Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1440

A Sandbox for Hard Choices: Using Simulation to Explore School Closure Scenarios and Their Consequences

Michael L. Chrzan
University of Maryland

Francis A. Pearman
Stanford University

School closures are often justified through seemingly neutral criteria such as enrollment or performance, but these metrics can unintentionally deepen educational disparities. This study uses a large urban district's administrative data to simulate 5,040 closure scenarios, systematically varying seven policy design principles, including proximity, enrollment, seat utilization, building quality, academic performance, disproportionality safeguards, and ordering of schools considered. By comparing the equity, fiscal, and operational outcomes of each scenario, we reveal three key findings: (1) safeguards explicitly designed to prevent disproportionality improve fairness but reduce cost savings and seat reductions needed to balance capacity and demand; (2) common criteria like enrollment do little to advance either efficiency or equity; and (3) how schools are ranked for evaluation is a surprisingly powerful policy lever. This work contributes to the field by showing how simulation can equip district leaders to anticipate the trade-offs embedded in closure decisions, moving policy design toward proactive fairness.

VERSION: April 2026

Suggested citation: Chrzan, Michael L., and Francis A. Pearman. (2026). A Sandbox for Hard Choices: Using Simulation to Explore School Closure Scenarios and Their Consequences. (EdWorkingPaper: 26-1440). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/xt1x-1787

A Sandbox for Hard Choices: Using Simulation to Explore School Closure Scenarios and Their Consequences

Michael L. Chrzan[1] Francis A. Pearman[2]

[1]Center for Educational Data Science and Innovation, University of Maryland

[2]Stanford University

Corresponding author: mlchrzan@umd.edu

Abstract

School closures are often justified through seemingly neutral criteria such as enrollment or performance, but these metrics can unintentionally deepen educational disparities. This study uses a large urban district's administrative data to simulate 5,040 closure scenarios, systematically varying seven policy design principles, including proximity, enrollment, seat utilization, building quality, academic performance, disproportionality safeguards, and ordering of schools considered. By comparing the equity, fiscal, and operational outcomes of each scenario, we reveal three key findings: (1) safeguards explicitly designed to prevent disproportionality improve fairness but reduce cost savings and seat reductions needed to balance capacity and demand; (2) common criteria like enrollment do little to advance either efficiency or equity; and (3) how schools are ranked for evaluation is a surprisingly powerful policy lever. This work contributes to the field by showing how simulation can equip district leaders to anticipate the trade-offs embedded in closure decisions, moving policy design toward proactive fairness.

Keywords: simulation, school closure, district leadership, explainable AI, education policy

1 Introduction

The decision to permanently close a public school represents one of the most highly contested, politically volatile, and structurally consequential actions a district administration can undertake. Historically, district leaders have justified these closures through a technocratic vocabulary emphasizing fiscal solvency, declining enrollment, demographic shifts, and facility underutilization (Hahnel and Marchitello, 2023). This approach often treats the impact of the closures as a secondary concern, something considered after an initial list of schools is generated based on seemingly neutral metrics, if considered at all.

However, as the frequency and scale of school closures have accelerated over the past two decades-spurred by the expansion of charter networks, fluctuating birth rates, and the expiration of pandemic-era stimulus funding-a robust and expanding body of empirical literature has repeatedly demonstrated that these seemingly neutral administrative criteria frequently replicate and exacerbate deep-seated historical inequities (Brazil and Candipan, 2022; Pearman and Greene, 2022; Ewing and Green, 2022). Modern educational research unequivocally indicates that the burdens of school closure are not distributed equally; rather, they disproportionately impact marginalized communities, Black and Hispanic students, and socioeconomically disadvantaged neighborhoods.

Simultaneously, the methodological landscape of public administration is undergoing a profound transformation. As districts increasingly turn toward algorithmic decision-making systems, machine learning heuristics, and data-driven composite scores to identify candidate schools for closure, the intersection of education policy, data science, and spatial analysis has become a critical frontier for academic inquiry (Mutgan and Tapia, 2025; Gillani et al., 2023). While computational tools promise optimized resource allocation and objective governance, without rigorous oversight, they risk laundering historical prejudice through the opacity of mathematical optimization (Caven, 2018). In response, contemporary policy design requires a shift toward the deployment of ”policy sandboxes”-advanced computational simulation environments to safely examine the potential downstream consequences and policy interactions of

administrative decisions prior to real-world implementation.

In a turbulent era where difficult choices are increasingly unavoidable, we argue that the methods used to make these choices must be interrogated to ensure they do not replicate historical disparities and offer a new way forward. This study demonstrates such a way by examining potential outcomes of specific school closure decision rationale and their potential interactions.

The objective of this research is to preemptively unmask the implications school closure criteria. We seek to provide compelling evidence and methodology that challenges policymakers to move toward a proactive stance on designing fair closure plans (Hahnel and Marchitello, 2023). Our primary research questions are:

1. To what extent do commonly used closure criteria (e.g., enrollment, academic performance) effectively generate the desired outcomes of a district closure process?

2. How do explicit equity safeguards alter the realization of the fiscal, operational efficiency, and equitable treatment goals of a closure process?

3. Which closure policy levers might enable districts to balance fiscal efficiency and fairness most effectively?

This research challenges the notion that there is a single ”best” or objective way to approach school closures. By simulating a variety of scenarios for our partner district, this framework demonstrates an avenue for district leaders to predict the conditions under which a given policy is likely to be beneficial or harmful in their specific, situated context (Ewing and Green, 2022). Empirically, this work challenges the assumption that ’neutral’ administrative metrics such as enrollment or building utilization naturally align with fiscal efficiency. Our simulations reveal a quantifiable tension: safeguards explicitly designed to limit disproportionate harm to marginalized communities successfully improve fairness but can constrain potential cost savings and resource reallocation if not examined carefully. This finding provides the necessary evidence to move beyond the illusion of optimization, forcing a transparent reckoning with the financial cost of rectifying historical inequity.

Methodologically, this study advances the field of educational policy analysis by demonstrating the utility of 'policy sandboxes'-computational simulation environments that allow leaders to test the downstream consequences of administrative decisions. Unlike traditional evaluations that assess the damage of closures retrospectively, our framework enables a shift toward proactive fairness. By embedding equity constraints directly into the algorithmic architecture, we demonstrate how districts can mathematically audit their decision logic for structural bias before a single school is identified for closure.

Ultimately, this study contributes to the field by: (1) establishing a 'policy sandbox' closure methodology that shifts closure planning toward proactive fairness; (2) empirically quantifying the trade-offs between fiscal solvency and demographic equity, revealing that common 'neutral' metrics often fail to achieve either; and (3) identifying the ranking order of schools as a decisive policy lever that districts can adjust to better align outcomes with their values.

2 Prior Work

2.1 Educational and Behavioral Impacts of School Closures

The empirical evaluation of school closures has historically concentrated on the immediate academic and behavioral impacts on the students directly displaced by the closure, alongside the collateral spillover effects on the incumbent students attending the receiving schools. As Ewing and Green (2022) note in their comprehensive review of nearly two decades of scholarship, the literature has been heavily indexed to the mass urban school closures of the mid-2010s in cities such as Chicago, Philadelphia, and Detroit, though shifting demographics demand an expansion of this scope to include rural and suburban contexts that literature is just now beginning to address (Chrzan et al., 2025). The consensus within the econometric and educational literature is overwhelmingly cautionary: the outcomes of displaced students are highly variable and almost entirely contingent upon the quality of the receiving institution.

A foundational study by de la Torre and Gwynne (2009), analyzing mass closures in Chicago Public Schools, established the baseline parameters for modern closure research. The researchers discovered that the largest negative impact on students' reading and mathematics achievement actually materialized in the academic year prior to the physical closure of the building. This phenomenon is largely attributed to the severe destabilization of the school environment following the public announcement of the closure list, which triggers rapid teacher attrition, administrative disengagement, and plummeting student morale. Following the closure, the researchers observed that because the vast majority of displaced students (roughly 80%) transferred to schools that were also academically weak and ranking in the bottom half of the district's standardized assessments, the closure failed to significantly improve their academic trajectories. The study cemented the principle that the success of any closure policy hinges not on the act of closing an underperforming school, but on the absorptive capacity and superior academic quality of the receiving schools.

In a highly rigorous evaluation of the 2013 mass closures in Philadelphia, Steinberg and MacDonald (2019) isolated the specific conditions under which these academic gains manifest. Utilizing student-level administrative data, they found that the aggregate effect of school closures on the average achievement of displaced students was strictly null. Achievement only increased among the subset of displaced students who managed to secure enrollment in significantly higher-performing schools following the closure. For students who moved laterally to schools of similar quality, the disruption of the transition yielded no academic dividends.

Despite the lack of an academic effect, the team did identify the behavioral consequences of displacement were detectable. Steinberg and MacDonald identified acute negative behavioral outcomes for displaced students in Philadelphia. School absences and disciplinary suspensions increased significantly for displaced students in the years following the closure. Crucially, the magnitude of these adverse behavioral outcomes scaled linearly with the geographic distance students were forced to travel to their new schools. As transit times increased, student engagement plummeted, highlighting the deep intersection between spatial geography and

educational continuity.

2.2 School Closure as Community Disruption

While psychometric evaluations of academic proficiency provide a quantitative measure of closure impacts, qualitative, phenomenological, and public health research emphasizes the profound socio-emotional trauma inflicted upon communities. Schools do not function exclusively as sites of academic instruction; they serve as highly complex community anchors, providing vital extracurricular programming, pediatric health services, food security, and safe civic gathering spaces.

The phenomenology of a school closure is routinely characterized by the impacted community not as a fiscal necessity, but as a profound civic betrayal and an institutional death. Gary Pierson (2025), in a qualitative case study of school consolidation in California, documented the multifaceted impact of school closures on the resilience of residents. Families of impacted students routinely reported a deep sense of mourning, the severing of trusted, long-term relationships with school personnel, and heightened daily stressors regarding safety concerns and unmanageable transportation logistics. Prior work by Ewing (2018) shared similar stories of impacted communities in Chicago. For historically disenfranchised students, particularly those requiring specialized educational services, Pierson further notes that the disruption of established continuity is severely detrimental. These populations depend heavily on highly localized service delivery models; when schools close, the intricate web of specialized care, paraprofessional support, and tailored pedagogical environments is shattered, often taking years to reconstruct in a receiving school.

2.3 Demographic Disproportionality and Spatial Injustice

The intersection of public administration, geographic spatial analysis, and racial equity forms the core of modern school closure critiques. Despite the persistent use of standard, seemingly race-neutral metrics-such as enrollment counts, seat utilization, and building maintenance

indices-closures consistently and overwhelmingly map onto deeply entrenched lines of racial segregation and spatial inequality.

Nationwide empirical analyses confirm that the burden of school closures is not distributed evenly across populations. Brazil and Candipan (2022) conducted a large geographic evaluation of traditional elementary school closures across 260 U.S. metropolitan areas between 2010 and 2016. Utilizing school attendance boundaries (SABs) to accurately capture the localized demographics of the exact populations served by the facilities, they found that school closures are significantly and positively associated with neighborhoods harboring higher percentages of Black residents and deeper socioeconomic disadvantage. Conversely, closures are associated with neighborhoods having lower percentages of White residents.

This complex relationship between school closures, racial isolation, and urban demographic shifts was definitively modeled by Pearman and Greene (2022). Integrating longitudinal data from the U.S. Census Neighborhood Change Database with the National Center for Educational Statistics, the authors investigated whether severing the historical neighborhood-school connection accelerated the displacement of incumbent residents. The findings isolated a striking and specific racial dimension to gentrification mechanics. Pearman and Greene found that the effects of school closures on subsequent patterns of gentrification were concentrated exclusively within Black neighborhoods. Specifically, the closure of a local public school increased the probability that a highly segregated Black neighborhood would experience gentrification by precisely 8 percentage points, and it increased the overall extent of that gentrification by 0.21 standard deviations. Crucially, despite running multiple alternative model specifications and falsification tests, the researchers found absolutely no evidence that school closures increased the likelihood or extent of gentrification in White or Latinx neighborhoods.

This macro-level national data is heavily corroborated by localized, state-level analyses. Hahnel and Marchitello (2023), examining the school closure process in California, highlight the severe fiscal mechanics driving disproportionality. Because district funding in California is inextricably linked to Average Daily Attendance rates, the state's steeply declining enrollment

forces brutal budgetary contractions. In analyzing the sites chosen for closure to offset these deficits, the authors found a stark racial disproportionality: schools serving low-income students and students of color are shuttered at vastly higher rates than those serving whiter, more affluent populations. Hahnel and Marchitello situate these disparate outcomes within a historical continuum of racially restrictive housing policies, purposeful neighborhood disinvestment, and systemic under-resourcing-all sociological phenomenon outside of school districts' control yet required for their consideration when adjusting the ecology of the schools they have placed in neighborhoods (Lu and Pearman, 2025). Hahnel and Marchitello advocate for local decision-makers to implement explicit equity strategies when making closure decisions.

2.4 The Illusion of Neutral Metrics

The mechanisms by which these profoundly disproportionate outcomes are rationalized and enacted rely heavily on the data-driven policies of modern public administration. Caven (2018) provides a critical sociological deconstruction of this process through a mixed-methods analysis of the 2013 mass closure of 24 public schools in Philadelphia. To generate the closure lists, district officials utilized seemingly objective, race-neutral metrics: building condition (Facilities Condition Index, or FCI), seat utilization rates, projected financial savings, and standardized academic performance.

Caven argues that this process of "quantification"—the reliance on strict numerical data to dictate policy—actively reproduces inequality by systematically stripping schools of their socio-historical context. Because academic underperformance is inextricably correlated with socioeconomic disadvantage, historic underinvestment, and residential segregation, utilizing it as a primary metric inevitably targeted the city's most marginalized communities. Caven's logistic regression demonstrated the stark reality of this data model: in a theoretical school of 100 students, the addition of a single Black student increased the likelihood of a closure recommendation by 3.1%, whereas a $1,000 increase in the surrounding neighborhood's median household income reduced the likelihood by 4.5%. By framing closures through the lens of

academic sanctions, the district effectively punished schools for the downstream results of poverty, laundering structural racism through technocratic data points.

Further, Caven notes a critical limitation to these typical closure processes. They note that when the district relied on the highly complex ”resource management” logics of building utilization algorithms, seating capacity formulas, and capital expenditure projections, affected communities found the mathematical ecosystem too opaque to effectively interpret or challenge. This dynamic cements the need for districts to use easily understandable decision criteria in their processes to include and invest the community in the institutions serving them.

2.5 Algorithmic Decision-Making

The reliance on quantification highlighted by Caven (2018) has rapidly evolved into the deployment of complex algorithmic systems and Artificial Intelligence (AI) across the public sector. From predictive policing and child-welfare risk assessments to resource allocation and school district portfolio management, algorithmic decision-making systems are increasingly utilized to process massive datasets and automate complex governance decisions (Wang, 2024). While these systems are routinely procured with the promise of overcoming flawed human heuristics and maximizing operational efficiency, they frequently encode, scale, and obfuscate historical prejudices, demanding a radical rethinking of how fairness is integrated into policy design.

The traditional approach to algorithmic fairness in machine learning—particularly in high-stakes environments like educational policy and financial compliance—has relied heavily on post-hoc mitigation (Vallarino, 2025). In this reactive paradigm, a model is trained to maximize a primary objective function, such as predictive accuracy or operational efficiency (e.g., maximizing cost savings and seat reductions in a closure scenario). The outputs generated by the algorithm are subsequently audited for disparate impact against protected groups. If bias is detected, downstream adjustments are applied, such as altering decision boundaries, reweighting the final predictions, or retroactively adjusting the thresholds for specific demographics.

However, contemporary computer science and ethical AI literature firmly rejects post-hoc patching as structurally insufficient and theoretically flawed (O'Neil, 2016; Crawford, 2021; Buolamwini, 2023). Because algorithmic bias is not merely a technical error, but a direct manifestation of deep socioeconomic inequalities and historical power structures embedded in the training data, retrofitting fairness at the end of the pipeline cannot rectify fundamental injustices (Vallarino, 2025). The literature increasingly demands a shift toward "proactive fairness". Proactive fairness requires embedding causality-aware decision-making, rigorous testing, and explicit equity constraints directly into the foundational architecture and the objective functions of the models before they are deployed.

In the highly sensitive context of school closures, a proactive fairness framework mandates that demographic disproportionality is not evaluated after an algorithmic heuristic generates a finalized list of schools to close based on utilization or academic rankings. Instead, as demonstrated in the methodologies of advanced simulation algorithms such as those in this study, the proportional demographic impact must be explicitly managed dynamically as the scenario is built. Structural constraints are needed to prevent algorithms from achieving its cost-saving targets through the mathematical exploitation of vulnerable populations. While proactive fairness might force a mathematical trade-off— reducing the maximum achievable operational efficiency or gross financial savings—it provides an essential, non-negotiable governance tool that aligns technical optimization with democratic values, institutional accountability, and regulatory compliance.

2.6 In Silico Experimentation and "Policy Sandboxes"

While machine learning excels at static prediction (e.g. identifying which districts or systems are at risk of failure as in Chrzan et al. (2025)), it often falls short in dynamic policy design (determining what specific interventions will work under shifting parameters). To bridge this "prediction-prescription gap," educational researchers have adapted Agent-Based Modeling and System Dynamics from the fields of epidemiology and computational economics to create what

are known as ”Policy Sandboxes” (Paz, 2025).

A policy sandbox is defined as a transparent, extensible simulation environment that allows institutions to conduct safe, in silico experimentation on policies before committing to costly, large-scale implementation. Unlike static regression models, sandboxes represent heterogeneous actors (students, schools, teachers) as individual ”agents” navigating complex institutional structures governed by sets of programmatic behavioral rules. For example, the CAPIRE Intervention Lab discussed by Paz (2025) utilizes a leakage-aware learning analytics pipeline overlaid on a curriculum graph to simulate thousands of student trajectories under varying bundles of policy interventions (e.g., changing academic support structures or financial aid policies) to view the long-term impact on dropout rates.

The necessity of these simulation environments is further underscored by empirical research demonstrating how seemingly straightforward closure strategies often fail in complex urban systems. For example, utilizing simulation models calibrated with administrative data from primary and lower secondary schools in Stockholm, Sweden, Mutgan and Tapia (2025) demonstrated that the common administrative practice of closing minority-dominated schools in minority-dominated neighborhoods to theoretically spur ethnic integration is largely ineffective. By testing various reallocation scenarios within their simulation, the study revealed that ethnic school segregation meaningfully decreases only if displaced students are deliberately reassigned to schools with different ethnic compositions (heterophilous allocation). When the simulations applied realistic reallocation criteria based on proximity to the nearest school or demographic similarity (homophilous allocation), city-level segregation experienced only marginal declines. This application illustrates precisely why policy sandboxes are essential: they reveal that the ultimate impact of a closure depends profoundly on the local opportunity structure and the explicit rules governing student reallocation, allowing decision-makers to test the validity of their assumptions before implementation.

In the specific context of school closures, our policy sandbox utilizes algorithmic heuristics to generate thousands of hypothetical closure scenarios. By systematically altering and

varying specific design principles and constraints suggested by the prior research discussed to this point—such as maximum transit time for displaced students, minimum academic excellence thresholds for receiving schools, rigid regional balance requirements, and strict demographic disproportionality caps—policymakers can observe the emergent outcomes of their choices.6 This high-volume combinatorial generation allows decision-makers to map the exact mathematical trade-offs between competing public values. For instance, a policy sandbox can quantify precisely how many millions of dollars in potential budget savings must be sacrificed to ensure that the demographic burden of closures does not fall disproportionately on Black and low-income students, or how strictly limiting the capacity of receiving schools impacts overall portfolio efficiency.

3 Data

This study draws on several administrative data sources provided by our partner district and linked at the school level. The analytic dataset was constructed by merging information on school characteristics, enrollment, capacity, student demographics, geographic location, and travel times across candidate school pairings.

School-level characteristics were obtained from the district's resource allocation budget sheet, which provided information on each school's operational cost, enrollment capacity for the 2024–25 academic year, free and reduced-price meal (FRPM) eligibility rates, region designation provided by the district, and Title I status. To account for realistic absorptive capacity, building-level program capacities were supplemented with an updated capacity file, in which each school's reported capacity was inflated by 10 percent. Schools for which updated capacity data were unavailable were excluded from the analytic dataset.

Student enrollment counts - the central metric in determining a school's capacity to operate as a welcoming school for students affected by closures - were drawn from the California Basic Educational Data System (CBEDS) (California Department of Education, 2024).

Grade-level enrollment records were aggregated to the school level and classified into three instructional bands: elementary (TK–5), middle (6–8), and high school (9–12). Students enrolled in vocational or transitional programs (grade code ”VT”) were excluded. Total school enrollment, band-specific enrollment, and FRPM-eligible enrollment counts were computed from these records. A school-level open seats variable was derived as the difference between the updated facility capacity and total CBEDS enrollment.

To support our equity analyses, student subgroup counts were also drawn from the CBEDS enrollment file. School-level counts were compiled for African American students, Hispanic/Latino students, English Language Learners (ELL), students with disabilities (SPED), students experiencing foster care or homelessness (combined as a housing and family instability indicator), and socioeconomically disadvantaged students (proxied by FRPM eligibility). District-wide totals for each subgroup were computed from aggregate records and used to derive district-level proportions, against which school-level subgroup concentrations could be benchmarked.

Per-school cost savings estimates associated with closure - a key outcome we measure for closure scenarios - were assigned based on school level, using values provided by the partner district. Elementary schools were assigned an estimated annual savings of $1,111,072; middle schools $1,245,503; high schools $1,384,453; and K–8 schools $1,587,379.

From these data, we generate school pairing data, reflecting all potential candidate/welcoming combinations based on grade levels served. Each record represents a directed pairing of a potential school closure candidate and a receiving/welcoming school. Pairings were restricted to schools present in the broader school characteristics dataset. It is from these pairings that schools are ranked for consideration for closure by our algorithm and processed (see Section 4.2 and Appendix A).

4 Methods

The study relies on a unified dataset integrating school-level enrollment demographics, building capacity and cost data, and GIS information for a large, diverse urban school district in the U.S.

We investigate potential policies that shape closure decisions. To do so, we developed a simulation framework that systematically generates closure scenarios under different criteria, both traditional and novel. The approach allows us to directly test how variations in these closure design principles create closure lists that we compare across three outcomes of interest: (1) demographic proportionality of the closed schools versus those not targeted, (2) number of seats removed to align the school portfolio with current demand, and (3) the cost-savings accrued as a result of the closures.

4.1 Design Principles

In total we simulated 5,040 closure scenarios. Each scenario was created by applying a distinct combination of values across the seven design principles described in Table 1. For each principle, we varied parameter values to make the policies more or less restrictive, including baseline runs where principles were effectively neutralized.

Given the greedy nature of our algorithm (discussed in Section 4.2), we examine multiple ordering rules for considering and selecting schools (Design Principle 7 in Table 1). Each of the 6 non-random orderings were based on metrics that could be used to rank the schools. After each school's score on the final metric used for ranking (described below) was calculated, they were converted to percentile ranks which were then used to sort the list of schools. Each simulation began with this sorting and then used the sorted list to examine schools one at a time for their eligibility for inclusion in the final scenario closure list based on the set of parameters for that simulation run. More information on each metric used in these orderings can be found in Appendix B.1.

The ordering scores were:

1. **Enrollment**: smallest student body to the largest.

2. **Academic**: lowest academic performance to the highest based on standardized test scores.

3. **Utilization**: schools with less enrollment compared to their building capacity to schools utilizing more of their capacity.

4. **Excellence**: a score combining academic metrics - namely measures of *Academic Performance* (same as the 'Academic' ordering), *Social and Emotional Learning*, and *School Culture and Climate*. The score was calculated using weights determined using community input. Schools were ranked from lowest to highest.

5. **Composite Score (No Equity)**: a score combining academic and resource-use metrics - namely the academic measures named under 'Excellence' as well as *Family Choice and Demand*, *Teacher Turnover*, *Student Enrollment*, and *Building Condition*. As with 'Excellence', the weights for combination were determined by community input and the schools were ranked from lowest to highest.

6. **Composite Score**: a score that combined the previous two composite metrics with measures of equity - namely *Number of Proximate Schools*, *Number of Specialized Programs* (e.g. specific language services), and the *Historical Inequality* of the neighborhood the school served, measured using the Upward Mobility Index by Chetty et al. (2018).

7. **Random**: Randomized orderings which serve as the category for this design principle.

4.2 Simulation

This simulation algorithm is designed to create school closure scenarios based on the design principles' values. It begins by going through the dataset of all school pairings, checking each candidate school against three design principle thresholds - building condition, excellence, and proximity - to determine its baseline eligibility to close. If a school meets these criteria – that is to

say, if the school has enough welcoming schools within the set thresholds for that simulation – further design principles will be considered, such as avoiding closing schools in the same geographical area or making sure to not exceed a given number of welcoming schools required for students to transfer.

Additionally, the algorithm can evaluate the impact of adding a school to a scenario on the scenario's overall disproportionality for specific student subgroups and ensures that the scenario approaches and then stays within acceptable bounds for equitable impact, rejecting schools if they move this disproportionality too far away from the goal, which for these runs is to be less than 3.5% of the district's proportion of the groups.

We consider impacts on six student subgroups of interest: Black students, Latinx students, low-SES students, Special Education students, Homeless and Foster Youth, and English Language Learners. These groups are broadly considered "historically underserved" by the district based on levels of their academic performance over time.

The algorithm iterates through the schools, updating cost savings, seats removed, and other relevant scenario metrics as it progresses, and the scenario is built. The process continues until specific goals are met, namely either reaching target savings or removing a target number of seats. Once the goals are satisfied or all schools have been evaluated, the algorithm returns detailed results, including the final scenario data, disproportionality statistics, and a summary of the scenario's outcomes.

Full details of the algorithm can be found in Appendix A.

4.3 Policy Interaction Detection via Friedman's H-Statistic

To identify meaningful pairwise interactions among design principles prior to specifying our interaction regression models, we employed Friedman's H-statistic (Friedman and Popescu, 2008). This approach was chosen because it is agnostic to functional form and does not require a priori specification of which interactions to test—a meaningful advantage given the combinatorial complexity of seven design principles across three outcomes.

The H-statistic quantifies the proportion of a model's predicted variance attributable to the joint behavior of two features beyond their individual additive contributions. Formally, for a pair of features j and k, the statistic is defined as:

$$H_{jk}^2 = \frac{\sum_i [\hat{f}_{jk}(x_{ij}, x_{ik}) - \hat{f}_j(x_{ij}) - \hat{f}_k(x_{ik}) + f_0]^2}{\mathrm{Var}(\hat{y})} \tag{1}$$

where $\hat{f}_{jk}$ is the two-way partial dependence function, $\hat{f}_j$ and $\hat{f}_k$ are one-way partial dependence functions, and f_0 is the mean prediction. A value of $H_{jk} = 0$ indicates a purely additive relationship between the two features, while larger values indicate stronger non-additive dependence. The H-statistics are best interpreted as relative rankings of interaction strength within a given outcome model rather than as absolute measures, as their magnitudes are sensitive to model specification and feature encoding choices.

We used random forest models (fitted via the `ranger` package in R with 800 trees) as the underlying flexible learners for computing these partial dependence estimates, as their ability to approximate complex, non-linear response surfaces makes them well-suited for detecting interactions that linear models would miss (Wright et al., 2024).

Because H-statistics are estimated from data and subject to sampling variability, we conducted a stratified bootstrap procedure with 300 resamples to construct 95% confidence intervals around each pairwise estimate. This allowed us to distinguish stable interaction signals from those that may reflect sampling noise. Following established heuristics in the literature, we treated pairs whose bootstrap confidence interval lower bound exceeded 0.05 as candidates for inclusion in subsequent interaction regression models, and we prioritized interaction terms where the median bootstrapped H-statistic exceeded 0.10 (a "Moderate" interaction threshold). To maintain interpretability and avoid overfitting, we limited inclusion to the three highest-ranked pairs per outcome.

While there are no universally established cutoffs specifically for the H-statistic, we interpret its magnitude using standard heuristics for variance-explained effect sizes (e.g., values $<$ 0.05 indicating negligible interactions, 0.10–0.20 as moderate, and $>$ 0.20 as strong), consistent

with applied interpretations in the literature (Blake-Mahmud and Struwe, 2019).

4.4 Modeling Design Principle Effects

4.4.1 Regression Specification and Marginal Effects

To quantify the independent and joint effects of each design principle on our three outcomes of interest—disproportionality, removed seats, and cost savings—we estimated a series of ordinary least squares (OLS) regression models. Each outcome was regressed on seven design principle predictors: closure ordering criterion (Simulation), disproportionality consideration (Disproportionality), excellence, facilities condition index (Building Quality), proximity (Time), welcoming school cap (Welcoming Cap), and regional boundary enforcement (Regions). All continuous outcomes were standardized prior to estimation to facilitate comparison of effect magnitudes across outcomes, and all design principle variables were treated as unordered factors with theoretically motivated reference categories (e.g., random ordering as the reference for the closure criterion). This base specification takes the form:

$$Y_i^m = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ki} + \varepsilon_i \tag{2}$$

where Y_i^m is one of the three standardized outcomes, X_{ki} represents the set of design principle indicators, and ε_i is an idiosyncratic error term. Average marginal effects (AMEs) were calculated from each model using the `margins` package in R, which averages the partial derivative of the predicted outcome with respect to each predictor across all observations (Leeper et al., 2024). AMEs provide an interpretable summary of the expected change in a standardized outcome associated with a given design principle relative to its reference category, averaging over the joint distribution of all other covariates.

4.4.2 Interaction Model Specification

The base additive specification assumes that the effect of each design principle is independent of all others—an assumption that, while analytically convenient, may obscure substantively important joint effects. For instance, our disproportionality safeguard may have a meaningfully different impact on a scenario's outcomes depending on whether closures are ordered by academic performance versus enrollment. To relax this assumption in a principled and computationally tractable way, we employed Friedman's H-statistic (Friedman and Popescu, 2008) as a data-driven screening tool to identify candidate interaction pairs prior to augmenting the regression specification (see Section 4.3). Results are visualized in Figure 4.

Pairs whose bootstrapped median H-statistic exceeded 0.05 and whose 95% bootstrap confidence interval lower bound was strictly positive were treated as candidates for inclusion as interaction terms. To maintain interpretability and guard against overfitting, we limited inclusion to the three highest-ranked pairs per outcome. Across all three outcomes, this procedure consistently identified the same three pairs as the strongest interaction signals: (1) closure ordering × disproportionality consideration, (2) closure ordering × regional boundary enforcement, and (3) disproportionality consideration × regional boundary enforcement. These three pairs were accordingly entered as two-way interaction terms in augmented regression models for each outcome:

$$\begin{aligned} Y_i^m = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ki} \\ + \gamma_1 (\text{Ordering} \times \text{Disproportionality}) \\ + \gamma_2 (\text{Ordering} \times \text{Regions}) \\ + \gamma_3 (\text{Regions} \times \text{Disproportionality}) + \varepsilon_i \end{aligned} \tag{3}$$

Because interaction terms render the main effect coefficients conditional—and therefore uninterpretable in isolation—we computed conditional marginal effects (CMEs) from the interaction models using the `margins` package (Leeper et al., 2024). For each interacted pair, we

calculated the marginal effect of one principle at each level of its interacting partner. This yields two complementary conditional quantities for each pair: the effect of activating a design parameter (disproportionality consideration or regional boundary enforcement) conditional on a given ordering criterion being in place, and the effect of adopting a particular ordering criterion conditional on a given design parameter being active. Both perspectives are analytically meaningful. The former reflects the incremental value of adding an equity safeguard or geographic constraint to an already-specified ordering mechanism, while the latter reflects the optimal choice of ordering criterion when a policy constraint is treated as a non-negotiable baseline. Together, the AMEs from the base models and the CMEs from the interaction models constitute the primary inferential evidence for the effects of individual and joint design principles on closure scenario outcomes.

5 Results

5.1 H-statistics

To better understand the potential complex interactions among the design principles, we first calculated pairwise Friedman's H-statistics to measure the strength of two-way feature interactions. Figure 4 illustrates the median H-statistics and their 95% bootstrap confidence intervals across our three primary outcomes: Disproportionality, Number of Removed Seats, and Total Savings.

The analysis reveals a stark contrast between a select few highly influential feature interactions and a long tail of negligible effects. For most variable pairings—such as those involving Building Quality, Welcoming Cap, Time, and Excellence—the interaction strengths remain strictly negligible or weak across all three predictive models, suggesting these features operate largely independently of one another. In contrast, interactions involving the Simulation, Disproportionality, and Regions features demonstrate substantial predictive importance.

The Disproportionality $\times$ Regions interaction is particularly striking; it uniquely exhibits

a "Very Strong" effect (median $H \geq 0.30$) when predicting the Disproportionality outcome itself, though its influence drops to weak or negligible in the Savings and Removed Seats models. Meanwhile, the Simulation $\times$ Disproportionality and Simulation $\times$ Regions interactions maintain consistent, moderate-to-strong importance across all three outcomes, highlighting their systemic role in the data.

Given the pronounced and consistent strength of these specific feature combinations, the three interaction terms—Disproportionality $\times$ Simulation, Simulation $\times$ Regions, and Disproportionality $\times$ Regions—were formally added to the main regression models to more accurately capture the potential policy interactions captured in our simulations.

5.2 Design Principle Effects

Our analysis of the simulated closure scenarios demonstrates how alternative design principles shape the equity, fiscal, and operational outcomes of school closure decisions. Across the modeled scenarios, several clear patterns emerge that highlight the inherent tension between achieving operational efficiency and ensuring demographic fairness with rare policies that achieve all desired outcomes.

Receiving School Quality and Continuity. As displayed in Figure 2, design principles intended to prioritize the quality of the receiving environment for displaced students frequently produce counterproductive results. For example, policies that most strictly cap the number of welcoming schools to preserve student cohorts severely constrain the district's ability to meet its fiscal targets (-0.50 SD, $p < 0.001$) and paradoxically worsen demographic disproportionality (-0.58 SD, $p < 0.001$). Similarly, requiring that displaced students transfer to receiving schools with strictly higher academic excellence backfires operationally and equitably; it depresses proportionality (-0.25 SD, $p < 0.001$), total savings (-0.23 SD, $p < 0.001$), and seat reductions (-0.21 SD, $p < 0.001$). In contrast, prioritizing receiving schools based on strictly better building quality is one of the few design principles that yields weak to insignificant effects across all three outcomes, having only a modestly positive effect on removing seats (0.08 SD, $p < 0.001$).

Equity Safeguards and Geographic Constraints. Further, policies that explicitly account for disproportionality meaningfully improve fairness. Implementing a strict disproportionality safeguard acts as a highly effective mechanism for protecting marginalized groups ($+1.0$ SD, $p < 0.001$). However, this improvement comes at a direct operational cost, significantly reducing the total financial savings (-1.4 SD, $p < 0.001$) and the number of surplus seats removed across the district (-1.5 SD, $p < 0.001$). Enforcing regional boundaries during the closure process ("Go by Regions") presents a similar dynamic, yielding positive effects on proportionality ($+0.5$ SD, $p < 0.001$) while actively hindering both savings (-0.2 SD, $p < 0.001$) and seat reduction goals (-0.3 SD, $p < 0.001$).

When analyzing the conditional marginal effects displayed in Figure 3, we find that the interaction between these two constraints mainly offers changes to their effects on the proportionality of the scenario. Conditional on enacting the disproportionality safeguard, the effect of considering regions stays about the same for savings and seats removed though removes the effect on disproportionality. Were the district to consider regions first instead, no effects on any of the three outcomes change direction but some changes in effect magnitude arise, with the largest changes being for the proportionality of the scenario (from $+1.0$ SD up to $+1.5$ SD if regions were ignored and down to $+0.5$ SD if regions were considered).

The Power of Algorithmic Ordering. The order in which schools are evaluated for closure exerts the strongest influence of all tested design principles. Traditional, "neutral" criteria perform poorly when applied as ranking mechanisms. Ordering schools strictly by lowest enrollment produces negative outcomes across the board, failing to improve proportionality (-0.35 SD, $p < 0.001$), savings (-0.27 SD, $p < 0.001$), or seat reductions (-0.49 SD, $p < 0.001$). While ordering by academic performance successfully reduces excess seats (0.07 SD, $p < 0.01$), it has no effect on savings (-0.01 SD, $p > 0.05$) yet severely worsens disproportionality (-0.49 SD, $p < 0.001$).

Additionally, the conditional marginal effects reveal that algorithmic ordering mechanisms are highly sensitive to other design principle's explicit constraints. For instance,

applying the disproportionality safeguard to the highly inequitable "Composite (No Equity)" ordering drastically corrects its proportionality effect, shifting it from -0.85 SD ($p < 0.001$) - the largest average negative effect on scenario proportionality - up to one of the strongest positive effects on scenario proportionality we discovered, 1.68 SD ($p < 0.001$). This demonstrates that strict algorithmic parameters can effectively neutralize the fairness penalties of flawed ranking metrics, albeit at the potential cost of reduced operational efficiency.

Fundamentally, we find that a fully integrated composite score—one that explicitly balances academic, operational, and historical equity metrics—is the only evaluated ordering approach that simultaneously exerts a positive average effect on proportionality (0.13 SD, $p < 0.001$), total savings (0.12 SD, $p < 0.001$), and seat removal (0.11 SD, $p < 0.001$). We see the power of this approach further exemplified in its conditional effects in Figure 3, where it maintains a positive effect when implemented after a decision on region consideration is made and even after disproportionality is considered. However, it is also worth noting that these benefits do not appear in the reverse - if our partner district used the composite ranking order and then considered other factors - such as ensuring use of the disproportionality safeguard - these gains would be lost and the prior tension of the trade-offs across the three outcomes would reappear.

6 Discussion

The simulation of 5,040 closure scenarios using administrative data from a single, large urban district highlights an apparent tension between fiscal efficiency and racial and socioeconomic equity. When the algorithm is explicitly constrained to avoid disproportionate demographic impacts, fairness improves significantly, but at a direct and measurable cost to potential savings and seat reductions. Conversely, common "neutral" design principles—such as prioritizing schools for closure based strictly on low enrollment or academic underperformance—fail to produce equitable outcomes and - often - do little to advance fiscal goals.

Furthermore, well-intentioned constraints aimed at improving the displaced student

experience often yield perverse operational results. For example, strictly capping the number of receiving schools to preserve student cohorts severely restricts the district's ability to meet operational targets and paradoxically worsens disproportionality. Requiring displaced students to attend receiving schools with strictly higher academic excellence similarly depresses proportionality, savings, and seat reductions.

The most promising policy lever identified in this study is the ranking mechanism, or "ordering," used to evaluate schools for closure consideration. Our findings demonstrate that a fully integrated composite score—which balances academic, operational, and historical equity metrics—is the only tested ordering approach that simultaneously improves proportionality, total savings, and seat removal. Stripping the equity metrics from this composite score completely eliminates its fairness benefits, plunging the proportionality outcomes into the negative.

Crucially, it must be recognized that the specific effect sizes, operational penalties, and policy interactions observed in these simulations are inextricably tied to the geographic topology, facility portfolio, and demographic distribution of our partner district. Because spatial distribution and historical segregation vary wildly across municipalities, these exact mathematical trade-offs cannot be universally generalized across all educational systems. However, this limitation precisely reinforces the central thesis of this work: there is no universal "best practice" or perfectly neutral algorithm for school closures. The interventions that successfully balance efficiency and fairness in one city may fail completely in another due to the localized spatial realities of the neighborhoods they serve. It is precisely for this reason we present not just the results from our work with one district but the algorithmic framework we developed such that these sandboxes can be used to help districts build their data-informed decision-making for their localized context.

Together, these findings illustrate both the difficulty of trade-offs and the possibility of identifying policies that allow districts to better balance competing demands. Safeguards against disproportionality can be costly but are effective. Traditional criteria such as enrollment perform poorly on all fronts. And ordering rules offer a promising path for reconciling fiscal responsibility

with equitable treatment of students in finding closure scenarios that meet district needs.

7 Limitations

While this study provides a robust computational framework for evaluating closure policies, several limitations must be acknowledged.

First, the simulated scenarios, effect sizes, operational penalties, and policy interactions are inextricably tied to the specific geographic topology, facility portfolio, and demographic distribution of a single partner district. Because localized spatial realities and historical segregation patterns vary widely across municipalities, these exact mathematical trade-offs cannot and should not be universally generalized to all educational systems. It is for this reason we share the algorithm we devised (Appendix A) so that the methodology can be replicated to ensure reliable, localized results.

Further, there is a discrepancy between the methods used for interaction detection and effect estimation. To identify policy interactions, the study relied on Friedman's H-statistic derived from flexible random forest models, which are highly capable of capturing complex, non-linear dynamics. However, the subsequent calculation of marginal effects relied on linear ordinary least squares (OLS) regression models. This linear specification may obscure substantively important joint effects that the random forest algorithm detected. This choice was made to align the report of effects with those expected in policy discourse and as the field of explainable AI continues to improve and converge with policy analysis, we hope further analysis can be done where the models are rectified. Notably - as a robustness check - we examined the SHAP (SHapley Additive exPlanations) dependence plots across our three interactions in Appendix B and did not find that these values contradicted our reported findings.

A final limitation here is - due to work with our partner - the simulations held the disproportionality threshold strictly constant, requiring that marginalized group representation not exceed 3.5% above the district average. Keeping this threshold static prevents an exploration of

how slightly more or less restrictive equity bounds might dynamically alter fiscal and operational outcomes and continues as a potential fruitful path for study.

8 Conclusion

As the frequency of school closures accelerates, district leaders are increasingly forced to make structurally consequential decisions that have historically placed disproportionate burdens on marginalized communities. This study contributes to the field by demonstrating how the deployment of ”policy sandboxes”—computational simulation environments—can equip administrators to preemptively anticipate the consequences of their design choices before they are implemented.

By quantifying the specific trade-offs between fiscal solvency and demographic equity, we illustrate that proactive fairness is not merely a theoretical ideal, but a workable mathematical constraint. Because the impacts of closure criteria are highly dependent on localized contexts, districts cannot rely on generalized heuristics imported from other municipalities. Instead, leaders must utilize localized simulation frameworks to map their own unique efficiency-equity trade-offs, moving the governance of school closures toward transparent, proactive policy design.

8.1 Future Work

Building upon the ”policy sandbox” framework established in this study, future research should focus on expanding the model’s applicability and refining its methodological architecture.

Future research must apply this simulation framework to a broader variety of educational contexts, including both other urban districts as well as rural and suburban districts. Because closure impacts are highly dependent on local opportunity structures, testing the sandbox in municipalities with different spatial distributions will help determine how localized realities shift the balance between efficiency and equity.

To address the current modeling limitations, future iterations of this work should align the

interaction detection and effect estimation phases. Utilizing non-linear regression techniques or advanced machine learning approaches for the final outcome modeling would ensure that the complex dynamics identified by the H-statistics are fully captured in the calculated marginal effects.

Acknowledgments

We are grateful to Mele Lau-Smith, Laura Wentworth, Ritu Khanna, and Moonhawk Kim for their extraordinary collaboration and teamwork throughout this research–practice partnership that made this study possible.

Table 1: Design Principles and Simulation Values

	Design Principle	Criteria for Closure	Values for Simulation
(1)	Proximity	Students must have access to a receiving school within [value].	[10, 15, 20, 25, None*] minutes
(2)	Academic Performance	Students can only be sent to a receiving school with [value] academic performance.	[Strictly better, Better than 10% less than the closed school's, Any*] ranking
(3)	Building Quality	Students can only be sent to a receiving school with [value] building quality.	[Strictly better, Better than 10% less than the closed school's, Any*] ranking
(4)	Cohort Continuity	There are enough open seats at [value] schools to fit the capacity of the closed school, limiting the number of possible welcoming schools for displaced students.	[2, 3, 5, None*]
(5)	Regional Balance	A school must be selected for closure from each district pre-defined region before a 2nd school within a region can be selected.	[TRUE, FALSE*]
(6)	Disproportionate Marginalized Group Impact	A school's closure must not increase the scenario's representation of marginalized groups (see *Note*) beyond 3.5% above the district average.	[Considered, Ignored*]
(7)	School Ordering	Order the schools are examined for closure in the algorithm.	[Academic, Enrollment, Utilization, Excellence, Composite Score (No Equity), Composite Score, Random*]

Note: This table outlines the seven policy design principles systematically varied to generate the 5,040 unique school closure scenarios. Each principle represents a specific parameter or constraint applied by the simulation algorithm when evaluating candidate schools for closure and assigning displaced students. Parameter values were varied across simulation runs to make the policies more or less restrictive, allowing for the observation of emergent outcomes based on different combinations of rules. Values marked with an asterisk (*) denote the reference category, representing baseline runs where that specific principle was effectively neutralized or ignored. The School Ordering principle dictates the initial ranking sequence used to evaluate schools sequentially. For the Disproportionate Marginalized Group Impact constraint, evaluated subgroups include Black students, Hispanic students, Homeless/Foster Youth, Special Education (SPED) students, English-Language Learners, and Low-Income students (proxied by Free/Reduced Price Meals eligibility).

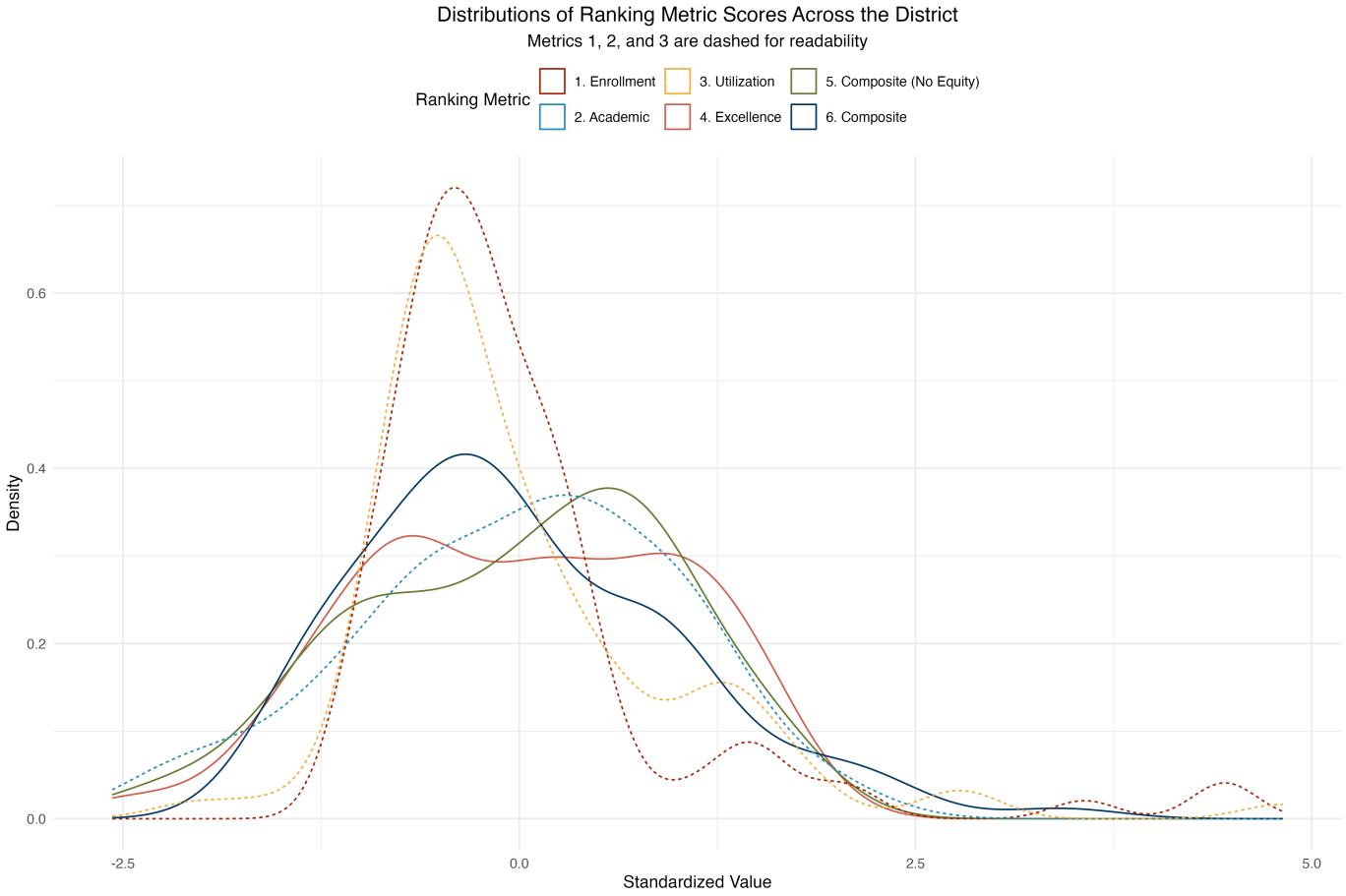

Figure 1: **Distributions of Ordering Metrics Standardized Scores.** This figure presents the density estimates for the six non-random sorting criteria utilized within the simulation framework to rank candidate schools for closure consideration. The evaluated metrics include Enrollment, Academic Performance, Capacity Utilization, Composite Excellence, a Composite score lacking equity measures, and a fully integrated Composite score. All underlying continuous variables were standardized prior to their conversion into the percentile ranks used by the algorithmic selection process, enabling direct structural comparison of the varying distributions across the district's portfolio.

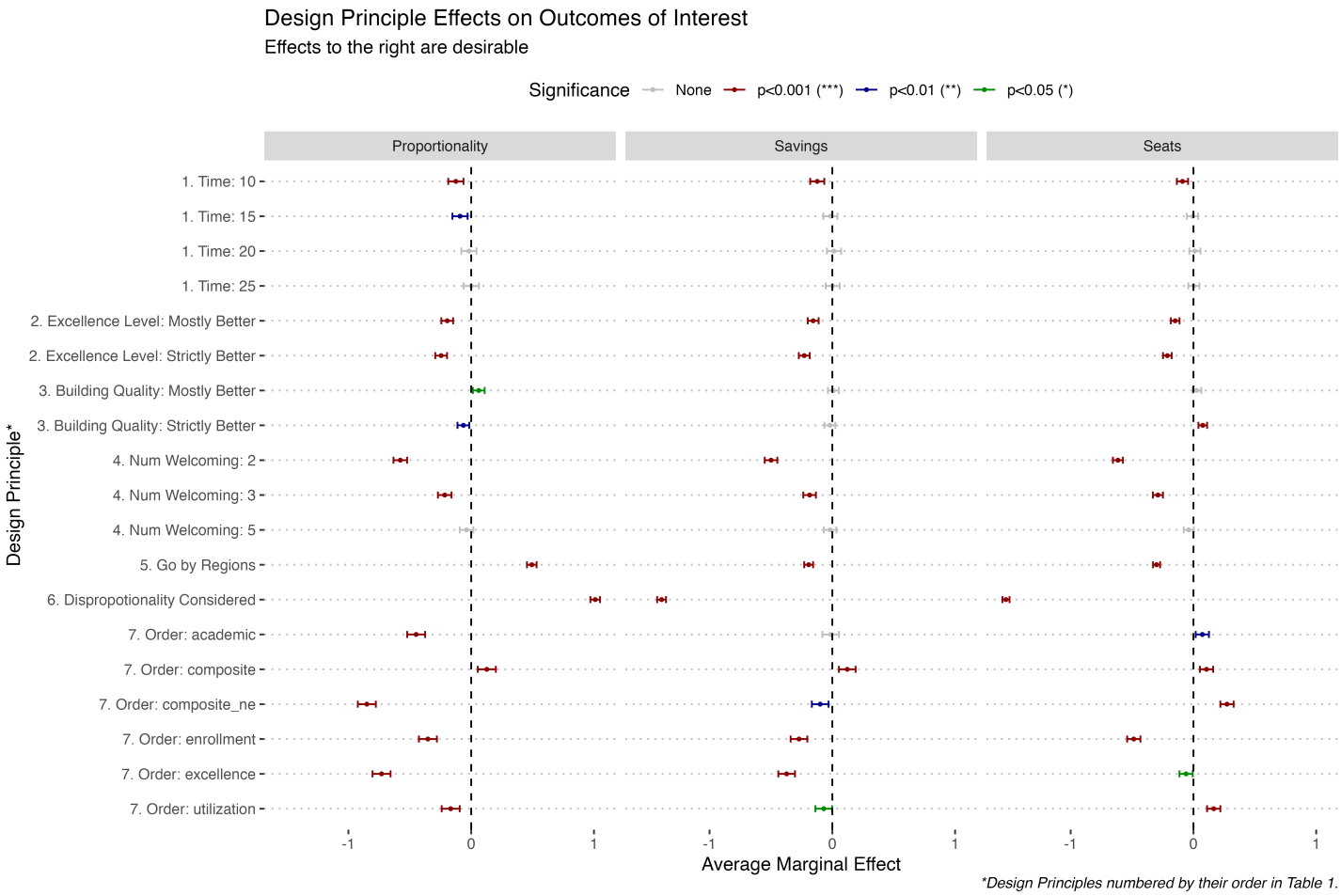

Figure 2: **Average Marginal Effects of Design Principles on Key Policy Outcomes.** This coefficient plot illustrates the estimated impact of various design principles on three outcome dimensions: Proportionality, Savings, and Seats. The independent variables (y-axis) are grouped by design principle categories, including geographic constraints (”Go by Regions”), equity metrics (”Disproportionality”), quality controls (”Excellence Level,” ”Building Quality”), capacity thresholds (”Num Welcoming”), sorting algorithms (”Order”), and temporal constraints (”Time”). The x-axis represents the average marginal effect. The vertical dashed line at zero indicates no effect; estimates falling to the right of this line represent desirable outcomes, while those to the left represent undesirable outcomes. Error bars indicate confidence intervals. Statistical significance is denoted by color and symbol: Red ($^{***}, p < 0.001$), Blue ($^{**}, p < 0.01$), Green ($^{*}, p < 0.05$), and Grey (non-significant).

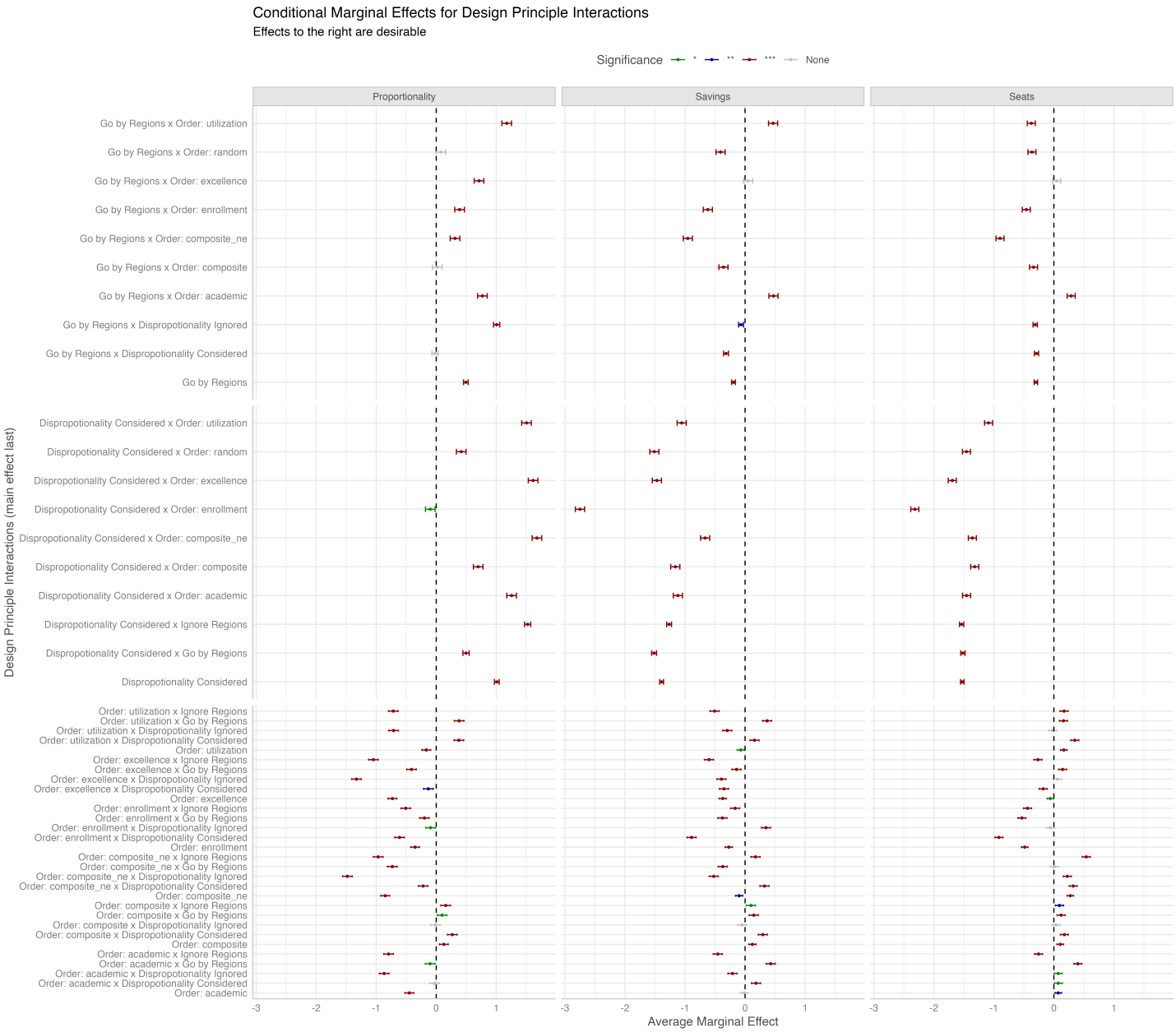

Figure 3: **Conditional Marginal Effects of Design Principles on Key Policy Outcomes.** This coefficient plot illustrates the estimated impact of various design principle interactions on three outcome dimensions: Proportionality, Savings, and Seats. The independent variables (y-axis) are grouped by the focal design principle of the interaction, with the last row of each group being the Average Marginal Effect of the design principle for comparison. The x-axis represents the marginal effect. The vertical dashed line at zero indicates no effect; estimates falling to the right of this line represent desirable outcomes, while those to the left represent undesirable outcomes. Error bars indicate confidence intervals. Statistical significance is denoted by color and symbol: Red ($^{***}, p < 0.001$), Blue ($^{**}, p < 0.01$), Green ($^{*}, p < 0.05$), and Grey (non-significant).

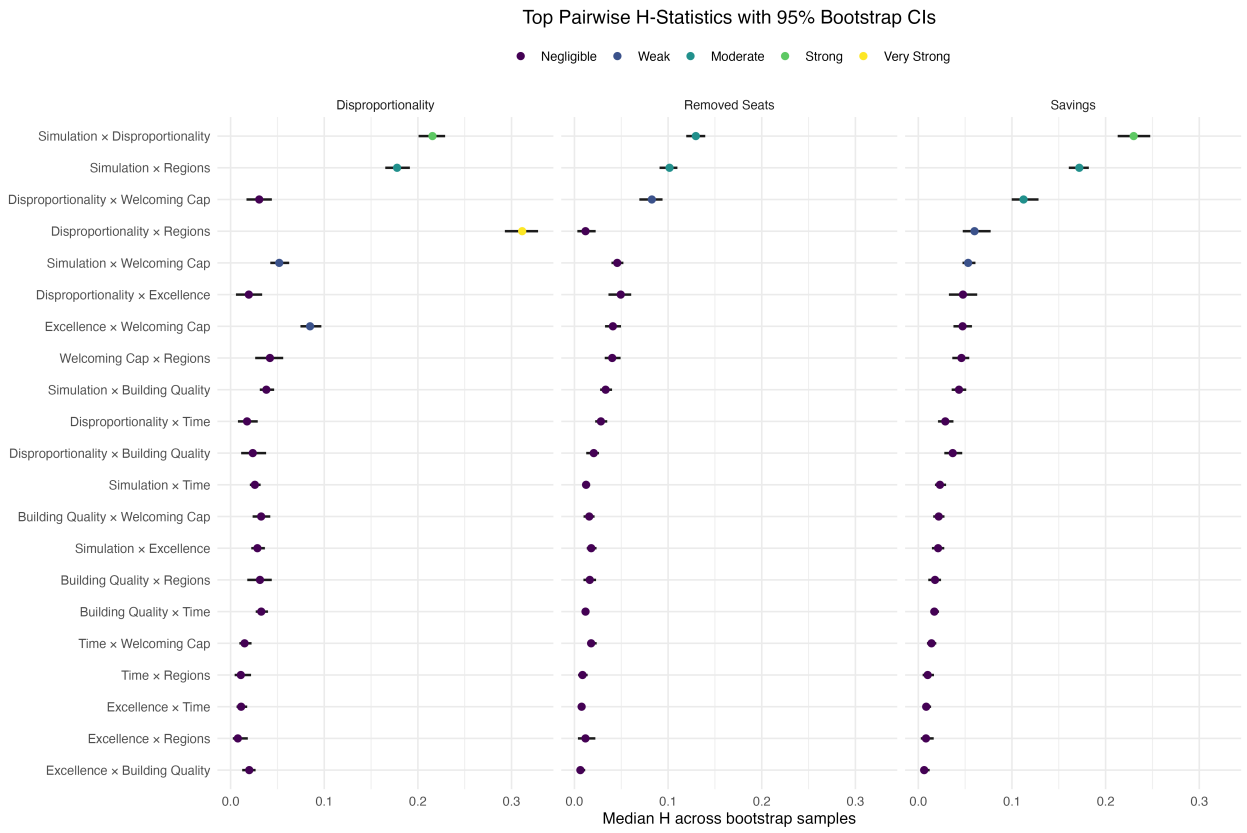

Figure 4: **Top pairwise feature interactions measured by H-statistics across three outcome models.** The plot displays the median Friedman's H-statistic (x-axis) for the two-way feature interactions (y-axis) across three models predicting the Disproportionality, Number of Removed Seats, and Total Savings of the scenarios based on the design principles for the simulation run. Points represent the median H-statistic derived across bootstrap samples, with horizontal lines indicating the corresponding 95% bootstrap confidence intervals. The color of each point denotes the categorical strength of the interaction effect, ranging from Negligible (purple) to Very Strong (yellow). Notably, interactions involving Simulation and Regions consistently demonstrate moderate to strong effects across all three outcomes, while the Disproportionality $\times$ Regions interaction exhibits a very strong, highly localized effect in the Disproportionality model.

References

Blake-Mahmud, J. and L. Struwe (2019, October). Time for a change: patterns of sex expression, health and mortality in a sex-changing tree. *Annals of Botany 124*(3), 367–377.

Brazil, N. and J. Candipan (2022, March). The neighborhood ethnoracial and socioeconomic context of public elementary school closures in U.S. metropolitan areas. *Social Science Research 103*, 102655.

Buolamwini, J. (2023). *Unmasking AI: a story of hope and justice in a world of machines* (First edition ed.). New York: Random House.

California Department of Education (2024). CBEDS Data about Schools & Districts.

Caven, M. (2018). Quantification, Inequality, and the Contestation of School Closures in Philadelphia. *Sociology of Education 92*, 21 – 40.

Chetty, R., J. Friedman, N. Hendren, M. Jones, and S. Porter (2018, October). The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility. Technical Report w25147, National Bureau of Economic Research, Cambridge, MA.

Chrzan, M. L., F. A. Pearman II, and B. W. Domingue (2025, June). Deeper Roots Before the Storm: Utilizing Machine Learning to Alert School Districts of Permanent School Closures. Master's thesis, Stanford University Graduate School of Education.

Crawford, K. (2021). *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

de la Torre, M. and J. Gwynne (2009, October). When Schools Close Effects on Displaced Students in Chicago Public Schools. Technical report, Consortium on Chicago School Research at the University of Chicago Urban Education Institute.

Ewing, E. L. (2018). *Ghosts in the schoolyard: racism and school closings on Chicago's South side*. Chicago: The University of Chicago Press.

Ewing, E. L. and T. L. Green (2022, January). Beyond the Headlines: Trends and Future Directions in the School Closure Literature. *Educational Researcher 51*(1), 58–65.

Friedman, J. H. and B. E. Popescu (2008, September). Predictive learning via rule ensembles. *The Annals of Applied Statistics 2*(3). arXiv:0811.1679 [stat].

Gary Pierson (2025). *Community Impact of School Closure: Implications for Access to Socially Valued Resources and Services in an Urban Community*. Dissertation, Claremont Graduate University.

Gillani, N., D. Beeferman, C. Vega-Pourheydarian, C. Overney, P. Van Hentenryck, and D. Roy (2023, August). Redrawing Attendance Boundaries to Promote Racial and Ethnic Diversity in Elementary Schools. *Educational Researcher 52*(6), 348–364.

Google (2023). Google Maps Distance Matrix API.

Hahnel, C. and M. Marchitello (2023, September). Centering Equity in the School-Closure Process in California. Technical report, Policy Analysis for California Education.

Leeper, T. J., J. Arnold, V. Arel-Bundock, J. A. Long, and Bolker (2024, July). margins: Marginal Effects for Model Objects.

Lu, A. and F. A. Pearman, II (2025, November). Schools Never Die: Toward a Dynamic Systems Theory of School Closure. EdWorkingPaper 25-1327, Annenberg Institute at Brown University.

Mutgan, S. and E. Tapia (2025, February). Can school closures decrease ethnic school segregation? Evidence from primary and lower secondary schools in Stockholm, Sweden. *Journal of Urban Affairs 47*(2), 404–427.

O'Neil, C. (2016). *Weapons of math destruction: how big data increases inequality and threatens democracy* (First edition ed.). New York: Crown.

Paz, H. R. (2025, November). CAPIRE Intervention Lab: An Agent-Based Policy Simulation Environment for Curriculum-Constrained Engineering Programmes. arXiv:2511.18145 [cs].

Pearman, F. A. and D. M. Greene (2022, July). School Closures and the Gentrification of the Black Metropolis. *Sociology of Education 95*(3), 233–253.

Steinberg, M. P. and J. M. MacDonald (2019, April). The effects of closing urban schools on students' academic and behavioral outcomes: Evidence from Philadelphia. *Economics of Education Review 69*, 25–60.

Vallarino, D. (2025, March). Causal GNNs and the Anthropology of Data: An Ethical Approach to Fairness and Transparency in Financial AI.

Wang, Y. (2024, November). Algorithmic decisions in education governance: implications and challenges. *Discover Education 3*.

Wright, M. N., S. Wager, and P. Probst (2024, November). ranger: A Fast Implementation of Random Forests.

A Algorithm Methodology

A.1 Problem Formulation

Let $\mathscr{S}_c$ denote the set of candidate schools for closure and $\mathscr{S}_r$ the set of potential /welcoming schools.

For each school $i \in \mathscr{S}_c$, we define:

- E_i: total enrollment
- E_i^{elem}, E_i^{mid}: elementary and middle school enrollment for K-8 schools
- Q_i^{ex}: academic excellence score rank
- Q_i^{fci}: facility condition index rank
- L_i: school level (elementary, middle, high school, K-8)
- $\mathbf{D}_i = (D_i^1, D_i^2, \ldots, D_i^k)$: demographic composition vector for k

For each receiving school $j \in \mathscr{S}_r$:

- C_j: total capacity
- S_j: open seats ($S_j = C_j - E_j$)
- T_{ij}: travel time from school i to school j

A.2 Design Principles

The algorithm enforces six design principles as constraints, described here. Values for thresholds, τ, used in the simulation runs are described in Table 1.

1. **Proximity Principle**: Students must have access to a receiving school within a set travel time threshold τ_{time}:

$$T_{ij} \leq \tau_{time}$$

2. **Academic Performance Principle**: Receiving schools are prioritized based on higher academic excellence scores. Students must transfer to schools with academic performance at least as good as their current school, within a threshold τ_{ex}:

$$Q_j^{ex} \geq Q_i^{ex} - \tau_{ex}$$

The exact values of τ_{ex} used as the values described in Table 1 are:

(a) "Strictly better": $\tau_{ex} = 0$

(b) "Better than 10% less than the closed school': $\tau_{ex} = 0.1$

(c) "Any": $\tau_{ex} = 1$

3. **Building Quality Principle**: Receiving schools are prioritized based on stronger facilities. Receiving schools must have facility conditions at least as good as the closing school, within threshold τ_{fci}:

$$Q_j^{fci} \geq Q_i^{fci} - \tau_{fci}$$

The exact values of τ_{ex} used as the values described in Table 1 are the same as those described for the **Academic Performance Principle**.

4. **Cohort Continuity Principle**: The number of possible receiving schools for displaced students is limited to n_{max} to minimize disruption and maintain community cohesion:

$$|\mathscr{R}_i^*| \leq n_{max}$$

5. **Regional Balance Principle**: School closures are distributed across geographic regions to prevent concentration of closures in any single area. This is enforced through regional constraints that can skip schools in regions that have already had a school added to the scenario's closure list.

6. **Disproportionality Principle**: The proportional demographic impact is explicitly managed as each school is added to the closure scenario. Let P_g be the district-wide proportion of group g and P_g^{scen} the proportion among students in closed schools. Then:

$$|P_g^{scen} - P_g| \leq \tau_{dispro} \quad \forall g \in \{1, \ldots, k\}$$

This constraint is monitored dynamically throughout scenario construction.

A.3 Objective Function

The algorithm seeks to achieve target goals for:

- Cost savings: G_{cost}
- Seat utilization improvement: Maximum acceptable underutilization rate G_{util}
- Minimum seats removed: G_{seats}

A.4 Algorithm Description

The algorithm proceeds through three main phases: candidate identification, candidate reduction, and student assignment. Schools are examined in order of a composite vulnerability score that combines enrollment, facility condition, and other factors.

A.4.1 Phase 1: Candidate Identification

For each school i under consideration for closure, we identify the set of feasible receiving schools $\mathscr{R}_i \subseteq \mathscr{S}_r$ that satisfy:

$$L_j \in \text{compatible}(L_i) \tag{A.1}$$

$$S_j > 0 \tag{A.2}$$

$$Q_j^{fci} \geq Q_i^{fci} - \tau_{fci} \tag{A.3}$$

$$Q_j^{ex} \geq Q_i^{ex} - \tau_{ex} \tag{A.4}$$

$$T_{ij} \leq \tau_{time} \tag{A.5}$$

The function compatible(L_i) returns the set of school levels that can serve students from level L_i. For example:

- If $L_i =$ Elementary, then compatible$(L_i) = \{$Elementary, K-8$\}$
- If $L_i =$ Middle, then compatible$(L_i) = \{$Middle, K-8$\}$
- If $L_i =$ K-8, then compatible$(L_i) = \{$Elementary, Middle, K-8$\}$

Candidate schools are ordered by capacity in descending order to prioritize larger receiving schools.

A.4.2 Phase 2: Candidate Reduction

To minimize disruption, we reduce $|\mathscr{R}_i|$ to the minimum number of schools needed to accommodate all students from school i. For standard school levels:

$$|\mathscr{R}_i^*| = \min \left\{ n : \sum_{j \in \mathscr{R}_i^{(n)}} S_j \geq E_i \right\}$$

where $\mathscr{R}_i^{(n)}$ denotes the n highest-capacity schools in $\mathscr{R}_i$.

For K-8 schools, we must separately ensure sufficient elementary and middle school seats:

$$\sum_{j \in \mathscr{R}_i^{(n)}: L_j \in \{\text{Elem, K-8}\}} S_j \geq E_i^{elem} \tag{A.6}$$

$$\sum_{j \in \mathscr{R}_i^{(n)}: L_j \in \{\text{Middle, K-8}\}} S_j \geq E_i^{mid} \tag{A.7}$$

A.4.3 Phase 3: Student Assignment

Students are assigned to receiving schools sequentially. For each receiving school $j \in \mathscr{R}_i^*$, we determine the number of students to assign based on available capacity and remaining enrollment. The assignment process accounts for both grade-level compatibility for K-8 schools and real-time capacity updates.

A.5 Equity Monitoring

After each school closure, the algorithm updates cumulative disproportionality metrics. For each demographic group g, we compute:

$$\Delta_g = \frac{\sum_{i \in \mathscr{C}} D_i^g}{\sum_{i \in \mathscr{C}} E_i} - P_g$$

where $\mathscr{C}$ is the set of closed schools. If $|\Delta_g| > \tau_{dispro}$ for any group g, and the cumulative enrollment of closed schools exceeds a buffer threshold, the algorithm terminates to prevent disproportionate impacts. For the simulations described in this paper, τ_{dispro} is held constant at 3.5%.

A.6 Termination Conditions

The algorithm terminates when any of the following conditions is met:

1. Cost savings goal achieved: savings $\geq G_{cost}$
2. Seat utilization goal achieved: underutilization $\leq G_{util}$

3. Seat removal goal achieved: seats removed $\geq G_{seats}$

4. Disproportionality threshold exceeded

5. All candidate schools examined

6. No feasible receiving schools found for current candidate

B Additional Tables and Figures

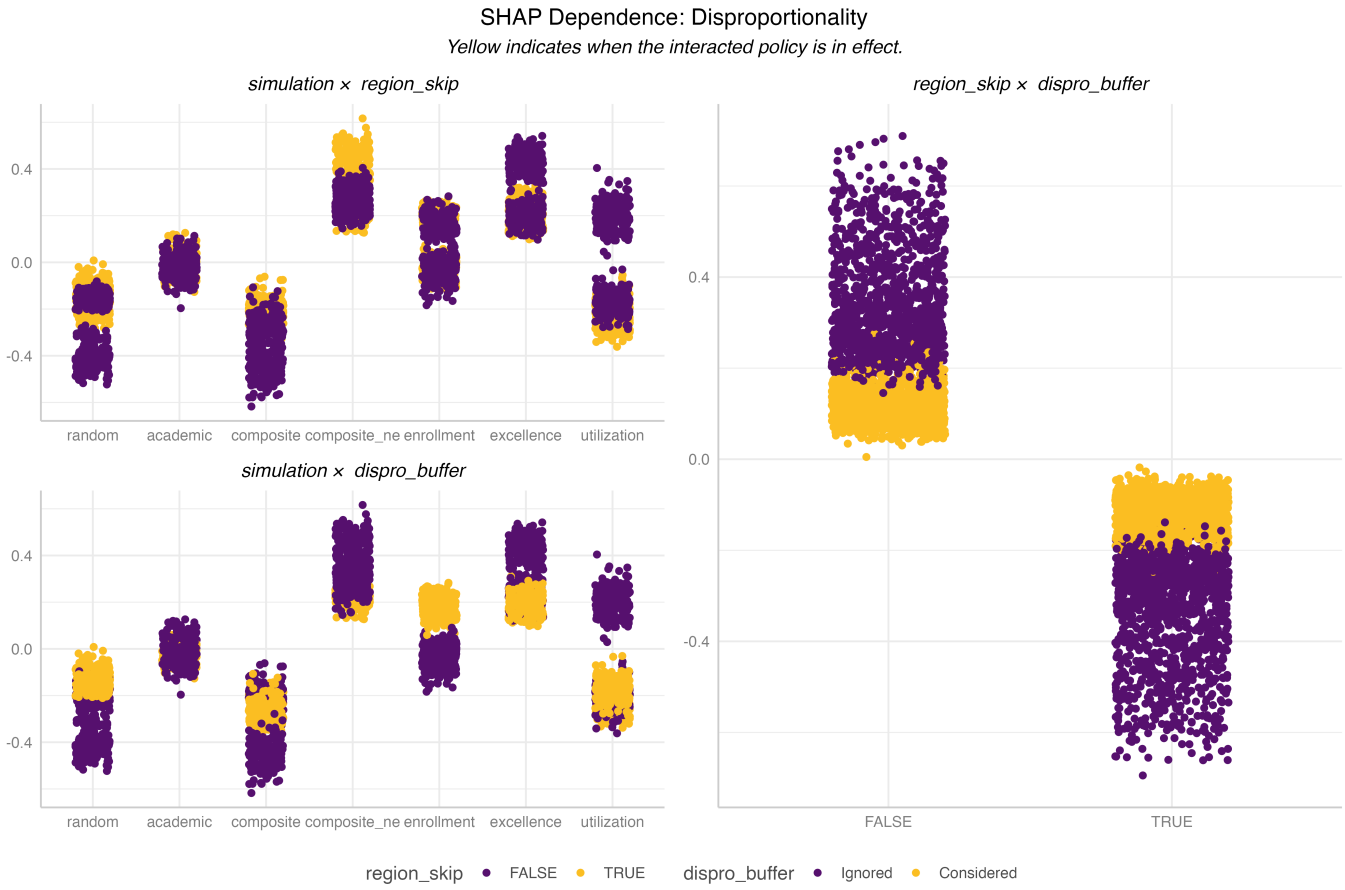

Figure B.1: **Interactive Effects of Design Principles on Scenario Disproportionality.** This figure presents SHAP (SHapley Additive exPlanations) dependence plots detailing the three strongest two-way feature interactions predicting the demographic proportionality of a closure scenario ($N = 5,040$). The y-axis represents the SHAP value, where a higher positive value indicates an undesirable increase in disproportionality, and a lower negative value indicates improved demographic fairness. The color grouping distinguishes when the interacting policy is active (Yellow: TRUE or Considered) versus inactive (Purple: FALSE or Ignored). As shown in the bottom-left and right panels, activating the disproportionality safeguard (dispro_buffer) systematically moderates SHAP values across most conditions, demonstrating its effectiveness in protecting marginalized groups and dampening an ordering's effect on disproportionate impact of the scenario. Notably, the composite_ne (No Equity) ordering mechanism produces massively high disproportionality when unchecked, but this inequitable effect is closer to neutralized when the dispro_buffer is applied.

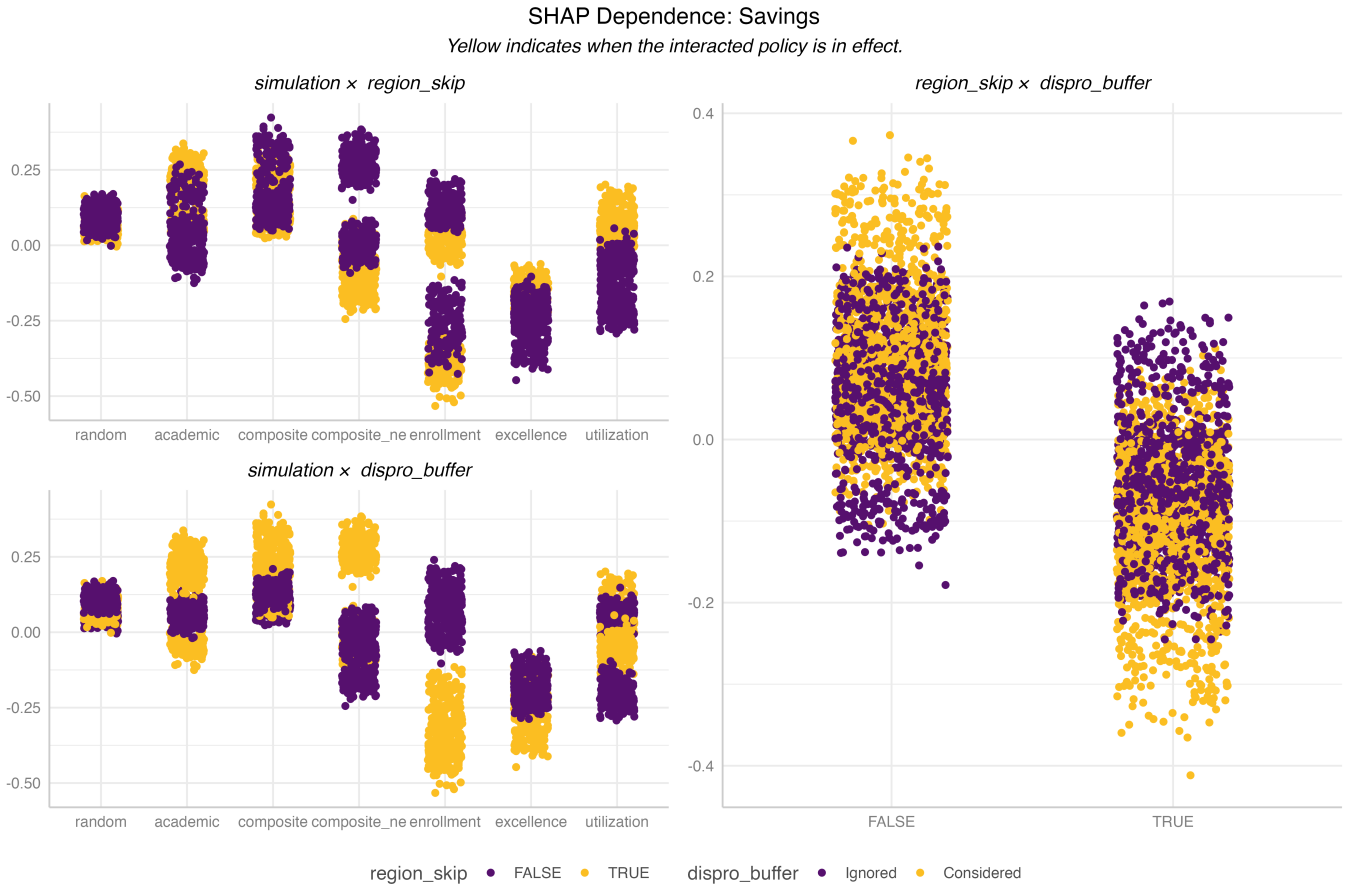

Figure B.2: **Interactive Effects of Design Principles on Scenario Cost Savings.** This figure illustrates the joint impact of school ordering (simulation), regional boundaries (region_skip), and equity safeguards (dispro_buffer) on the financial savings generated by closure scenarios, represented by each point on each figure ($N = 5,040$). Positive SHAP values along the y-axis denote a desirable increase in cost savings within a scenario attributable to the features. Across all three interactions we see that when regions and disproportionality are explicitly considered, they appear to have the effect of increases the dispersion of the other principles' effects on savings when they have any effect.

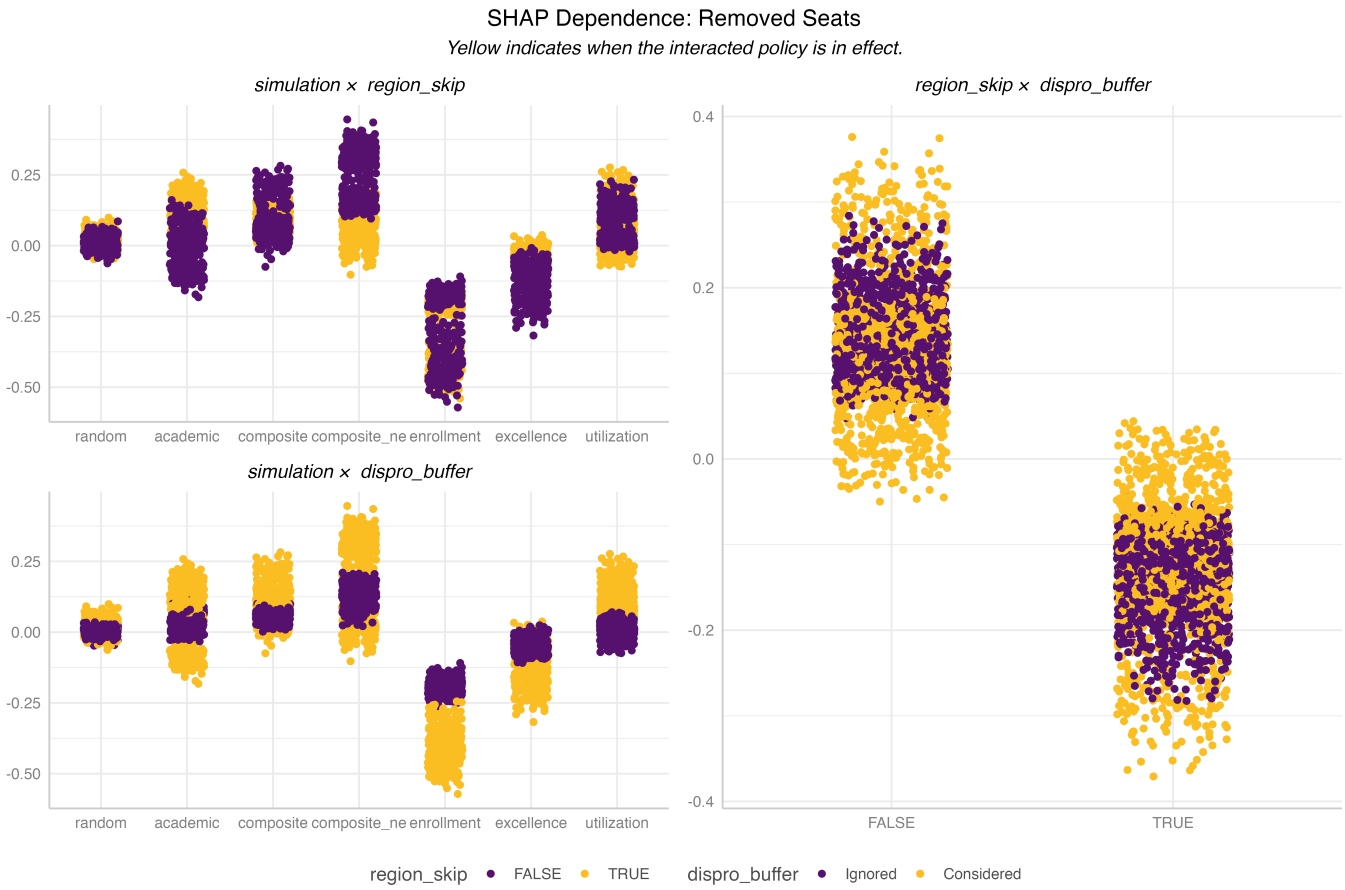

Figure B.3: **Interactive Effects of Design Principles on Scenario Seat Removal.** This figure displays SHAP dependence plots demonstrating how equity constraints (dispro_buffer) and regional boundaries (region_skip) moderate the influence of school ordering mechanisms (simulation) on the operational goal of removing excess seats. The y-axis represents the SHAP value, indicating a feature's contribution to increasing (positive) or decreasing (negative) the number of removed seats. Similar to the moderating effects observed on cost savings, activating the disproportionality safeguard (Yellow: Considered) acts as an operational expansionist. It visually expands the SHAP values across the effects of the sorting algorithms, drastically increasing the variance and increasing the large capacity reductions that would be possible under unconstrained orderings like utilization or academic (Purple: Ignored).

Table B.1: Simulation Order Metrics

Metric	Definition	Calculation/Notes	Source
Academic Performance - How students perform academically in core subject areas currently and over time. (*Excellence*)	Average state assessments scores on English Language Arts and Math performance and growth.	Measured as the average student performance on statewide assessments and average year-over-year growth on statewide assessments. Each measure was first standardized then the resultant combined metric was then standardized.	State-provided school dashboard and district administrative files.
School Culture and Climate - A school community's perception of belonging, safety, and academic learning. (*Excellence*)	The percentage of families, staff, and students responding favorably to survey questions about a sense of belonging, safety, or academic support for learning.	Computed as the percentage of respondents who agreed or strongly agreed with positive statements on the school culture and climate survey, averaged across the student, staff, and family surveys. Any school with a missing value was assigned a value equal to the district average. * District administrative files.	
Socio-Emotional Development - Skills such as social awareness, self-management, growth mindset, and self-efficacy. (*Excellence*)	The percentage of students responding favorably to survey questions related to social awareness, self-management, growth mindset, or self-efficacy.	Measured as the average percent of respondents who agreed or strongly agreed with positive evaluations for growth mindset, self-efficacy, self-management, and social awareness. Any school with a missing value was assigned a value equal to the district average.	District administrative files.
Family Choice and Demand for the School - The level of demand for a school from families as shown by school choices listed on a school application. (*Effective Resource Use*)	The percentage of applicants ranking the school as one of their top three choices in their school application.	Scores were calculated as the percentage of applicants who rated the school in top three.	
Teacher Turnover - How often teachers leave their jobs and are replaced by others. (*Effective Resource Use*)	The percentage of teachers who leave a school.	Teacher turnover was averaged across school years 2023-23 and 2023-24.	District administrative files.
Student Enrollment - Student enrollment refers to the number of students attending a school as a percentage of the school's ideal capacity. (*Effective Resource Use*)	A school's 2023-2024 school year enrollment compared to its ideal enrollment.	Calculated as the average capacity of each school minus the total enrollment.	
Building Condition - The condition of a school facility including its systems (e.g., heating, ventilation, and air conditioning, electrical, plumbing, roof, etc.), the interior and exterior of each building, and open space. (*Effective Resource Use*)	The school building's facility condition index (FCI) score. The FCI is an aggregate measure of the condition of all individual systems in a given facility. A lower FCI score indicates better school building conditions.	Schools with multiple locations were assigned the average FCI score for each of its locations.	(Data from VFA Facility Condition Assessment and District administrative files.
School Access - The availability of schools in a neighborhood. (*Equity*)	Average driving distance between the three closest schools with the same grade span.	School access was measured as the distance between a given school and the three nearest schools in the same grade span adjusted by the transition grade student population density of the focal school's zip code.	Addresses provided by District Administrative Files were used in Google's Maps API (Google, 2023) to generate the estimated driving time between all candidate/receiving pairings described in Section 3.
Program Access - The availability of educational programs in a school. (*Equity*)	Percentage of students in each school participating in Language programs, Special Education programs, or Career Technical Education and Pathway programs.	Program access was measured as the unduplicated counts of ELL students, special education students, Language Pathways students, AVID students, and low-SES students.	District administrative files.
Historical Inequities - Challenges and disparities rooted in a school's or community's history that affect educational opportunities today. (*Equity*)	The average amount of historical neighborhood opportunity experienced by students in each school. This measure is gathered from the Opportunity Insight Lab's "upward mobility index," defined in terms of the eventual earnings of children who grew up in households in in the 25th percentile of income distribution Chetty et al. (2018).	Student home addresses were geocoded then merged with "upward mobility" data from the opportunity atlas. Scores for schools were then computed as the average amount of opportunity experienced by students in each school. This measure captures historical conditions that lead to opportunity.	Student home addresses gathered from SFUSD administrative files combined with neighborhood data from opportunityatlas.org.

Note: The "Composite" score order metric is calculated as $C = \sum_{i=1}^{10} w_i x_i$, where w_i represents the community-assigned weight for metric x_i and i is the row in this table. The "Composite (No Equity)" score included all but the last 3 rows and is calculated similarly.