for each group so that it refers $\lfloor BN_a \rfloor$ of its students, which tests Proposition 2. Knowledge-enhanced explainable fake-review detection illustrates how heterogeneous information and feature-level evidence can be organized to support transparent classification decisions [32].

D. Metrics

For every group we compute the miss and false-flag rates of the decisions on the test split and average them over splits; the worst-group rate is the largest group average, as in the original framework. The worst-group automated error (WAE) is the larger of the worst-group miss and false-flag rates. We also report between-group gaps, the realized referral rate, the certified level $\hat{\alpha}$, the overflow frequency, and paired Wilcoxon signed-rank tests on the per-split worst-group error.

VI. Results

A. Reproducing the Equalized Layer

Table II reports the conformal layers at $\alpha_1 = \alpha_0 = 0.1$. Our reimplementation reproduces the pattern reported for TRUE-EWS [21]: ECC keeps all twelve group-by-outcome cells between 0.897 and 0.909, with referral rates of 0.317, 0.556, and 0.386 against published values of 0.316, 0.554, and 0.385, whereas marginal conformal prediction covers 3.1% of at-risk White law students and label-conditional prediction covers 62.4% of on-track Non-White law students. Group-wise Beta calibration lowers the worst-group expected calibration error of the raw GBDT from 0.076, 0.076, and 0.044 to 0.062, 0.062, and 0.035, and the test AUCs in Table I agree with the original study within 0.001. The same experiment exposes the problem addressed here: the guarantee at $\alpha = 0.1$ requires a third to more than half of each cohort to be reviewed.

Table 2CONFORMAL LAYERS AT A1 = A0 = 0.1

	STAR		**HS&B**		**LSAC**	
Method	Min cov.	Ref.	Min cov.	Ref.	Min cov.	Ref.
Marginal CP	0.609	0.201	0.765	0.518	0.031	0.008
Group-cond. CP	0.579	0.216	0.780	0.519	0.001	0.081
Label-cond. CP	0.774	0.295	0.857	0.547	0.624	0.334
ECC (reproduced)	**0.900**	**0.317**	**0.897**	**0.556**	**0.899**	**0.386**

Min cov.: smallest of the four group-by-outcome coverages $P(y \in C \mid A = a, Y = y)$; Ref.: referral rate. Means over 30 splits.

B. Workload Structure

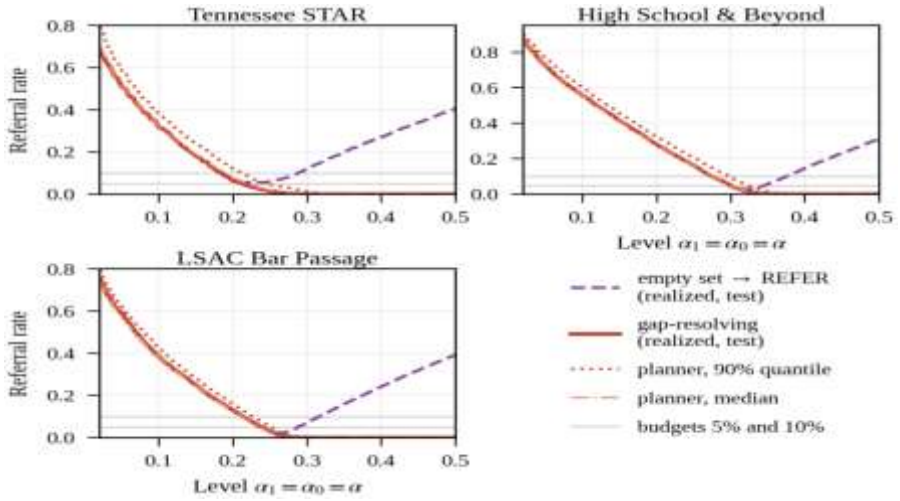


Figure 2. Referral rate on the test split as a function of a common level (mean over 30 splits). Under the original rule (dashed) empty sets are referred and the workload is U-shaped; with gap resolution (solid) it decreases monotonically to zero. Dotted and das

Fig. 2 plots the referral rate on the test split against a common level. Under the original rule the curve is U-shaped: its minimum is 0.053 on STAR (at $\alpha \approx 0.23$), 0.027 on HS&B ($\alpha \approx 0.33$), and 0.021 on LSAC ($\alpha \approx 0.27$), and it climbs back to 0.41, 0.31, and 0.39 at $\alpha = 0.5$ because the empty sets of the gap are referred. With gap resolution the curve is monotone and reaches zero at $\alpha = 0.30$, 0.36, and 0.30, so every budget is feasible. The planner median deviates from the realized curve by 0.002–0.003 on average, and the 90% quantile lies above it by the margin that protects against overflow. Planning with the original rule finds no level that meets a 5% budget in any split on STAR and HS&B and fails in one split on LSAC; at 10% it still fails in 3 of 30 splits on STAR.

C. Fairness under Capacity

Table 3WORST-GROUP MISS AND FALSE-FLAG RATES AT EQUAL WORKLOAD

	STAR		**HS&B**		**LSAC**	
Method	Miss	FF	Miss	FF	Miss	FF
Referral budget B = 10%						
GBDT (point)†	0.251	0.510	0.350	0.428	0.318	0.632
ThreshOpt-EO†	0.263	0.290	0.336	0.323	0.270	0.304
Capacity-cut ECC	0.172	0.412	0.304	0.375	0.244	0.525
Conf. deferral	0.183	0.433	0.304	0.375	0.253	0.578
Quota deferral	0.179	0.450	0.306	0.374	0.254	0.569
Label-CP + rec.	0.246	0.360	0.314	0.361	0.284	0.545

Workload-prop.	0.228	0.221	0.269	0.279	0.226	0.240
RAFCP (ours)	**0.191**	**0.187**	**0.272**	**0.277**	**0.209**	**0.218**
Referral budget B = 20%						
Capacity-cut ECC	0.119	0.309	0.249	0.325	0.187	0.489
Conf. deferral	0.133	0.354	0.250	0.325	0.200	0.513
Quota deferral	0.124	0.377	0.252	0.320	0.202	0.494
Label-CP + rec.	0.184	0.283	0.262	0.316	0.222	0.473
Workload-prop.	0.182	0.179	0.228	0.233	0.183	0.197
RAFCP (ours)	**0.141**	**0.144**	**0.229**	**0.232**	**0.170**	**0.173**

† No referral; identical for both budgets. FF: false flag. Group rates are averaged over 30 splits before the worst group is taken; all other methods refer exactly $K = \lfloor BN \rfloor$ students.

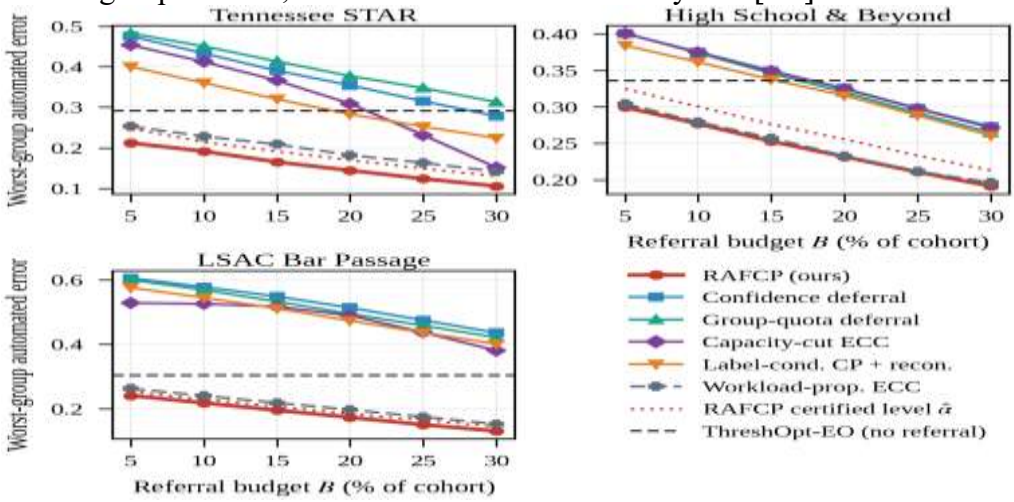


Figure 3Worst-group automated error (the larger of the worst-group miss and false-flag rates) against the referral budget. All referral-based methods refer exactly K = ⌊BN⌋ students; the dotted red line is the level certified by RAFCP.

Table III and Fig. 3 compare the methods at equal workload. At a 10% budget RAFCP holds the worst-group miss and false-flag rates at 0.191 and 0.187 on STAR, 0.272 and 0.277 on HS&B, and 0.209 and 0.218 on LSAC, with between-group gaps of at most 0.017. Confidence deferral and group-quota deferral leave worst-group false-flag rates of 0.433–0.450 on STAR and 0.569–0.578 on LSAC, and ThreshOpt-EO, which refers no one, reaches a WAE of 0.290, 0.336, and 0.304. The ordering holds at every budget (Fig. 3): RAFCP has the lowest WAE in all 18 dataset–budget combinations, and its certified level lies 0.01–0.04 above its realized WAE.

The failures of the baselines have identifiable causes. The point rule behind the deferral baselines flags every student whose risk exceeds the overall prevalence, which on LSAC includes 63.2% of on-track Non-White students before any deferral, and deferral removes only the students near the boundary, most of whom belong to the other group. Capacity-cut ECC fails in the same way:

on LSAC it reviews 11.8% of White but only 0.4% of Non-White students, whose risks lie far from the overall prevalence, and leaves 52.5% of on-track Non-White students flagged. RAFCP instead spends capacity where its group-wise thresholds leave ambiguity, which is mostly in the higher-prevalence group. At a 10% budget it refers 18.3% of Black and 6.6% of other students on STAR, 15.1% of Non-White and 9.0% of White students on LSAC, and 10.7% versus 9.5% on HS&B, where base rates are close. Per split, RAFCP attains a lower worst-group error than confidence deferral, group-quota deferral, capacity-cut ECC, and label-conditional CP in at least 27 of 30 splits on every dataset at the 10% and 20% budgets (Wilcoxon $p < 10^{-5}$). Workload-proportional ECC, which also works inside the conformal framework, is the strongest baseline. It trails RAFCP in WAE by 0.037 on STAR and 0.022 on LSAC at 10% (22 of 30 splits, $p \leq 2 \times 10^{-4}$), as Proposition 2 predicts, but on HS&B, where the two groups have base rates of 0.370 and 0.416, the two methods are statistically indistinguishable ($p = 0.66$ and 0.10 at 10% and 20%).

D. Priority Weights

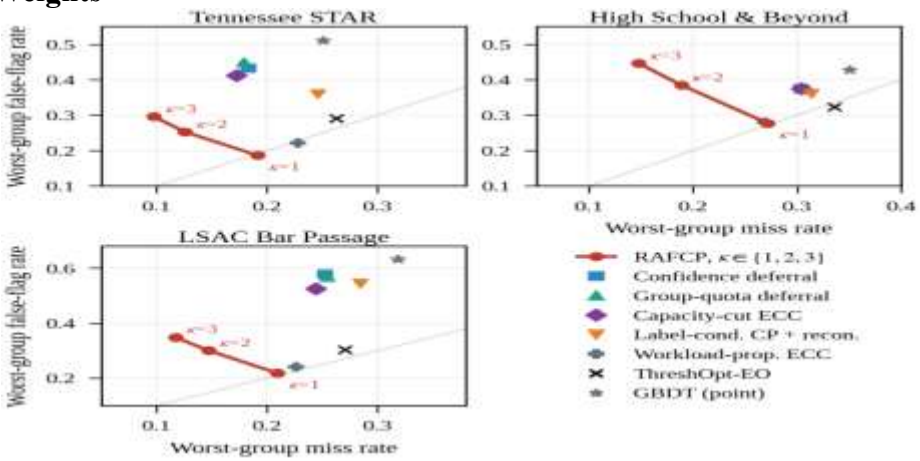


Figure 4. Worst-group miss rate against worst-group false-flag rate at a 10% referral budget. Raising the priority κ moves RAFCP toward stronger protection of at-risk students; the gray diagonal marks equal rates.

Fig. 4 shows the trade-off between worst-group misses and false flags at a 10% budget. Raising κ moves RAFCP along a curve that protects at-risk students more strongly. With $\kappa = 3$ the worst-group miss rate falls to 0.097 on STAR and 0.117 on LSAC, while the worst-group false-flag rate rises to 0.296 and 0.349; both coordinates remain below those of confidence deferral (0.183 and 0.433 on STAR, 0.253 and 0.578 on LSAC). On these two cohorts, RAFCP with $\kappa \in \{2, 3\}$ improves on confidence deferral, group-quota deferral, and capacity-cut ECC in both coordinates at once. On HS&B the symmetric configuration already does so (0.272 and 0.277 against 0.304–0.306 and 0.374–0.375), and larger values of κ trade errors almost one for one because the score is weak (AUC 0.741). For every κ and budget the between-group gaps stay below 0.025.

E. Certificates, Overflow, and Ablation

Table 4CERTIFICATES AND ABLATION AT A 5% REFERRAL BUDGET

Variant	Feas.	α̂	Overfl.	Ref.	WAE
Tennessee STAR					
RAFCP	**30/30**	**0.248**	**3%**	**0.049**	**0.212**

w/o gap resolution	0/30	0.229	53%	0.049	0.228
w/o safety margin	30/30	0.221	30%	0.049	0.212
w/o reconciliation	30/30	0.248	–	0.017	0.249
Adaptive level	–	0.214	–	0.049	0.212
High School and Beyond					
RAFCP	**30/30**	**0.325**	**10%**	**0.050**	**0.300**
w/o gap resolution	0/30	0.325	23%	0.050	0.300
w/o safety margin	30/30	0.304	37%	0.050	0.300
w/o reconciliation	30/30	0.325	–	0.016	0.319
Adaptive level	–	0.302	–	0.050	0.300
LSAC Bar Passage					
RAFCP	**30/30**	**0.253**	**7%**	**0.050**	**0.241**
w/o gap resolution	29/30	0.254	7%	0.050	0.243
w/o safety margin	30/30	0.238	53%	0.050	0.241
w/o reconciliation	30/30	0.253	–	0.025	0.258
Adaptive level	–	0.240	–	0.050	0.241

Feas.: splits in which the planner finds a level that meets the budget; $\hat{\alpha}$: mean certified level; Overfl.: share of splits in which the planned rule exceeds the budget; Ref.: realized referral rate; WAE: worst-group automated error. Table IV dissects RAFCP at the tightest budget, 5% of the cohort. The certified levels are 0.248, 0.325, and 0.253, and the realized WAE lies below them. Removing the safety margin ($\delta = 0.5$) lowers the certified level by 0.015–0.027 but lets the plan overflow in 30%–53% of the splits, so the certificate no longer holds reliably, whereas RAFCP overflows in 3%–10%. Without gap resolution no level is feasible on STAR and HS&B; the planner falls back to the level with the smallest workload, the plan overflows in 53% and 23% of the splits, and on STAR the WAE rises from 0.212 to 0.228. Without reconciliation the planned rule uses only 1.7%, 1.6%, and 2.5% of the cohort instead of 5%, and its WAE rises by 0.017–0.037. The adaptive variant, which selects the level on the cohort with the calibration thresholds, ends with the same decisions after reconciliation, but its level depends on the calibration scores. Across all budgets, RAFCP overflows in 7.2%, 11.7%, and 7.2% of the splits on average (3%–17% per budget), close to the target $\delta = 0.1$. For the planned rule, the per-group mean of $m_a - \hat{\alpha}$ and $f_a - \hat{\alpha}$ lies between −0.018 and +0.004, within sampling error of the bound, and after reconciliation between −0.044 and −0.013.

F. Intersectional Groups

Table 5INTERSECTIONAL GROUPS: WORST-GROUP MISS AND FALSE-FLAG RATES

	STAR		HS&B		LSAC	
Method	Miss	FF	Miss	FF	Miss	FF
Referral budget B = 10%						
Conf. deferral	0.301	0.540	0.568	0.424	0.264	0.616
Quota deferral	0.314	0.546	0.528	0.424	0.265	0.608
Label-CP + rec.	0.346	0.469	0.597	0.392	0.297	0.586
RAFCP (ours)	**0.205**	**0.211**	**0.289**	**0.293**	**0.212**	**0.219**
Referral budget B = 20%						
Conf. deferral	0.254	0.455	0.517	0.368	0.204	0.554
Quota deferral	0.262	0.462	0.427	0.372	0.207	0.531
Label-CP + rec.	0.290	0.392	0.543	0.342	0.231	0.516
RAFCP (ours)	**0.156**	**0.165**	**0.241**	**0.258**	**0.172**	**0.178**

Four groups per dataset: race × free lunch (STAR), URM × income (HS&B), and race × sex (LSAC). Crossing the primary group with the secondary attribute gives four groups and eight calibration cells per dataset. At $\alpha = 0.1$ the intersectional ECC keeps every cell between 0.891 and 0.934 with referral rates of 0.371, 0.583, and 0.393, again close to the values reported for the original framework. Under capacity (Table V) the advantage of RAFCP widens: at a 10% budget its WAE is 0.211, 0.293, and 0.219, against 0.540, 0.568, and 0.616 for confidence deferral and 0.469, 0.597, and 0.586 for label-conditional CP. With smaller cells the certified levels are higher (0.235, 0.315, and 0.232), which is the statistical price of guaranteeing four groups instead of two.

G. The Cost of Equity

Equal error rates are not free in pooled terms. At a 10% budget the balanced pooled error, the mean of the cohort-level miss and false-flag rates, is 0.182, 0.273, and 0.210 for RAFCP and 0.170, 0.275, and 0.190 for confidence deferral; across all budgets the difference ranges from −0.003 to +0.023. The increase of 0.012 on STAR and 0.020 on LSAC buys a reduction of the worst-group error by 0.24 and 0.36. This is the familiar tension between group-level and pooled performance [6], [7], here expressed in counselor time: RAFCP spends the budget where the model is least certain for the group it serves worst, rather than where it would remove the most errors overall.

VII. Discussion

Capacity as an input. RAFCP turns the staffing of a counseling office into a certified equalized level that can be stated in advance and checked afterwards. With a 10% budget, a district using

the STAR model can certify that neither group's miss or false-flag rate exceeds about 0.21, and the realized worst-group rates were 0.19. Because the planner needs only training data and the group sizes of the incoming cohort, the certificate can be issued before the term begins and recomputed when staffing changes. Equal workload is not equal protection. Proposition 2 and the experiments show that dividing counselor time in proportion to group size, a rule that is easy to justify administratively, certifies a weaker worst-case guarantee whenever groups differ in base rates or score separation. The equalized solution gives more review time to the group for which the model is less certain, and the allocation itself is an auditable output. Limitations. The guarantees rely on exchangeability between the calibration data and the screened cohort within every cell; a cohort that differs from earlier years, for example after a policy change, requires recalibration on recent data. The decision layer uses group membership to set thresholds, which some jurisdictions may restrict even though the model itself excludes protected attributes. Small cells make thresholds coarse and certificates conservative, as the intersectional results show. We assume that counselors resolve referred cases correctly and we model a single resource; limited intervention slots for flagged students and imperfect human review would add constraints. Finally, the three cohorts are historical research datasets, and their outcomes are proxies for the needs that a deployed EWS would serve.

VIII. Conclusion

Fair early-warning guarantees are only as useful as the capacity available to honor them. RAFCP makes this capacity explicit: it resolves empty prediction sets without loss of validity, certifies an equalized level from training data alone, and delivers it at exactly the budgeted workload. On three U.S. cohorts from early grades to professional licensure, it reduced the worst-group automated error under a 10% referral budget by 0.10 to 0.36 relative to confidence-based deferral and kept the gap between groups below 0.02. Future work will address shifts between cohorts, joint planning of review and intervention capacity, and models of counselor accuracy.

References

F. Zhou, "Moment-constrained side-force allocation for wind-assisted merchant ships: Hydrodynamic performance of retractable anti-leeway fins and flap rudders," Innovation and Technology Studies, vol. 3, no. 1, pp. 27–41, 2026.

H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison, "A machine learning framework to identify students at risk of adverse academic outcomes," in *Proc. 21st ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2015, pp. 1909–1918, doi: 10.1145/2783258.2788620.

R. Yu, H. Lee, and R. F. Kizilcec, "Should college dropout prediction models include protected attributes?" in *Proc. 8th ACM Conf. Learning @ Scale*, 2021, pp. 91–100, doi: 10.1145/3430895.3460139.

R. S. Baker and A. Hawn, "Algorithmic bias in education," *Int. J. Artificial Intelligence in Education*, vol. 32, no. 4, pp. 1052–1092, 2022, doi: 10.1007/s40593-021-00285-9.

J. Gardner, C. Brooks, and R. Baker, "Evaluating the fairness of predictive student models through slicing analysis," in *Proc. 9th Int. Conf. Learning Analytics & Knowledge*, 2019, pp. 225–234, doi: 10.1145/3303772.3303791.

J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Proc. 8th Innovations in Theoretical Computer Science Conf. (ITCS)*, LIPIcs, vol. 67, 2017, pp. 43:1–43:23, doi: 10.4230/LIPIcs.ITCS.2017.43.

G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5680–5689.

M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, vol. 29, 2016, pp. 3315–3323.

V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. New York, NY, USA: Springer, 2005, doi: 10.1007/b106715.

X. Yue and R. Zhu, "An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study," World Journal of Management Science, vol. 2, no. 1, pp. 63–72, 2024.

V. Vovk, "Conditional validity of inductive conformal predictors," in *Proc. Asian Conf. Machine Learning*, PMLR, vol. 25, 2012, pp. 475–490.

M. Sadinle, J. Lei, and L. Wasserman, "Least ambiguous set-valued classifiers with bounded error levels," *J. American Statistical Association*, vol. 114, no. 525, pp. 223–234, 2019, doi: 10.1080/01621459.2017.1395341.

Y. Romano, R. F. Barber, C. Sabatti, and E. J. Candès, "With malice toward none: Assessing uncertainty via equalized coverage," *Harvard Data Science Review*, vol. 2, no. 2, 2020, doi: 10.1162/99608f92.03f00592.

E. Jones, S. Sagawa, P. W. Koh, A. Kumar, and P. Liang, "Selective classification can magnify disparities across groups," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

D. Teng, "TRIAD: Deadline- and reliability-aware scheduling for post-earthquake structural assessment workflows," World Journal of Engineering Research, vol. 3, no. 6, pp. 55–66, 2025.

A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024.

Y. Jin and E. J. Candès, "Selection by prediction with conformal p-values," *J. Machine Learning Research*, vol. 24, no. 244, pp. 1–41, 2023.

D. Madras, T. Pitassi, and R. Zemel, "Predict responsibly: Improving fairness and accuracy by learning to defer," in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 6147–6157.

H. Mozannar and D. Sontag, "Consistent estimators for learning to defer to an expert," in *Proc. 37th Int. Conf. Machine Learning*, PMLR, vol. 119, 2020, pp. 7076–7087.

E. Straitouri, L. Wang, N. Okati, and M. Gomez Rodriguez, "Improving expert predictions with conformal prediction," in *Proc. 40th Int. Conf. Machine Learning*, PMLR, vol. 202, 2023, pp. 32633–32653.

Q. Fang, "TRUE-EWS: Trustworthy early warning with equalized conformal guarantees for equitable student support in U.S. education," *Educational Research and Human Development*, vol. 1, no. 2, pp. 44–54, 2024, doi: 10.61784/erhd4059.

M. Kull, T. Silva Filho, and P. Flach, "Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers," in *Proc. 20th Int. Conf. Artificial Intelligence and Statistics*, PMLR, vol. 54, 2017, pp. 623–631.

J. Chen, X. Zhang, and X. Han, "Explainable identification of tax-related financial reporting anomalies in U.S. public companies: Cross-fitted tax-consistency residuals and Owen-value Shapley propagation," World Journal of Economics and Business Research, vol. 1, no. 1, pp. 33–45, 2023.

A. B. Krueger, "Experimental estimates of education production functions," *Quarterly J. Economics*, vol. 114, no. 2, pp. 497–532, 1999, doi: 10.1162/003355399556052.

C. E. Rouse, "Democratization or diversion? The effect of community colleges on educational attainment," *J. Business & Economic Statistics*, vol. 13, no. 2, pp. 217–224, 1995.

C. Kleiber and A. Zeileis, *Applied Econometrics with R*. New York, NY, USA: Springer, 2008, doi: 10.1007/978-0-387-77318-6.

J. Wang, Y. Ge, J. Liu, and B. Jiang, "Early warning of liquidity stress in U.S. banks with seasonally aligned frailty hazards and delayed-feedback conformal alarms," Journal of Trends in Finance and Economics, vol. 1, no. 1, pp. 85–100, 2024.

T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, "A survey on datasets for fairness-aware machine learning," *WIREs Data Mining and Knowledge Discovery*, vol. 12, no. 3, Art. no. e1452, 2022, doi: 10.1002/widm.1452.

G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 3146–3154.

H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, and M. Madaio, "Fairlearn: Assessing and improving fairness of AI systems," *J. Machine Learning Research*, vol. 24, no. 257, pp. 1–8, 2023.

J. Zou, E. Feng, P. Li, and G. Wang, "Horizon-aware group-conditional conformal prediction: Post-hoc shrinkage for sparse multi-horizon coverage auditing," Machine Learning with Applications, vol. 26, art. 101038, 2026, doi: 10.1016/j.mlwa.2026.101038.

J. Chen et al., "KE-MLLM: A Knowledge-Enhanced Multi-Sensor Learning Framework for Explainable Fake Review Detection," Applied Sciences, vol. 16, art. 2909, 2026, doi: 10.3390/app16062909.