# COVERAGE IS NOT FAIRNESS: AUDITING LEARNED CONTROLS IN PARTICIPATORY BUDGETING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

In participatory budgeting, spending-based group fairness scores can report a perfect result even when nearly all voters receive no funded project they approved. We examine one such measure, cumulative share deficit (CSD), alongside voter coverage, and construct a valid allocation with CSD 0.0000 that excludes 99.0% of voters. Next, we hold the weights and feature definitions of learned project scores fixed and replay them through four allocation rules. At the unpenalized score, iterative payment reduces exclusion from 0.1470 to 0.0530 and increases welfare to 1.74× the direct-score value. Adding approval-count completion brings exclusion to 0.0427 and welfare to 1.99×. Neither payment rule produces a statistically detectable CSD improvement. On 18 held-out Warsaw series, learned endowments reduce CSD from 0.1992 to 0.1736, but a post-lock, four-parameter age table fitted on Warsaw training data has lower mean CSD (0.1624). Two locked external tests, one of learned priority and one of endowments on 32 projected-support series in two new cities, do not confirm transfer (decision: `falsified`). Finally, a post-hoc replay without refitting finds no detectable series-level difference on 58 native-approval series and unchanged allocations on 36. These results support reporting spending-based fairness alongside voter coverage, and learned interventions alongside the allocation changes they produce.

## 1 INTRODUCTION

Learning can enter a participatory budgeting (PB) procedure at several points, even when the fairness objective stays the same. A model can change voter endowments, reorder affordable projects, or score projects for direct budget fill. Where such a learned intervention enters the procedure determines whether it changes the funded set and which allocations it can produce.

Under the Method of Equal Shares (MES), learned endowments and project priorities retain approver-funded payments during the payment phase. Direct scoring removes this requirement. Small changes to endowments or priorities may leave the funded set unchanged. The three classes also learn different objects from different features, so comparing their fitted outcomes does not isolate the effect of payment.

We first decompose cumulative share deficit (CSD) and construct an allocation with zero CSD that excludes almost all voters. Next, we fix the score weights and feature definitions while changing the allocation rule. Each rule generates its own history, so this replay measures the rule change together with the history it induces. We then fit policies in Warsaw and test them elsewhere. Learned project priority fails its three-city test. Endowment estimates on projected support sets in two new cities are favorable in direction but fail the locked criteria. On 58 native-approval series, the fixed Warsaw endowment map shows no detectable series-level difference from Equal Shares and produces identical allocations on 36. No external study confirms transfer. Audits should report coverage and allocation changes, not CSD alone.

**Contributions.**

- We show that CSD can be zero while exclusion approaches one: the metric does not count distinct voters served. In a replay with fixed score weights and feature definitions, iterative payment reduces exclusion and raises welfare at the unpenalized score, and approval-count



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

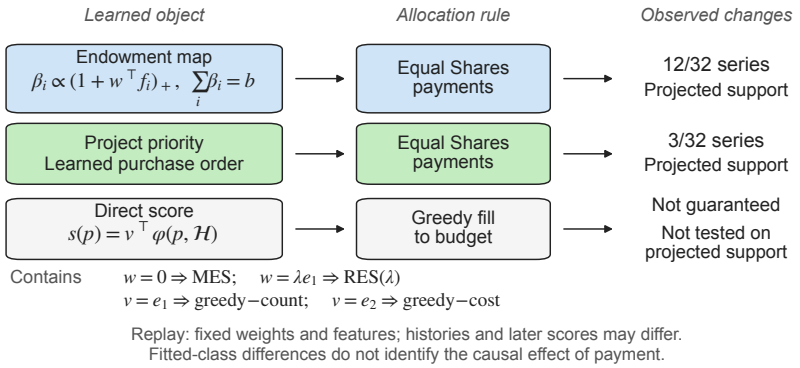



Figure 1: Where the three learned interventions enter the allocation process. Endowment learning changes voter budgets inside Equal Shares, and project-priority learning changes the order of affordable purchases. Both retain Equal Shares payments but change allocations on only 12 and 3, respectively, of the 32 projected-support series. Direct scoring fills the budget without the payment rule and was not evaluated on that corpus. Differences between fitted classes do not identify a causal effect. In the controlled replay, score weights and feature definitions stay fixed, but later scores can differ because each rule generates its own deficit history.

completion adds coverage. The CSD changes are not statistically detectable. The coverage gains weaken or reverse at welfare-regularized scores.

- The fixed endowment map reproduces the unmodified Equal Shares allocation on most external series. These series have fewer projects on average. That association does not establish that all endowment changes would leave their allocations unchanged.
- The Warsaw results do not establish a benefit from context. A post-lock static age table reaches lower CSD. Both locked external studies fail, and a post-hoc native replay without refitting finds no detectable series-level difference.

## 2 Problem setting and learned interventions

A *repeated participatory budgeting instance* consists of elections $I_1, \ldots, I_T$ and a fixed set of groups $G$. Each election is

$$I_t \;=\; (P_t,\; c_t,\; b_t,\; N_t,\; A_t)$$

where $P_t$ contains projects with costs $c_t : P_t \to \mathbb{R}_{>0}$, $b_t$ is the budget, $N_t$ is the voter set, and $A_t(i) \subseteq P_t$ is voter $i$'s approval ballot. A voter with recorded demographics belongs to a group $g(i) \in G$. We write $N_t^{\mathrm{cov}} \subseteq N_t$ for these voters. Ballots without demographic coverage still affect the allocation, but not the group ledger. The groups persist across elections even though the individual voters may change. An outcome sequence $W_1, \ldots, W_T$ satisfies $c_t(W_t) \le b_t$. In this corpus, $G$ is the set of recorded age-bracket $\times$ sex cells.

**Definition 1** (Attribution and deficit)**.** *In edition $t$, group $g$ receives $u_t(g) = \sum_{p \in W_t:\ n_t^{\mathrm{cov}}(p) > 0} c_t(p)\, n_{t,g}(p)/n_t^{\mathrm{cov}}(p)$, where $n_t^{\mathrm{cov}}(p)$ counts covered approvers of $p$ and $n_{t,g}(p)$ those in $g$; its entitlement is $e_t(g) = c_t(W_t)\,|g_t|/|N_t^{\mathrm{cov}}|$. The* cumulative share deficit *after $T$ editions is*

$$\mathrm{CSD}_T(g) \;=\; \frac{\sum_{t \le T} \big(e_t(g) - u_t(g)\big)}{\sum_{t \le T} e_t(g)}.$$

We attribute spending only for projects with at least one approver whose demographics are recorded, but entitlement includes all funded projects. A funded project with no covered approver would therefore create an artificial deficit for every group. We found no such project across 54 scored elections and 4 policies. The largest budget share spent on such projects is 0.00%, so this attribution issue does not affect the reported comparisons.

Entitlement uses the amount spent rather than the nominal budget. This separates underspending from unequal distribution. If every funded project has a covered approver, the unnormalized gaps $e_t(g) - u_t(g)$ sum to zero across groups. For each series, we accumulate CSD over its scored elections and take the largest group value. We then average these series-level maxima with equal weight. A policy carries its own history through $\mathcal{H}_{t-1}$, but the reported ledger begins with the scored period. Exclusion is also averaged equally across series, whereas welfare is summed.

CSD measures how spending is distributed; it is not a standard PB axiom. It combines the share attribution used by Equal Shares (Peters & Skowron, 2020) with a longitudinal equality-of-resources ledger (Maly et al., 2023). The ledger compares cumulative group shares but does not require an allocation to serve many distinct voters. Proposition 1 states this limitation exactly.

We report two additional outcomes. *Welfare* counts (voter, funded project) approval pairs, or equivalently the number of approved funded projects summed over voters. The *exclusion rate* is the share of voters who receive none of their approved projects. Neither enters the objective except through the soft welfare target in Section 2.1.

### 2.1 THREE LEARNED INTERVENTION CLASSES

Figure 1 shows where each learned object enters the allocation process. Appendix C gives the feature formulas and optimizer details.

**Learned endowments (payment retained).** Equal Shares purchases projects at the smallest price $\rho$ for which their approvers can cover the cost. Approver $i$ pays $\min(\beta_i, \rho)$ from a personal endowment. Standard MES uses uniform $\beta_i = b_t/|N_t|$. The learned variant sets $\beta_i = (b_t/|N_t|)\big(1 + w^\top f(g(i), \mathcal{H}_{t-1})\big)_+$, and rescales the values to sum to $b_t$, falling back to uniform endowments if all values are zero. A voter without recorded demographics receives the uniform value before the common rescaling. The six features record the group's carried deficit per capita, deficit normalized by cumulative entitlement, size relative to an equal split, mean ballot length, mean cost per approval, and approval overlap. The last three are expressed as deviations. Clamping, rescaling, and the fallback ensure nonnegative endowments summing to the budget. Each reported MES-family outcome uses the same approval-count completion (Algorithm 1). Payment-phase winners are paid for entirely by their approvers. Completion is outside this payment accounting.

**Learned project priorities (payment retained).** This class changes which affordable project is bought next, while retaining Equal Shares balances and approver-funded payments. Setting every learned utility to zero reproduces ordinary MES exactly. The learned priority changes the order of purchases, not the payment rule. In the fresh-city study, it changes allocations on only 3 series.

**Learned direct scores (payment discarded).** We score each project as $s(p) = v^\top \phi(p, \mathcal{H}_{t-1})$ and fund projects in descending score order subject to the budget. Ties are broken by ascending cost and then project identifier. The five features are approval share, approvals per unit cost, cost share, a deficit-weighted share of $p$'s covered approvers, and the concentration of $p$'s support across groups. This class has no proportionality guarantee. It is a direct-selection comparison, not a one-factor ablation of the payment rule.

**Warsaw fitting and the soft welfare target.** We fit the endowment and direct-score classes on the training elections with covariance matrix adaptation evolution strategy (CMA-ES). The initial means are RES$(1)$ for endowments and greedy approvals-per-cost for direct scores. These vectors initialize the search but do not automatically remain among the fitted candidates. We replay and report every baseline separately. For each class, target, and seed, one weight vector is optimized jointly across the training series and frozen before held-out evaluation. The loss is the equal-weighted mean of the series-level worst-group CSD plus $\mu \max(0, \tau - \text{welfare}/\text{welfare}_{\text{MES}})$. This penalty does not impose a hard welfare constraint. It is zero for any candidate whose welfare ratio meets the target. Once the target exceeds every attainable welfare ratio, increasing it further adds the same offset to every candidate, so their ranking no longer changes (Appendix E).

### 2.2 CONTAINMENT OF DEPLOYED RULES

Each learned class can reproduce different hand-designed rules. We check the following identities by running the policies:

$$w = \mathbf{0} \;\Rightarrow\; \text{MES}, \qquad w = \lambda e_1 \;\Rightarrow\; \text{RES}(\lambda), \qquad v = e_1 \;\Rightarrow\; \text{greedy by approval count.}$$

The first two identities hold with zero numerical error on all 19 series for $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$. The direct class does not contain MES. Appendix Table 16 lists every tested RES intensity. These values concern only the endowment class. In the direct class, $v = e_1$ and $v = e_2$ recover greedy approval count and greedy approvals per cost, with exact winner-set equality on all 132 elections. The project-priority code also reproduces ordinary MES at zero project utility on every multi-city election in the locked study.

The direct approval-count check found a counting error. The learned score had counted only ballots with demographic coverage, whereas the deployed rule counts every ballot. This difference reordered projects in four elections.

## 3 WHY LOW CSD CAN COEXIST WITH EXCLUSION

We first ask what a low CSD score tells us about voter coverage. We show that CSD can be zero while almost no voter is served, then derive the support condition imposed by the Equal Shares payment phase. To examine how the allocation rule affects coverage, we replay one fixed score function through different rules. Section 4 compares the fitted classes separately. Those classes use different objects and features, so their differences cannot isolate the effect of approver-funded payment.

**Proposition 1** (Representative-support degeneracy)**.** *Assume every group and every funded project's covered-approver set is nonempty. Let $s_{t,g} = |g_t|/|N_t^{\mathrm{cov}}|$, let $q_{t,g}(p) = n_{t,g}(p)/n_t^{\mathrm{cov}}(p)$ be group $g$'s share among project $p$'s covered approvers, and let $E_g = \sum_t c_t(W_t)s_{t,g} > 0$. Then*

$$\mathrm{CSD}_T(g) = 1 - \sum_t \sum_{p \in W_t} \underbrace{\frac{c_t(p)s_{t,g}}{E_g}}_{\text{weights sum to } 1} \frac{q_{t,g}(p)}{s_{t,g}}.$$

*If $|q_{t,g}(p)/s_{t,g} - 1| \le \epsilon$ for every funded project and group, then $|\mathrm{CSD}_T(g)| \le \epsilon$. Yet for any positive rational group shares and $\delta > 0$, an instance exists with $\mathrm{CSD}(g) = 0$ for every group and exclusion at least $1 - \delta$.*

The construction funds one project that costs the entire budget and whose approvers match the electorate's group shares. All other voters approve only an unfunded dummy project. With 8 supporters among 800 voters, the implemented ledger reports CSD 0.0000 and exclusion 99.0%. Spending shares match entitlements exactly, yet almost everyone receives nothing. This construction demonstrates a limitation of CSD; it is not an outcome of a fitted policy. Appendix H provides the proof and executable checks.

The learned endowment policies do not approach this extreme in the Warsaw data. Their payment phase must satisfy the following support condition. Greedy completion is not subject to it.

**Proposition 2** (Support floor)**.** *Let $W_{\mathrm{pay}}$ be the projects bought during the Equal Shares payment phase on instance $(P, c, b, N, A)$ under endowments $\beta \ge 0$ with $\sum_i \beta_i = b$. Then every $p \in W_{\mathrm{pay}}$ satisfies $\sum_{i \in A(p)} \beta_i \ge c(p)$. Consequently, if $\beta_i \le \kappa \cdot b/|N|$ for all $i$, then*

$$\frac{|A(p)|}{|N|} \;\ge\; \frac{c(p)}{\kappa\, b} \qquad \textit{for every } p \in W_{\mathrm{pay}}.$$

Appendix G proves the proposition.

During the payment phase, more expensive projects require proportionally more approvers. Raising the endowment cap weakens this support floor linearly in $\kappa$. Direct score-and-fill has no corresponding restriction.

The fitted map's per-series cap ratio averages 7.40 and reaches 13.63. Its support floor is therefore much weaker than the fixed $\kappa = 1$ reference in Table 9. A large $\kappa$ leaves little restriction, which is why we report the observed cap.

The unpenalized direct-score policy does not systematically choose the least approved projects. It chooses fewer, larger projects: 11.0 per election, compared with 30.1 under Equal Shares. Their supporters more closely match the electorate, with cost-weighted total-variation distance 0.0664 versus 0.0773. This match keeps attribution close to entitlement even as exclusion rises from 0.0438 to 0.1470.

Proposition 2 lets us measure the share of spending on projects that the uniform-endowment payment phase could not buy at $\kappa = 1$. That share is 0.29% for Equal Shares and 0.64% for learned endowments, compared with 25.06% for the unpenalized direct-score policy. For uniform Equal Shares, such spending comes from completion. Learned endowments can exceed the uniform cap, so their value does not identify the allocation phase. Table 9 reports all tested rules.

### 3.1 Frozen-score replay through alternative allocation rules

Retraining direct scores under the static floor from Proposition 2 closes 59% of the exclusion gap to Equal Shares at $\kappa = 1$, but only 27% of the welfare gap. The floor reduces the observed coverage damage without reproducing the learned-endowment outcomes. Loosening the cap to $\kappa = 2$ worsens both welfare and exclusion (Table 7).

In a separate exploratory replay, we keep the learned object fixed. We hold the unpenalized policy's score weights and feature definitions constant and apply four allocation rules (Table 11; Appendix Figure 5). Direct greedy fill gives CSD 0.1228, welfare 586,114, and exclusion 0.1470. A one-shot $\kappa = 1$ filter lowers exclusion to 0.0935. Iterative payment with uniform voter balances lowers it to 0.0530 and multiplies welfare by 1.74×. Adding approval-count completion gives exclusion 0.0427 and welfare 1.99× the direct value. Relative to direct fill, the completed intervention lowers exclusion by 0.1043 (95% CI [-0.1313, -0.0773], $p < 0.001$, 17/18 series). CSD changes by 0.0054 (95% CI [-0.0227, +0.0341], $p = 0.721$).

This replay differs from standard Mes. It orders affordable projects by the fixed learned score rather than minimum price. Later scores depend on the deficit history generated by each rule's earlier allocations. Thus, fixing the score function does not fix the sequence of score values. At the unpenalized score, the one-shot floor reduces exclusion less than iterative payment, but their difference includes the histories induced by the two rules.

**The payment effect depends on the fitted score.** Figure 2 and Table 11 show that, at the unpenalized score, iterative payment lowers exclusion to 0.0530 and raises welfare 1.74×. Approval-count completion further lowers exclusion to 0.0427 and raises welfare 1.99×. Neither CSD contrast is statistically detectable. At both welfare-regularized scores, payment alone instead lowers welfare and raises exclusion; completion attenuates the coverage loss but does not consistently raise welfare. The corresponding payment-based CSD contrasts worsen. The replay therefore explains one score's coverage failure. It does not establish that payment protects coverage or fairness in general.

Proposition 1 explains why the outcomes can move in opposite directions. CSD does not record how many distinct voters are served, so the objective cannot reward the coverage change directly. The fitted sweeps span 0.0140 and 0.0717 (Appendix Figure 4), but those widths depend on this corpus, class, and optimizer. Only the frozen-score replay holds score weights and feature definitions fixed while changing the allocation rule.

We next compare the fitted classes on this corpus and ask how much of their performance requires the learned intervention.

## 4 What the Warsaw fit changes

The four hand-designed rules cover a narrow range of held-out outcomes (Appendix Figure 4). Both learned classes produce results outside that range. The unpenalized direct-score fit is not the best observed endpoint: a nearby regularized fit has both lower CSD and higher welfare.

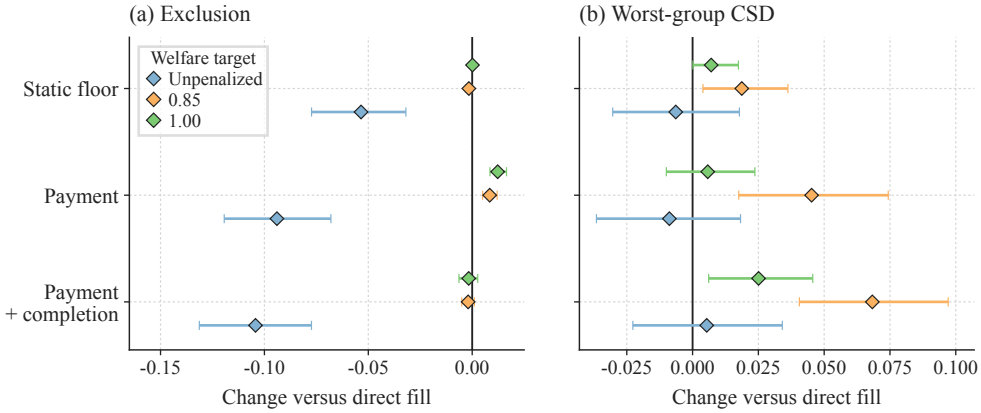



Figure 2: Paired effects of changing the allocation rule, relative to direct greedy fill, on 18 held-out Warsaw series. Score weights and feature definitions stay fixed, but each rule generates its own history and therefore may produce different later scores. Points are mean paired differences and bars are 95% paired percentile-bootstrap intervals. **(a)** At the unpenalized score, every alternative rule lowers exclusion and the intervals exclude zero. At the regularized scores, these effects weaken or reverse. **(b)** No kernel yields a statistically detectable improvement in worst-group deficit; the regularized contrasts mostly worsen it. The unpenalized replay shows a coverage gain that CSD does not directly measure.

| Rule | Mean worst-group CSD ↓ | Welfare ↑ | Exclusion ↓ |
|---|---|---|---|
| Recorded outcome | 0.2189 | 917,379 | 0.0830 |
| Greedy (approval count) | 0.2173 | 918,508 | 0.0813 |
| Greedy (approvals/cost) | 0.1991 | 1,319,325 | 0.0448 |
| MES | 0.1992 | 1,293,610 | 0.0438 |
| RES ($\lambda$=1) | 0.1951 | 1,286,793 | 0.0436 |
| LLMRule (published count score) | 0.2007 | 1,291,397 | 0.0466 |
| *Endowment family (mechanism retained)* | | | |
| Learned endowment map, no welfare penalty | 0.1736 | 1,273,257 | 0.0414 |
| *Static demographic reference (exploratory, post-lock)* | | | |
| Four-bracket age lookup | 0.1624 | 1,261,342 | 0.0452 |
| *Outcome family (mechanism discarded)* | | | |
| Learned, soft target 1.00 | 0.1728 | 1,303,057 | 0.0447 |
| Learned, soft target 0.85 | 0.1153 | 1,198,311 | 0.0450 |
| Learned, no welfare penalty | 0.1228 | 586,114 | 0.1470 |

Table 1: Held-out results. The recorded row replays public winner sets; all other rows are executable counterfactuals. The four reported hand-designed rules (greedy approvals/cost, MES, and RES($\lambda$=0.5, 1.0)) span $[0.1951, 0.1992]$. We replay LLMRule (Thach et al., 2026) without refitting. CSD and exclusion are series-weighted; welfare is summed. The age lookup is a post-lock exploratory audit, not a confirmatory or cross-city result.

### 4.1 REGULARIZATION AND HETEROGENEITY

The unpenalized direct-score fit has the largest train–test gap: +0.0531, compared with +0.0057 at soft target 1.00. The target-1.00 fit reduces mean worst-group CSD by 0.0264 (95% CI [-0.0462, -0.0076], raw $p = 0.013$; Holm $p = 0.023$ in the 3-contrast direct-score/endowment/senior-tilt family). Targets 1.02 and 1.05 show no detectable effect. Seed spread is 0.0040 at the headline target and grows toward the aggressive end, as the seed band of Appendix Figure 4 and Appendix Table 17 both show. The full sweep and the secondary family of related tests appear in Appendices E and I.

**Ablation and district variation.** Removing cost share, approvals per cost, or carried deficit worsens the direct-score model in the single-refit comparison. Removing cohort concentration improves CSD by 0.0095. The model may be exploiting this proxy, but the ablation does not identify a single cause. When we rescore the same allocations with age-only or sex-only groups, both learned policies remain below MES ($p \leq 0.0130$). These are dependent, secondary exploratory evaluations, not new fits. The direct-score policy loses on 4 districts (Ochota, Żoliborz, Wilanów, Białołęka), and the endowment average also reflects larger wins than losses. Appendix tables report every ablation, partition, and district result.

## 5 LEARNED ENDOWMENTS ON HELD-OUT WARSAW ELECTIONS

The hand-designed RES rollover rule adjusts endowments with one scalar. Its CSD barely differs from Equal Shares (0.1951 versus 0.1992). This result concerns that particular parametrization; it does not establish that changing endowments is ineffective.

The fitted endowments reach 0.1736, compared with 0.1992 for Equal Shares. The reduction is 0.0255 (95% CI [-0.0453, -0.0081], raw $p$ = 0.012; Holm-adjusted $p$ = 0.023, 11/18 series). Welfare falls by 1.6%, while exclusion falls from 0.0438 to 0.0414. The policy retains the approver-funded payment phase and its charge accounting because it belongs to the Equal Shares family by construction. We do not attribute standard priceability or an unweighted proportionality axiom to the nonuniform endowments or greedy completion.

The published LLMRule count score reaches CSD 0.2007 without refitting. The endowment map's CSD is lower by 0.0271 (secondary exploratory $p$ = 0.002).

**Direct scoring has lower CSD at a common soft target.** The main table compares an unpenalized endowment fit with a welfare-regularized direct-score fit. These fits use different objectives. In a separate frontier sweep, we give both classes the same soft target, positive penalty, and number of generations. Their dimensions, evaluation counts, learned objects, and features still differ. At target 1.00 for both, learned endowments reach 0.1876 against 0.1728 for direct scores, a paired difference of +0.0148 in favor of direct scoring (95% interval [-0.0004, 0.0298], secondary exploratory $p$ = 0.080, 5/18 series). Only exclusion favors the payment-based class, 0.0433 against 0.0447. This common-target comparison favors direct scoring on CSD, but it does not identify the effect of payment because the two policy classes remain structurally different.

**Static demographic controls reach lower mean CSD.** The senior tilt reaches 0.1841 (raw $p$ = 0.007; Holm $p$ = 0.022), the six-feature map 0.1736, and a history-free version 0.1693. These means do not establish a benefit from temporal memory. We also fit a four-parameter age lookup on the training split in a post-lock exploratory analysis. It reaches 0.1624. In the corrected post-hoc replay, the six-feature contextual model's CSD minus the lookup's CSD is +0.0112 (95% CI [-0.0105, +0.0301], unadjusted exploratory $p$ = 0.311, 5/1/12 wins/ties/losses), so we find no statistically detectable series-level difference between the policies.

The lookup has higher exclusion and lower welfare. Exclusion is 0.0452 versus 0.0414 for the learned map and 0.0438 for MES; welfare is 1,261,342 (2.5% below MES). Multipliers [0.0564, 0.0528, 0.0895, 20.0855] leave almost no endowment for the younger brackets and concentrate it on the oldest. We use this lookup as an exploratory reference.

Across 3 six-feature fits, held-out CSD ranges from 0.1729–0.1738. In the disjoint-series check, the learned map reaches 0.1941 versus 0.2242 for MES (17/19 wins). However, 18 of the 19 series are Warsaw districts and the overlapping training folds are not independent replications. Appendix Tables 19 and 20 report all folds and the coefficient-box sensitivity. Appendix K contains the no-refit non-Warsaw stress set.

## 6 EXTERNAL TRANSFER TESTS

We locked two studies before opening their held-out ballots: learned priority and endowments evaluated on projected support sets. Each lock fixes the corpus, fitted weights, rule, and decision thresh-

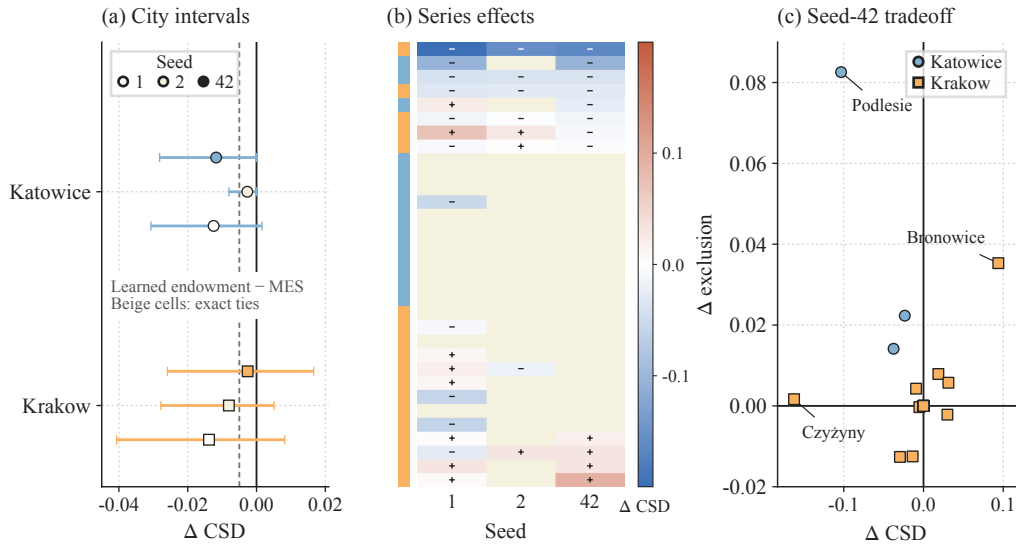



Figure 3: Fresh-city results over projected support sets ($n = 32$ series). The endowment map changes allocations on 12/32 series. **(a)** Endowment-minus-MES city means use 95% paired bootstrap intervals. The three fixed-seed city-macro effects are -0.0131, -0.0054, and -0.0072. At the primary seed, the Holm-adjusted $p$ values are 0.5000 for Katowice and 0.8398 for Krakow. **(b)** Per-series effects use beige for exact ties, $+/-$ for sign, and the side strip for city. **(c)** Primary-seed CSD and exclusion changes are descriptive, not causal: negative CSD change favors learned endowments, while positive exclusion change is adverse. Projection discards cumulative point magnitude and ordinal rank.

olds. The opening receipts record the lock digests. The native replay below is a separate post-hoc analysis. Appendix L gives the artifacts and conditions.

The corrected post-hoc replay of the first locked study, learned priority inside Equal Shares, remains *falsified*. The temporal city-macro CSD contrast misses the threshold, some city estimates do not point downward, and priorities change fewer allocations than matched endowments. Welfare and exclusion safeguards pass, but the study finds no stable fairness benefit.

The second test applies the fixed learned endowments in two new cities. We convert cumulative and ordinal ballots to projected support sets, keeping the listed projects but discarding point magnitudes and ranks. The lock covers 228 elections across 32 series; 96 elections are scored. It uses three fixed seeds $1, 2, 42$ and equal-city aggregation.

The endowment map changes an allocation on only 12/32 series. 20 series reproduce Equal Shares exactly. Conditional on an allocation change, the primary-seed CSD contrast is -0.0175, compared with -0.0066 across all series (8 wins and 4 losses). The corresponding conditional effects over all seeds lie in [-0.0228, -0.0175]. Because this subset is defined by the observed outcome, the comparison is descriptive and does not alter the locked test. On most fresh-city series, the fitted intervention leaves the allocation unchanged. These exact ties also help explain why the Katowice interval reaches 0.0000.

Across all series, endowment-minus-MES estimates are negative in both cities for every seed. Equal-city macro changes are -0.0131, -0.0054, and -0.0072. The primary result contains 8 wins, 20 exact ties, and 4 losses. The all-series effect is directionally favorable but seed-sensitive.

The lock evaluates eight conditions. Direction, macro-effect, welfare, and actuation pass. The failures are city-level uncertainty, primary Holm significance, the exclusion safeguard, and seed spread. The primary intervals extend to 0.0000 in Katowice and 0.0166 in Krakow, while the Holm-adjusted $p$ values are 0.5000 and 0.8398. Katowice exclusion increases by +0.0078 and +0.0085 at seeds 1 and 42. The locked decision is therefore falsified. The small number of allocation changes helps explain the uncertainty, but it does not establish successful generalization.

**Native approval ballots, without projection.** Gdynia and Łódź provide a post-hoc, no-refit replay on 58 native-approval series. The frozen Warsaw fit gives 0.1345 versus 0.1384 (difference -0.0039; 95% paired interval [-0.0114, 0.0031], $p = 0.316$, 12/58 series); neither city differs detectably from zero. The map exactly matches Equal Shares on 36 series, whose median is 6.2 projects per scored election versus 22.2 when it acts. The estimated effect remains near zero across the reported granularity grid (Table 21). Thresholds change sample size and composition, so this grid does not isolate the role of low actuation.

Projection changes ballot lengths and approval overlap, both of which are map features. The native inert rate (36/58) nearly matches the projected study (20/32); median project counts are 18.0 under projection and 47.5 in Warsaw. Series with fewer projects are more likely to have unchanged allocations. Other differences between cities remain unresolved.

## 7 Related work

Research on PB studies proportionality and repeated representation (Aziz et al., 2015; Peters & Skowron, 2020; Lackner, 2020; Elkind et al., 2023). Learned systems fit project-scoring aggregation rules (Fairstein et al., 2024b) or evolve executable priority programs (Thach et al., 2026); mixed rules alter MES budgets after project selection (Baychkov et al., 2026b). We learn endowments before project selection and, in a separate replay, fix the project-score parameters while changing the allocation rule. We do not claim to introduce learned PB rules or nonuniform MES budgets. Our audit also connects to mechanism-informed learning (Maruo et al., 2026) and tests of reward hacking (Skalse et al., 2022); Appendix J details the PB comparison.

## 8 Limitations

**Evidence scope.** The corrected Warsaw replay provides within-city evidence, not a fresh holdout test. The projected-support study covers 32 series in two new cities, but changes few allocations and has seed sensitivity, city-level uncertainty, and increases in exclusion. It does not establish transfer to native ballots. The post-hoc native replay finds no detectable series-level difference and leaves allocations unchanged on 36 series. All external cities are Polish. Both locked studies fail, and the native replay is not confirmatory. The post-lock age lookup also prevents us from attributing the Warsaw reduction to learned context.

**Identification and deployment limits.** The frontier compares fitted classes with different learned objects and features. It cannot isolate the causal effect of the payment rule. Weight vectors are not identifiable, so we do not interpret individual coefficients. Direct scoring has no proportionality axiom. The frozen-score replay explains one Warsaw coverage failure, not a general safeguard. Fixed ballots also cannot reveal voters' strategic responses. Administrative age-by-sex groups describe voters incompletely. Changing the partition changes the objective, and demographic missingness may be nonrandom (Appendix B).

## 9 Conclusion

CSD measures group spending shares, not the number of voters served. At the fixed unpenalized score, payment reduces exclusion from 0.1470 to 0.0530 and raises welfare to 1.74× the direct-score value. Completion brings exclusion to 0.0427 and welfare to 1.99×. Neither CSD contrast is statistically detectable at this score, and both worsen CSD at the two regularized targets. Reporting coverage alongside CSD makes this difference visible.

In Warsaw, the fixed endowment map lowers CSD but costs 1.6% in welfare. A post-lock age lookup reaches CSD 0.1624 with worse exclusion and welfare. Both locked studies fail, including the projected-support city-level uncertainty criterion. The post-hoc native replay finds no detectable difference and identical allocations on 36/58 series. Neither a benefit from learned context nor transfer is established. Report objective value, coverage, and allocation changes separately. Replay fixed score parameters across allocation rules to help distinguish the score's contribution from the rule's.

## REPRODUCIBILITY STATEMENT

Manuscript macros and tables are checked against authenticated stored results. The primary Warsaw results use the corrected v2 fit and evaluation artifacts; the post-lock contextual comparison uses a separately recorded paired audit. These are corrected post-hoc replays, not fresh untouched holdout tests. Appendix B describes the corrections and retained evidence. The statistical figures use the same stored outputs. The public corpus (Stolicki et al., 2020) is split and frozen before fitting. A 415-file integrity manifest pins the frozen Warsaw, multi-city, and native cross-city corpus files to official Pabulib commit `2f4321f` (the artifact records the full hash), and every file matches the recorded size and SHA-256 digest. Both locked studies and the separate native replay write a manifest for their held-out corpus; only the first two have pre-opening protocol locks.

Run JSONs record every fitting seed. Analysis scripts or companion manifests fix the resampling seeds. The optimizer is a 120-line CMA-ES implementation that depends only on NumPy. None of the experiments requires a GPU, network connection, or API key. Three verification scripts reproduce the reference replay and the endowment/direct-score containments (Section 2.2). One of these checks found the ballot-counting defect described above.

## ETHICS STATEMENT

This study reuses a public participatory-budgeting corpus and collects no new human-subject data. The fairness analysis uses age and sex fields released with the ballots. These administrative categories are incomplete descriptions of voters and poor substitutes for communities. We report only aggregate group outcomes and do not infer missing demographic attributes. The learned rules are research prototypes, not deployment recommendations.

The worst-served group is nearly fixed in this corpus. It is men aged 60 or older in 13/18 series, voters aged 60 or older in 14/18, and men of some age in 17/18. Minimizing the largest deficit therefore usually creates pressure to redirect spending toward older men in these data.

This pattern does not show that older men are disadvantaged more broadly. CSD compares the attributed spending attracted by a group's ballots with its share of turnout. Older voters in this corpus submit shorter ballots, which attract less attributed spending under the metric. Because the maximization usually selects the same demographic group, the objective also behaves more like a single-group deficit than the notation suggests. Any deployment would require an explicit local choice of groups and attribution rules. Our partition should not be inherited by default.

## AI USE STATEMENT

Generative AI tools were used during hypothesis refinement, study design, implementation, data preparation, result interpretation, proof checking, literature search, and manuscript editing. The authors verified the resulting claims against the code, stored outputs, proofs, and cited sources and accept responsibility for the submission.

## REFERENCES

Hugh Adams, Srijoni Majumdar, and Evangelos Pournaras. Fair compromises in participatory budgeting: A multi-agent deep reinforcement learning approach. In *Machine Learning, Optimization, and Data Science*, volume 16467 of *Lecture Notes in Computer Science*, pp. 50–65, 2026. doi: 10.1007/978-3-032-21477-5_4. URL https://link.springer.com/chapter/10.1007/978-3-032-21477-5_4.

Haris Aziz, Markus Brill, Vincent Conitzer, Edith Elkind, Rupert Freeman, and Toby Walsh. Justified representation in approval-based committee voting. In *Proceedings of AAAI 2015*, 2015. doi: 10.1609/aaai.v29i1.9324.

Haris Aziz, Barton E. Lee, and Nimrod Talmon. Proportionally representative participatory budgeting: Axioms and algorithms. In *Proceedings of AAMAS 2018*, 2018. doi: 10.65109/qhuq3213.

Anton Baychkov, Markus Brill, and Jannik Peters. Utilitarian guarantees for the method of equal shares. In *Proceedings of AAAI 2026*, 2026a. doi: 10.1609/aaai.v40i20.38707.

Anton Baychkov, Markus Brill, and Markus Utke. Mixed voting rules for participatory budgeting, 2026b. URL https://arxiv.org/abs/2606.23320.

Gerdus Benade, Swaprava Nath, Ariel Procaccia, and Nisarg Shah. Preference elicitation for participatory budgeting. In *Proceedings of AAAI 2017*, 2017. doi: 10.1609/aaai.v31i1.10563.

Niclas Boehmer, Piotr Faliszewski, Łukasz Janeczko, Andrzej Kaczmarczyk, Grzegorz Lisowski, Grzegorz Pierczyński, Simon Rey, Dariusz Stolicki, Stanisław Szufa, and Tomasz Wąs. Guide to numerical experiments on elections in computational social choice, 2024. URL https://arxiv.org/abs/2402.11765.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia. Introduction to computational social choice. In *Handbook of Computational Social Choice*. Cambridge University Press, 2016. doi: 10.1017/cbo9781107446984.002.

Markus Brill, Rupert Freeman, Svante Janson, and Martin Lackner. Phragmén's voting methods and justified representation. *Mathematical Programming*, 2023. doi: 10.1007/s10107-023-01926-8.

Laurent Bulteau, Noam Hazon, Rutvik Page, Ariel Rosenfeld, and Nimrod Talmon. Justified representation for perpetual voting. *IEEE Access*, 2021. doi: 10.1109/access.2021.3095087.

Yves Cabannes. Participatory budgeting: a significant contribution to participatory democracy. *Environment and Urbanization*, 2004. doi: 10.1177/095624780401600104.

Nikhil Chandak, Shashwat Goel, and Dominik Peters. Proportional aggregation of preferences for sequential decision making. *Journal of Artificial Intelligence Research*, 2026. doi: 10.1613/jair.1.18660.

Vincent Conitzer, Rupert Freeman, and Nisarg Shah. Fair public decision making. In *Proceedings of EC 2017*, 2017. doi: 10.1145/3033274.3085125.

Sankarshan Damle and Boi Faltings. LLMs for resource allocation: A participatory budgeting approach to inferring preferences, 2025. URL https://arxiv.org/abs/2508.06060.

Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David C. Parkes, and Sai S. Ravindranath. Optimal auctions through deep learning. *Communications of the ACM*, 2021. doi: 10.1145/3470442.

Edith Elkind, Svetlana Obraztsova, and Nicholas Teh. Temporal fairness in multiwinner voting, 2023. URL https://arxiv.org/abs/2312.04417.

Edith Elkind, Tzeh Yuan Neoh, and Nicholas Teh. Temporal elections: Welfare, strategyproofness, and proportionality, 2024. URL https://arxiv.org/abs/2408.13637.

Edith Elkind, Svetlana Obraztsova, Jannik Peters, and Nicholas Teh. Verifying proportionality in temporal voting, 2025. URL https://arxiv.org/abs/2502.05949.

Equal Shares team. Method of equal shares: real-world deployments in aarau and wieliczka. https://equalshares.net, 2023.

Brandon Fain, Ashish Goel, and Kamesh Munagala. The core of the participatory budgeting problem. In *Proceedings of WINE 2016*, 2016. doi: 10.1007/978-3-662-54110-4_27.

Roy Fairstein, Gerdus Benadè, and Kobi Gal. Participatory budgeting designs for the real world. In *Proceedings of AAAI 2023*, 2023. doi: 10.1609/aaai.v37i5.25699.

Roy Fairstein, Reshef Meir, and Kobi Gal. Proportional participatory budgeting with projects interaction, 2024a. URL https://arxiv.org/abs/2106.05360.

Roy Fairstein, Dan Vilenchik, and Kobi Gal. Learning aggregation rules in participatory budgeting: A data-driven approach, 2024b. URL https://arxiv.org/abs/2412.01864.

Piotr Faliszewski, Jarosław Flis, Dominik Peters, Grzegorz Pierczyński, Piotr Skowron, Dariusz Stolicki, Stanisław Szufa, and Nimrod Talmon. Participatory budgeting: Data, tools, and analysis, 2023. URL https://arxiv.org/abs/2305.11035.

Bailey Flanigan, Paul Gölz, Anupam Gupta, Brett Hennig, and Ariel D. Procaccia. Fair algorithms for selecting citizens' assemblies. *Nature*, 2021. doi: 10.1038/s41586-021-03788-6.

Rupert Freeman, Seyed Majid Zahedi, and Vincent Conitzer. Fair and efficient social choice in dynamic settings. In *Proceedings of IJCAI 2017*, 2017. doi: 10.24963/ijcai.2017/639.

Ashish Goel, Anilesh K. Krishnaswamy, Sukolsak Sakshuwong, and Tanja Aitamurto. Knapsack voting for participatory budgeting. *ACM Transactions on Economics and Computation*, 2019. doi: 10.1145/3340230.

Artur Gorokh, Siddhartha Banerjee, and Krishnamurthy Iyer. Near-efficient allocation using artificial currency in repeated settings. In *Proceedings of the 12th International Conference on Web and Internet Economics*, 2016. URL https://sidbanerjee.orie.cornell.edu/publication/gorokh-16-near/.

D Ellis Hershkowitz, Anson Kahng, Dominik Peters, and Ariel D. Procaccia. District-fair participatory budgeting, 2021. URL https://arxiv.org/abs/2102.06115.

Pallavi Jain, Krzysztof Sornat, and Nimrod Talmon. Participatory budgeting with project interactions. In *Proceedings of IJCAI 2020*, 2020. doi: 10.24963/ijcai.2020/54.

Nathanael Jo, Bill Tang, Kathryn Dullerud, Sina Aghaei, Eric Rice, and Phebe Vayanos. Fairness in contextual resource allocation systems: Metrics and incompatibility results. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11837–11846, 2023. doi: 10.1609/aaai.v37i10.26397. URL https://ojs.aaai.org/index.php/AAAI/article/view/26397.

Raphael Koster, Jan Balaguer, Andrea Tacchetti, Ari Weinstein, Tina Zhu, Oliver Hauser, Duncan Williams, Lucy Campbell-Gillingham, Phoebe Thacker, Matthew Botvinick, and Christopher Summerfield. Human-centred mechanism design with democratic ai. *Nature Human Behaviour*, 6:1398–1407, 2022. doi: 10.1038/s41562-022-01383-x.

Sonja Kraiczy and Edith Elkind. An adaptive and verifiably proportional method for participatory budgeting, 2023. URL https://arxiv.org/abs/2310.10215.

Martin Lackner. Perpetual voting: Fairness in long-term decision making. In *Proceedings of AAAI 2020*, 2020. doi: 10.1609/aaai.v34i02.5584.

Martin Lackner and Piotr Skowron. *Multi-Winner Voting with Approval Preferences*. Springer, 2022. doi: 10.1007/978-3-031-09016-5_1.

Martin Lackner, Jan Maly, and Simon Rey. Fairness in long-term participatory budgeting. In *Proceedings of IJCAI 2021*, 2021. https://www.ijcai.org/proceedings/2021/0042.pdf.

Maaike Los, Zoé Christoff, and Davide Grossi. Proportional budget allocations: Towards a systematization. In *Proceedings of IJCAI 2022*, 2022. doi: 10.24963/ijcai.2022/57.

Jan Maly, Simon Rey, Ulle Endriss, and Martin Lackner. Fairness in participatory budgeting via equality of resources, 2023. URL https://arxiv.org/abs/2205.07517.

Ryota Maruo, Tomohiko Yokoyama, Ayumi Igarashi, and Koh Takeuchi. Mechanism-informed learning for fair division. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, pp. 3420–3422, 2026. doi: 10.65109/HTOP9978. URL https://ifaamas.csc.liv.ac.uk/Proceedings/aamas2026/pdfs/HTOP9978.pdf.

Nicholas Mattei and Toby Walsh. Preflib: A library for preferences. In *Proceedings of ADT 2013*, 2013. doi: 10.1007/978-3-642-41575-3_20.

Farhad Mohsin, Ao Liu, Pin-Yu Chen, Francesca Rossi, and Lirong Xia. Learning to design fair and private voting rules. *Journal of Artificial Intelligence Research*, 75:1139–1176, 2022. doi: 10.1613/jair.1.13734. URL https://doi.org/10.1613/jair.1.13734.

Georgios Papasotiropoulos, Seyedeh Zeinab Pishbin, Oskar Skibski, Piotr Skowron, and Tomasz Wąs. Method of equal shares with bounded overspending, 2025. URL https://arxiv.org/abs/2409.15005.

David Parkes and Ariel Procaccia. Dynamic social choice with evolving preferences. In *Proceedings of AAAI 2013*, 2013. doi: 10.1609/aaai.v27i1.8570.

Dominik Peters and Piotr Skowron. Proportionality and the limits of welfarism. In *Proceedings of EC 2020*, 2020. doi: 10.1145/3391403.3399465.

Dominik Peters, Grzegorz Pierczyński, and Piotr Skowron. Proportional participatory budgeting with additive utilities, 2022. URL https://arxiv.org/abs/2008.13276.

Simon Rey, Felicia Schmidt, and Jan Maly. The (computational) social choice take on indivisible participatory budgeting, 2025. URL https://arxiv.org/abs/2303.00621.

Simon Rey et al. pabutools: Participatory budgeting tools for python (v1.2.3). https://pypi.org/project/pabutools/, 2024.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. *Advances in Neural Information Processing Systems 35*, 2022. doi: 10.52202/068431-0687.

Piotr Skowron and Adrian Górecki. Proportional public decisions. In *Proceedings of AAAI 2022*, 2022. doi: 10.1609/aaai.v36i5.20454.

Gogulapati Sreedurga, Mayank Ratan Bhardwaj, and Yadati Narahari. Maxmin participatory budgeting. In *Proceedings of IJCAI 2022*, 2022. doi: 10.24963/ijcai.2022/70.

Dariusz Stolicki, Stanisław Szufa, and Nimrod Talmon. Pabulib: A participatory budgeting library, 2020. URL https://arxiv.org/abs/2012.06539.

Stanisław Szufa, Piotr Faliszewski, Piotr Skowron, Arkadii Slinko, and Nimrod Talmon. Drawing a map of elections in the space of statistical cultures. In *Proceedings of AAMAS 2020*, 2020. doi: 10.65109/dvqm2845.

Nimrod Talmon and Piotr Faliszewski. A framework for approval-based budgeting methods. In *Proceedings of AAAI 2019*, 2019. doi: 10.1609/aaai.v33i01.33012181.

Nicholas Teh. The price of proportional representation in temporal voting, 2026. URL https://arxiv.org/abs/2605.11157.

Nguyen Thach, Xingchen Sha, and Hau Chan. Large language models for designing participatory budgeting rules. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, 2026. doi: 10.65109/LGEP3560. URL https://www.ifaamas.org/Proceedings/aamas2026/pdfs/LGEP3560.pdf.

Joshua C. Yang and Noemi Scheurer. Komitee equal shares: Choosing together as voters and as groups with a co-designed virtual budget algorithm, 2025. URL https://arxiv.org/abs/2510.02040.

# A NOTATION

| Symbol | Meaning |
|---|---|
| $I_t = (P_t, c_t, b_t, N_t, A_t)$ | edition $t$: projects, costs, budget, voters, ballots |
| $A_t(i) \subseteq P_t$ | the approval ballot of voter $i$ |
| $A_t(p) \subseteq N_t$ | the approvers of project $p$ |
| $N_t^{\mathrm{cov}} \subseteq N_t$ | voters with the demographics needed for the group ledger |
| $G$ | the group partition (age $\times$ sex cells the ballots record) |
| $g_t$ | covered members of group $g$ among edition-$t$ voters |
| $W_t \subseteq P_t$ | the funded set, $c_t(W_t) \leq b_t$ |
| $u_t(g)$ | money attributed to $g$ (Definition 1) |
| $e_t(g)$ | entitlement of $g$, its share of money spent |
| $\mathrm{CSD}_T(g)$ | cumulative share deficit of $g$ after $T$ editions |
| $\mathcal{H}_{t-1}$ | history available to a policy entering edition $t$ |
| $\beta_i$ | endowment of voter $i$; uniform $\beta_i = b_t/\lvert N_t\rvert$ is MES |
| $w \in \mathbb{R}^6$ | endowment-family weights |
| $v \in \mathbb{R}^5$ | outcome-family weights |
| $\rho$ | Equal Shares price level at which a project is bought |
| $\kappa$ | largest endowment multiple, $\max_i \beta_i \lvert N_t\rvert / b_t$ |
| $\tau, \mu$ | soft welfare target and its penalty weight |
| $\lambda$ | rollover intensity of RES |

# B CORPUS AND SPLITS

**Construction.** We build the fitting corpus from the frozen Pabulib snapshot (Stolicki et al., 2020). We keep nonempty approval elections with recorded winners, at most 300,000 voters, and demographic data on at least half of the ballots. We then group elections by (city, district). A group qualifies as a *series* if its longest consecutive run contains at least three editions. The corpus contains 19 series and 132 elections, comprising 821,572 ballots and 8,366 projects from 2016–2025. We exclude the Warsaw citywide contest and a Wawer subunit because both aggregate ballots that already appear in the district series.

**Fitting corpus versus available corpus.** We fit the headline Warsaw weights only on the primary corpus above. Applying the same filter to the complete public repository admits many additional series, mostly outside Warsaw, including histories in Gdynia and Łódź that are long enough for the temporal protocol. An earlier version used an incomplete local download and incorrectly attributed its limited coverage to the public data. We corrected the download without refitting the headline weights. The locked projected-support study uses separate pooled-temporal fits trained on pre-2023 histories from Gdynia, Warsaw, and Łódź. For the post-hoc native-approval replay, we use the frozen Warsaw headline fit. We report no fit on the complete corpus.

**Corrected-result lineage.** The primary Warsaw results come from a saved corrected replay with duplicate project tokens removed within each ballot and a deterministic project order. Canonicalization removes one repeated approval token in Wola 2021; no ballot is removed. The corrected corpus contains 6,577,175 approval tokens across 821,572 ballots. We retain the original artifacts and authenticate the corrected fit and evaluation records separately. Because the original results had already been inspected, the corrected Warsaw and first multi-city analyses are post-hoc replays, not new untouched holdout tests. The first multi-city decision remains *falsified*. The projected-support external study and the native-approval replay retain their separate recorded lineages. The manuscript reconstruction uses saved estimates and intervals; it does not rerun fitting or statistical tests.

**Missing demographics.** Every ballot contributes to the *allocation*. When age or sex is missing, the approvals still affect funding, but the ballot cannot enter the cohort ledger because we cannot assign it to a group. Project scores therefore count all ballots, while the fairness ledger counts only covered ballots. Confusing these counts caused the defect described in Appendix F. The filter requires demographics on half of an election's ballots. We do not model why fields are missing. If coverage depends on age or sex, for example through the voting channel, the ledger represents a nonrandom subsample, and that selection affects every deficit, entitlement, and CSD value. We do not test this missingness assumption.

**Splits.** The temporal split fits on editions through 2022 and evaluates on 2023–2025: 75 training editions and 54 held-out editions across 18 series. The Łódź citywide contest enters the fitting corpus only in 2023 and thus has no training editions. We exclude that series from the temporal split and use it for the no-refit city-out replay in Section K. The district series available in the complete repository are held out entirely for Appendix L. We use no Łódź data to fit the primary temporal policies. Repeated analysis of the citywide series limits its role to a descriptive transfer case rather than external validation.

A separate five-fold disjoint-series check uses the complete history of every series. We hold Łódź out in one fold and include it in training in the other four. This check evaluates disjoint series, so it tests neither temporal nor Warsaw-to-Łódź transfer. We save each split to disk and reload it to keep the boundaries fixed across runs.

Here, held-out editions are excluded from fitting the shared weight vector. Allocation still uses the edition-$t$ ballots, project costs, and budget once they are available. The recorded winner set is never a policy feature. Each counterfactual policy replays a series in order and updates $\mathcal{H}_t$ using only its own selected outcome after scoring edition $t$.

## C Feature definitions

Every feature is dimensionless. Setting $w = 0$ in the endowment parametrization recovers MES. In the outcome family, $v = e_1$ and $v = e_2$ recover the two baselines, while the zero vector does not. Let $n = |N_t|$, $b = b_t$, and write $\bar{\ell}$ for the mean ballot length over covered voters.

**Endowment family.** For group $g$ with carried deficit $D_{t-1}(g) = \sum_{\tau<t}(e_\tau(g) - u_\tau(g))$ and carried entitlement $E_{t-1}(g) = \sum_{\tau<t} e_\tau(g)$:

$$f_1(g) = \frac{\max(0, D_{t-1}(g))}{|g_t| \cdot (b/n)} \qquad \text{deficit per capita, in units of the base endowment}$$

$$f_2(g) = \operatorname{clip}\left(\frac{D_{t-1}(g)}{E_{t-1}(g)}, -1, 1\right) \qquad \text{signed deficit as a fraction of entitlement owed}$$

$$f_3(g) = \frac{|g_t|}{n_{\text{cov}}} - \frac{1}{|G|} \qquad \text{size relative to an equal split}$$

$$f_4(g) = \frac{\bar{\ell}_g}{\bar{\ell}} - 1 \qquad \text{ballot length relative to the electorate}$$

$$f_5(g) = \frac{\bar{c}_g}{\bar{c}} - 1 \qquad \text{mean cost per approval, relative}$$

$$f_6(g) = \cos\big(\phi_g, \phi_{\text{all}}\big) - \overline{\cos} \qquad \text{approval-profile overlap, mean-centred over groups}$$

Here $\phi_g$ is the vector of per-project approval frequencies within $g$. History enters only through $f_1$ and $f_2$. A voter without recorded demographics receives raw endowment $b/n$. For a covered voter, $w = \lambda e_1$ gives raw endowment $b/n + \lambda \max(0, D_{t-1}(g))/|g_t|$. The common budget normalization then recovers RES exactly.

**Outcome family.** For project $p$ with $\text{sh}(p) = |A_t(p)|/n$ counted over *all* ballots and $\text{cs}(p) = c_t(p)/b$:

$$\varphi_1(p) = \text{sh}(p) \qquad \text{approval share}$$

$$\varphi_2(p) = \frac{\text{sh}(p)/\text{cs}(p)}{\max_q \text{sh}(q)/\text{cs}(q)} \qquad \text{approvals per unit cost, normalized to } [0,1]$$

$$\varphi_3(p) = \text{cs}(p) \qquad \text{cost share}$$

$$\varphi_4(p) = \sum_g \hat{w}_g \cdot \frac{n_{t,g}(p)}{n_t^{\text{cov}}(p)} \qquad \text{support drawn from lagging groups}$$

$$\varphi_5(p) = \sqrt{\frac{\sum_g \big(s_g(p) - 1/|G|\big)^2}{1 - 1/|G|}} \qquad \text{concentration of support across groups}$$

If at least one group has a positive deficit, the weights satisfy $\hat{w}_g \propto \max(0, D_{t-1}(g))$ and sum to one. Otherwise, all weights are zero. The term $s_g(p)$ is the share of $p$'s covered approvers in $g$. Only $\varphi_4$ uses history. A ratio over covered approvers is set to zero when the project has no covered approver. Setting $v = e_1$ recovers greedy approval count, and $v = e_2$ recovers greedy approvals per unit cost.

The supporter-composition diagnostic in Section 3 uses $q_{t,g}(p)$ for group $g$'s share among the demographically covered approvers of project $p$, and $s_{t,g} = |g_t| / \sum_h |h_t|$ for its share among covered voters. We report

$$\mathrm{TV}_t(p) = \frac{1}{2} \sum_g |q_{t,g}(p) - s_{t,g}| \,,$$

as a raw-cost-weighted average over funded projects. We compute this quantity only on held-out editions after replaying each policy's training prefix. The optimizer uses this descriptive measure neither as a feature nor as an objective.

## D Mechanism implementation

**Algorithm 1** Equal Shares with general endowments, and completion

**Require:** instance $(P, c, b, N, A)$, endowments $\beta \geq 0$
$W \leftarrow \emptyset$; $\mathcal{A} \leftarrow \{p : A(p) \neq \emptyset,\ c(p) > 0\}$
**while** $\mathcal{A} \neq \emptyset$ **do**
  **for all** $p \in \mathcal{A}$ **do**
    **if** $\sum_{i \in A(p)} \beta_i < c(p)$ **then**
      $\mathcal{A} \leftarrow \mathcal{A} \setminus \{p\}$ ▷ never affordable; endowments only shrink
    **else**
      $\rho_p \leftarrow \min\{\rho : \sum_{i \in A(p)} \min(\beta_i, \rho) \geq c(p)\}$
    **end if**
  **end for**
  **if** $\mathcal{A} = \emptyset$ **then break**
  $p^\star \leftarrow \arg\min_{p \in \mathcal{A}} \rho_p$ ▷ ties by ascending cost, then id
  $\beta_i \leftarrow \beta_i - \min(\beta_i, \rho_{p^\star})$ for $i \in A(p^\star)$
  $W \leftarrow W \cup \{p^\star\}$; $\mathcal{A} \leftarrow \mathcal{A} \setminus \{p^\star\}$
**end while**
▷ completion: Equal Shares is not exhaustive, so spend the remainder
$r \leftarrow b - c(W)$
**for all** $p \notin W$ with $A(p) \neq \emptyset$ in descending $|A(p)|$, ties by ascending $c(p)$, then id **do**
  **if** $c(p) \leq r$ **then**
    $W \leftarrow W \cup \{p\}$; $r \leftarrow r - c(p)$
  **end if**
**end for**
**return** $W$

Completion spends the remaining budget by approval count and falls outside the payment rule. It accounts for the nonzero uniform Equal Shares value in Table 9; the learned-endowment value uses the same uniform reference floor, not the learned cap. We find $\rho_p$ in one pass over the sorted approver endowments. The absolute tolerance is $10^{-9}$ for aggregate affordability and $10^{-12}$ for water-filling breakpoints.

## E Optimization

We fit both arms with the same self-contained CMA-ES implementation. It uses 120 lines of NumPy and runs without a GPU or dependencies on SciPy, cma, or PyTorch.

The headline endowment fit and its two additional seeds use different numbers of generations. Their range gives an informal check of optimizer stability, with no estimate of variance under matched optimization budgets.

**Algorithm 2** Outcome family, with an optional support floor

**Require:** instance, weights $v$, optional $\kappa$
$E \leftarrow \{p : A(p) \neq \emptyset\}$
**if** $\kappa$ given **then**
  $E \leftarrow \{p \in E : |A(p)|/n \geq c(p)/(\kappa b)\}$
**end if**
$W \leftarrow \emptyset$; $r \leftarrow b$
**for all** $p \in E$ in descending $v^\top \varphi(p)$, ties by ascending $c(p)$, then id **do**
  **if** $c(p) \leq r$ **then**
    $W \leftarrow W \cup \{p\}$; $r \leftarrow r - c(p)$
  **end if**
**end for**
**return** $W$

| | Endowment arm | Outcome arm |
|---|---|---|
| Search dimension | 6 | 5 |
| Population $\lambda = 4 + \lfloor 3 \ln d \rfloor$ | 9 | 8 |
| Parents $\mu = \lfloor \lambda/2 \rfloor$ | 4 | 4 |
| Initial step $\sigma_0$ | 0.4 | 0.4 |
| Generations | 30 (headline); 25 (frontier/stability) | 25 |
| Objective evaluations | 271 headline; 225 frontier; 226 stability | 200 per target |
| Weight box | $[-10, 10]$ | $[-10, 10]$ |
| Penalty weight $\mu$ | 0 headline; 0/2.0 frontier | 0/2.0 frontier |
| Seeds | 42, 1, 2 | 42, 1, 2, 3 |
| Corrected-run wall clock | Not recorded | Not recorded |

Table 2: Optimizer settings. In Section 4.1, "same optimization budget" means 25 generations with population 8, as in the frontier fits used for comparison. The original fits used an 18-GB Apple M3 Pro CPU, without a GPU. The corrected fit artifacts do not record elapsed times, so we do not report corrected-run wall-clock ranges. Endowment evaluation replays iterative Equal Shares; the outcome arm uses a greedy pass.

Each of the five primary disjoint-series fits uses 25 generations and the coefficient box $[-10, 10]$. All five reach the boundary. We therefore ran a matched post hoc set with box $[-40, 40]$ and separate artifact tags, but four of the wider fits also reach their boundary. Table 20 compares held-out outcomes. These results support neither coefficient interpretation nor a claim of independent replication.

We leave the features unstandardized. Election-level ratios are dimensionless, and the two history features retain their longitudinal scale. No fit uses a restart. We initialize the means at the hand-designed references $\text{RES}(1)$ for endowments and greedy approvals-per-cost for outcomes. Because CMA-ES need not return its initial mean as a candidate, we replay and report every baseline separately.

**Why distinct soft welfare targets can return identical policies.** The penalty is $\mu \max(0, \tau - \text{ratio})$. When the unconstrained optimum exceeds $\tau$, it incurs no penalty and remains optimal. If $\tau$ exceeds every attainable welfare ratio, each candidate incurs $\mu(\tau - \text{ratio})$. Increasing $\tau$ then adds the same constant to every objective value, leaving the minimizer unchanged. The trade-off changes only in the intermediate regime. This formulation explains the identical rows in Table 8 without implying a failed search.

We use $\tau = 1$ as the headline outcome-arm operating point because its training target equals MES approval-count welfare. Neither this choice nor the penalty weight $\mu = 2$ was preregistered. Table 8 reports the full tested target grid. The headline selection remains exploratory.

## F CONTAINMENT VERIFICATION

The first run failed the third identity on 4 of 132 elections. The learned score counted approvals only on demographically covered ballots, while the deployed rule counted every ballot. These counts

| Identity | Scope | Tolerance | Result |
|---|---|---|---|
| $w = \mathbf{0} \Rightarrow$ MES | 19 series | exact | max $\lvert\Delta\rvert = 0$ |
| $w = \lambda e_1 \Rightarrow$ RES($\lambda$), 5 values | 19 series | exact | max $\lvert\Delta\rvert = 0$ |
| $v = e_1 \Rightarrow$ greedy-by-count | 132 elections | set equality | 132/132 |
| $v = e_2 \Rightarrow$ greedy-by-cost | 132 elections | set equality | 132/132 |
| environment vs. frozen replay | 19 series, 3 rules | $5 \times 10^{-4}$ | 19/19 |

agree when coverage is complete, but partial coverage can change the project order. We corrected the implementation by separating allocation counts from ledger counts (Appendix B). The containment test found the defect, which code review had not detected. Without that check, the wrong counts could have produced plausible results.

## G PROOF OF PROPOSITION 2

*Proof.* Let $W_{\text{pay}}$ be the outcome of Algorithm 1 before completion, and let $p \in W_{\text{pay}}$ be bought at price level $\rho$. By the purchase rule, $p$ enters $W_{\text{pay}}$ only when its approvers can cover its cost at that level:

$$\sum_{i \in A(p)} \min(\beta_i^{(p)}, \rho) \;\geq\; c(p),$$

where $\beta^{(p)}$ are the endowments remaining when $p$ is bought. Endowments are non-increasing along the run, since each purchase only subtracts, so $\beta_i^{(p)} \leq \beta_i$ for the initial vector $\beta$. Also $\min(x, \rho) \leq x$. Chaining these,

$$c(p) \;\leq\; \sum_{i \in A(p)} \min(\beta_i^{(p)}, \rho) \;\leq\; \sum_{i \in A(p)} \beta_i^{(p)} \;\leq\; \sum_{i \in A(p)} \beta_i.$$

Now suppose $\beta_i \leq \kappa\, b/|N|$ for every $i$. Then $\sum_{i \in A(p)} \beta_i \leq |A(p)| \cdot \kappa b/|N|$, and combining with the display above gives $c(p) \leq |A(p)|\kappa b/|N|$, i.e. $|A(p)|/|N| \geq c(p)/(\kappa b)$, which is the claim. □

The bound is attained when every approver of $p$ still holds exactly $\kappa b/|N|$ and spends the full amount on $p$, as can happen when none has paid for an earlier winner. Otherwise the bound is generally slack. Under the stated cap, the proposition restricts payment-phase purchases but does not determine a fitted policy's choices. Section 3.1 tests that region constraint separately and finds that it does not reproduce the mechanism.

## H REPRESENTATIVE SUPPORT: PROOF AND EXECUTABLE CHECK

*Proof of Proposition 1.* For a fixed group $g$, Definition 1 and the definitions of $q_{t,g}(p)$ and $E_g$ give the following, where $\omega_{t,p,g} = c_t(p)s_{t,g}/E_g$ and $r_{t,p,g} = q_{t,g}(p)/s_{t,g}$:

$$\begin{aligned}\mathrm{CSD}_T(g) &= 1 - \frac{\sum_t \sum_{p \in W_t} c_t(p) q_{t,g}(p)}{E_g} \\ &= 1 - \sum_t \sum_{p \in W_t} \frac{c_t(p) s_{t,g}}{E_g} \frac{q_{t,g}(p)}{s_{t,g}} = 1 - \sum_t \sum_{p \in W_t} \omega_{t,p,g} r_{t,p,g}.\end{aligned}$$

Because $\sum_{t,p} \omega_{t,p,g} = \sum_t c_t(W_t) s_{t,g}/E_g = 1$,

$$|\mathrm{CSD}_T(g)| = \left| \sum_{t,p} \omega_{t,p,g}(1 - r_{t,p,g}) \right| \leq \sum_{t,p} \omega_{t,p,g} |1 - r_{t,p,g}| \leq \epsilon.$$

For the construction, let the positive rational group shares be $(s_g)_g$ and choose a positive integer $m$ such that every $ms_g$ is an integer. Choose a larger electorate size $n$ with every $ns_g$ integral and $m/n \leq \delta$. Create one project of cost $b$, approved by exactly $ms_g$ voters from each group, and a budget-feasible dummy project, also of cost $b$, approved by every other voter. Fund only the representative project; it exhausts the budget, so the dummy remains unfunded. Its approver

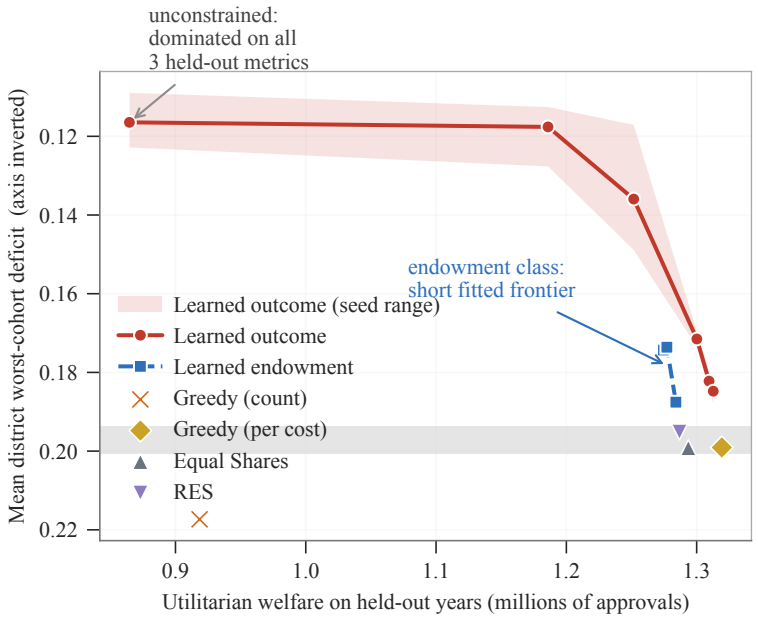



Figure 4: Held-out fairness–welfare results for the fitted policy classes. The grey band covers four hand-designed rules: greedy approvals/cost, MES, and RES($\lambda$=0.5, 1.0). Its width is 0.0040. Both learned classes extend beyond the band. Direct scores span 0.0717 (red, with variation across 4 seeds), while learned endowments span 0.0140 (blue). The classes learn different objects from different features, so this comparison cannot identify the causal effect of the payment rule. Mild regularization improves both fairness and welfare over the unpenalized direct-score point.

composition is $q_g = s_g$, so the identity gives $\mathrm{CSD}(g) = 0$ for every group. Exactly $m$ of $n$ voters approve a funded project, hence exclusion is $1 - m/n \geq 1 - \delta$. □

The executable construction uses 8 equal groups. We keep one supporter per group and increase the electorate, evaluating each case with the same `group_outcome`, `cumulative_share_deficit`, and `exclusion_rate` functions used in the experiments.

| Voters | Winner supporters | Worst \|CSD\| | Exclusion |
|---|---|---|---|
| 80 | 8 | 0.0000 | 90.0% |
| 200 | 8 | 0.0000 | 96.0% |
| 400 | 8 | 0.0000 | 98.0% |
| 800 | 8 | 0.0000 | 99.0% |

Table 3: Representative-support construction from `iclr_degeneracy.py`. CSD remains exactly zero in the production ledger as exclusion approaches one. This construction tests the metric's limits. It is neither a fitted result nor a candidate policy.

## I STATISTICAL METHODOLOGY

We use the *series* as the inferential unit. Editions within a district share a project pool and electorate, and the districts themselves belong to one city. The intervals describe dispersion across districts in this corpus. They cannot remove common-city shocks or support city-level population inference. For each policy contrast, we first calculate the difference in held-out worst-group CSD within each series, then take an equally weighted mean across series. The reported 2.5–97.5 paired percentile-bootstrap interval uses 10,000 resamples of these series pairs.

The exact paired test enumerates all $2^{18}$ sign assignments of the series differences under the null. It reports the two-sided fraction whose absolute mean is at least as large as the absolute observed mean. The test assumes sign-exchangeable paired differences under the null and treats neither elections nor voters as independent replicates. We display values below $0.001$ as $p < 0.001$. A "win" is a series with a negative difference. We report wins alongside the mean because large improvements in a few series can lower that mean.

In the disjoint-series analysis, the pooled 19-series mean, win count, and paired bootstrap interval are descriptive because all series in a held-out fold share one fitted map. As a sensitivity check, we average the paired effects within each fold and enumerate all $2^5$ sign assignments of the five fold means. The fold-block $p$-value has a minimum two-sided resolution of $0.0625$. These five fits use overlapping training sets and cannot count as independent replications. This analysis also uses each series' complete history, so it tests neither temporal nor cross-city transfer.

We apply Holm–Bonferroni correction to the 3 focal held-out Warsaw comparisons reported in the main text: direct scoring against MES, learned endowments against MES, and the senior tilt against MES. The full welfare-target grid, closest-prior comparison, matched-target check, allocation-kernel intervention, and three group partitions remain secondary exploratory analyses and are unadjusted unless a separate protocol states otherwise. The partition analyses re-score the same fitted policies, which makes them dependent. Neither the focal contrasts nor the frozen-score intervention was preregistered. Adjusting the values controls multiplicity within this reported family, but the evidence remains exploratory. In the post-lock audit, the age lookup changes CSD relative to the senior tilt by -0.0217 (95% CI [-0.0382, -0.0078], secondary exploratory $p$ = 0.007, 13/0/5 wins/ties/losses).

## J EXTENDED RELATED WORK

We expand the comparison in Section 7 by examining what each method learns, which mechanism it retains, and how it is evaluated.

**Participatory budgeting as a social-choice object.** Indivisible participatory budgeting extends approval-based committee selection to projects with costs, bringing many of the same axiomatic questions into a budgeted setting (Aziz et al., 2018; Talmon & Faliszewski, 2019). Surveys and systematizations review the available rules (Rey et al., 2025; Los et al., 2022), and handbooks describe their place in computational social choice (Brandt et al., 2016; Lackner & Skowron, 2022). Fairness targets vary. Prior work studies justified representation (Aziz et al., 2015; Peters et al., 2022), the core (Fain et al., 2016), egalitarian or maximin objectives (Sreedurga et al., 2022), equality of resources (Maly et al., 2023), and sequential-payment rules such as Equal Shares and Phragmén-style methods (Peters & Skowron, 2020; Brill et al., 2023).

A proportional rule may leave money unspent, so exhaustiveness requires a separate design choice (Papasotiropoulos et al., 2025). We evaluate Algorithm 1 with its approval-count completion included. Fair public decisions outside budgeting (Conitzer et al., 2017) and sortition-style panel selection (Flanigan et al., 2021) also raise group fairness questions, but they lack the monetary attribution on which CSD is built.

**Fairness across repeated decisions.** Dynamic social choice studies obligations to voters who lose across a sequence of decisions (Parkes & Procaccia, 2013; Freeman et al., 2017). Perpetual voting formalizes the problem for single-winner sequences (Lackner, 2020), including justified-representation analogues (Bulteau et al., 2021). Related work extends JR/PJR/EJR to sequences of unit-cost decisions (Elkind et al., 2023; 2024) and examines verification complexity (Elkind et al., 2025), sequential proportionality (Skowron & Górecki, 2022; Chandak et al., 2026), adaptive settings (Kraiczy & Elkind, 2023), and the price of temporal representation (Teh, 2026). Long-term PB for voter types is the closest antecedent (Lackner et al., 2021). These papers mainly analyze axioms and worst cases, often on synthetic instances. We fit policies within a deployed mechanism and evaluate them by replaying recorded ballots.

Repeated artificial-currency allocation also uses personalized endowments (Gorokh et al., 2016). There, the endowments parameterize an all-pay auction for one repeated resource and permit incentive and efficiency guarantees. We fit endowments from PB data for Equal Shares over combinatorial

public projects. Our evaluation measures group outcomes on recorded ballots and does not model strategic bids.

Mixed-rule work gives MES a set of projects selected earlier in the same election. Four hand-designed "pre-allocation" methods charge those projects to voters before rebalancing the remaining budgets (Baychkov et al., 2026b). The analysis proves EJR-style guarantees and evaluates within-election Greedy–MES mixtures on 313 Pabulib instances. Our map uses current group features and, optionally, prior editions to set endowments before any project is selected. We compare this payment mechanism with direct winner scores by measuring held-out fairness, welfare, coverage, and susceptibility to proxy gaming. We claim neither the first nonuniform MES budgets nor a new EJR guarantee.

Yang & Scheurer (2025) change the virtual-budget population by combining equal-budget voters with weighted "value agents" for deliberatively chosen impact fields. Participants elicit and co-design the weights for one election. This changes what the MES ledger represents without fitting endowments from past allocation outcomes or learning a policy over repeated elections.

**Empirical participatory budgeting.** Pabulib standardizes observed PB instances (Stolicki et al., 2020), following the PrefLib tradition (Mattei & Walsh, 2013). The surrounding infrastructure includes analysis tools (Rey et al., 2024), methodological guidance (Faliszewski et al., 2023; Boehmer et al., 2024), and maps of the instance space (Szufa et al., 2020). Empirical studies examine ballot formats and elicitation (Fairstein et al., 2023; Goel et al., 2019; Benade et al., 2017), spatial fairness across districts (Hershkowitz et al., 2021), interactions in voter behavior (Jain et al., 2020; Fairstein et al., 2024a), and the utilitarian cost of Equal Shares (Baychkov et al., 2026a). The last quantity enters the trade-off controlled by our soft welfare target.

Equal Shares is used in practice (Equal Shares team, 2023), within a longer policy history of participatory budgeting (Cabannes, 2004). Existing empirical designs generally study one election or analyze many elections independently. We link successive editions of the same district.

**Learning allocation objects in participatory budgeting.** Learning PB aggregation rules predates this paper. Fairstein et al. (2024b) adapt Set Transformers to produce project scores. They train from example outcomes or explicit objectives, recover or blend Approval Voting, Chamberlin–Courant, and PAV-like rules, and test on Warsaw Pabulib elections. At artifact freeze, we found no public implementation of that model. Reconstructing its architecture and training would require a new implementation, so we discuss the method without claiming to reproduce it. Thach et al. (2026) use LLM-guided evolution to generate executable project-priority programs for utilitarian and Strong-EJR-approximation objectives. They evaluate more than 600 PB instances and also use the learned priorities to complete Equal Shares.

Both prior methods learn which projects to select. Our payment-retaining family learns Equal Shares endowments, which we evaluate with a held-out longitudinal ledger. We use the direct-outcome family to diagnose objective gaming through counterfactual replay. We do not propose it for deployment.

Other PB systems learn at different stages. Adams et al. (2026) train voter agents to revise cumulative ballots while keeping Greedy or Equal Shares fixed. Damle & Faltings (2025) prompt language models to infer preferences and select feasible portfolios. These methods support voter decisions and reason from language to allocations, respectively. Neither learns an internal parameter of a recorded allocation mechanism.

For an executable comparison, we transcribed the two published approval-ballot scores from Thach et al. (2026) and replayed both without refitting. Count satisfaction matches our welfare definition, while cost satisfaction does not. We report both scores to avoid selecting one after observing the results.

Beyond PB, deep learning has been used to design auctions, voting rules, and redistribution mechanisms fitted to human preferences (Dütting et al., 2021; Mohsin et al., 2022; Koster et al., 2022). We learn a low-dimensional parameter of an existing proportional payment procedure, with exact containment checks that recover its hand-designed references. In the closest mechanism-informed example outside PB, Maruo et al. (2026) learn missing chore preferences through differentiable ap-

| Frozen published score | Mean worst-group CSD ↓ | Welfare ↑ | Exclusion ↓ |
|---|---|---|---|
| LLMRule, cost satisfaction | 0.2142 | 1,009,756 | 0.0689 |
| LLMRule, count satisfaction | 0.2007 | 1,291,397 | 0.0466 |

Table 4: Closest-prior learned rules replayed on the held-out editions using the published formulas without retraining. The count-satisfaction row supplies the comparison in Table 1 because it uses the same approval-count welfare statistic.

| Rule | Mean worst-group CSD ↓ | Welfare ↑ |
|---|---|---|
| Deployed outcome | 0.1324 | 320,936 |
| MES | 0.1702 | 449,934 |
| Greedy (approvals/cost) | 0.0700 | **464,787** |
| Learned endowment map | 0.1326 | 447,075 |
| Learned outcome policy | **0.0620** | 405,464 |

Table 5: Warsaw-fitted policies replayed without refitting on Łódź in 2023–2025. Both learned policies improve on Equal Shares in this city-out case. Greedy approvals per cost is nearly as fair and has higher welfare, while the city's deployed rule has lower CSD than Equal Shares.

proximations to adjusted winner, round robin, and moving knife. They learn preference information, while we learn endowments for a fixed payment procedure.

Contextual resource-allocation metrics can disagree even in the absence of strategic behavior or optimization failure. Allocation parity and outcome parity are generally incompatible (Jo et al., 2023). We examine a narrower empirical case. CSD attributes spending by supporter composition, so direct optimization over winner sets can match group shares while serving few voters. The tested payment-retaining class does not reach the same corner under the same objective. In Section 3.1, we first decompose the metric and give an executable construction. We then compare policy classes and test changes to the allocation kernel with frozen scores.

**Properties retained during payment.** Learning the endowment parametrization preserves budget feasibility and approver-funded accounting during the payment phase. Section 3.1 measures the empirical consequence of that restriction. Nonuniform endowments and greedy completion prevent us from invoking standard priceability.

## K CROSS-CITY TRANSFER AND BREADTH STRESS TEST

We fit the primary temporal policies on Warsaw districts through 2022. The Łódź citywide contest enters the fitting corpus in 2023, so we exclude it from that split. The policies below have seen none of its ballots or projects and no data after 2022. These citywide elections are also an order of magnitude larger than a Warsaw district election. We replay the frozen policies without retraining to obtain one descriptive transfer case across city, time, and scale.

The headline endowment map reaches 0.1326, compared with 0.1324 for the deployed rule. Its welfare is substantially higher, but its deficit is no lower. The incumbent therefore gives a more informative reference than Equal Shares in this case. This single series records performance on one unseen city, scale, and demographic profile and cannot establish external validity.

**Superseded breadth stress test.** We ran this analysis before correcting the corpus in Appendix B. The incomplete local download made the citywide case appear to be the only non-Warsaw series with three consecutive editions. The scores remain valid replays of frozen policies on observed elections. For evidence of breadth, the native-approval study in Appendix L supersedes this analysis by using the complete repository and full temporal protocol.

The earlier stress test lowers the minimum run length to one and leaves the approval-ballot, recorded-winner, demographic-coverage, and voter-count filters unchanged. This outcome-independent rule selects 17 series and 21 elections from Gdynia, Lublin, and Łódź. We replay the same frozen

| Scope | S/E | MES | Endow. | Δ | W/MES | Outcome | W/MES |
|---|---|---|---|---|---|---|---|
| Gdynia | 10/12 | 0.2024 | 0.2024 | +0.0000 | 1.000 | 0.1966 | 0.973 |
| Lublin | 2/2 | 0.1648 | 0.1648 | +0.0000 | 1.000 | 0.1243 | 0.975 |
| Łódź | 5/7 | 0.2019 | 0.1721 | -0.0299 | 0.994 | 0.1688 | 0.904 |
| All | 17/21 | 0.1979 | 0.1891 | -0.0088 | 0.994 | 0.1799 | 0.910 |

Table 6: No-refit results for every eligible shorter non-Warsaw series in the superseded stress set. S/E denotes series/elections. CSD gives each series equal weight; welfare (W) is summed within each scope and normalized to MES. "Outcome" denotes the headline soft-target-1.00 policy. The scores use full runs and are not temporal holdout estimates.

Warsaw fits over each complete run, with no refitting or policy selection after inspection. Most runs contain one edition, and all series come from only 3 cities. Table 6 is descriptive and includes neither confidence intervals nor $p$-values.

Across all selected series, the endowment map lowers mean CSD by 0.0088, reduces welfare by 0.6%, and lowers exclusion by 0.0014. The map returns exactly the MES allocations in Gdynia and Lublin, so the entire aggregate CSD change comes from Łódź. The headline outcome policy lowers mean CSD in each city, reduces overall welfare by 9.0%, and raises exclusion by 0.0101. The frozen result file also contains the aggressive outcome point and every baseline. These results extend the search for failures within the superseded stress set. The main longitudinal experiment still covers only one city.

## L EXTERNAL PROTOCOLS, LOCKS, AND THE NATIVE-APPROVAL STUDY

**Protocol contents and lineage.** The learned-priority and projected-support endowment studies each write a protocol lock and one-way opening receipt before evaluation. The locks fix the corpus manifest, fitted-weight digests, seeds, aggregation, decision thresholds, and comparison directions. The projected-support study also publishes the lock SHA-256 in a separate anchor. We evaluated the native-approval replay without a pre-opening lock, so it remains post-hoc. Table 14 traces the three versions of the projected-support study. These versions belong to a single external study.

Version v1 stopped before computing any metric because it encountered cumulative ballots under the locked approval-only condition. Version v2 used the locked support-set projection but also stopped before computing a metric: the Unicode source label *Kraków* did not match the literal analysis token *Krakow*. Version v3 added the source-to-analysis city alias and changed only this identity check. It kept the same corpus bytes, projection, weights, statistics, and decision thresholds. Neither v1 nor v2 produced a contrast, so neither restart could depend on an observed outcome. The v3 decision applies the thresholds fixed before opening its data. We created these locks after the in-domain Warsaw analysis, which remains exploratory.

A whole-tree lock also fixes the presentation code, including the scripts that generate manuscript numbers and figures. Later paper edits therefore appear as tree drift even when scientific inputs stay unchanged. The audit treats these changes separately. Corpus files, fitted weights, the corpus manifest, lineage records, and evaluation modules must remain byte-identical. The supplied `iclr_lock_drift.py` audit fails when any of them changes. Manuscript generators and figure scripts may change after evaluation, with each changed file named in the audit. These presentation edits cannot alter a decision that has already been computed and stored.

One evaluation module changed after publication of the locks. We amended the cohort ledger to sum project costs in a fixed project order, removing the dependence of floating-point accumulation on set iteration order. The audit reports decision-critical drift because the locked decisions used the earlier bytes. At the time, regenerating the macros from the original frozen results reproduced the then-reported manuscript values, and the replay gate matched the frozen reference implementation on all 19 series. This check did not cover the later corrected replays described in Appendix B. The amendment record in `iclr_lock_amendments.json` contains the locked and amended digests, the reason, and these checks. The audit accepts a decision-critical difference only if it matches that record exactly. Any other difference on a decision-critical file fails, as does an amendment record that declares an effect on the recorded decision.

| Condition (no welfare penalty) | Mean worst-group CSD | Welfare | Exclusion |
|---|---|---|---|
| Outcome family, unconstrained | 0.1228 | 586,114 | 0.1470 |
| + support floor $\kappa = 1$ | 0.1058 | 775,230 | 0.0858 |
| + support floor $\kappa = 2$ | 0.1036 | 693,752 | 0.0975 |
| Learned endowment map (reference) | 0.1736 | 1,273,257 | 0.0414 |
| MES (reference) | 0.1992 | 1,293,610 | 0.0438 |

Table 7: Retrained static-floor control for Proposition 2. We fit the outcome family with *no welfare penalty*. The support floor adds a one-shot coverage restriction without iterative approver payment. At $\kappa = 1$, all three reported outcomes improve over the unconstrained policy, yet welfare remains roughly forty percent below Equal Shares and exclusion roughly twice as high. Table 11 instead holds the learned score weights fixed and introduces payment dynamics separately.

**Projected support sets.** Katowice uses cumulative ballots and Krakow uses ordinal ballots. Neither is a native approval instance. The locked projection keeps every listed project and discards point magnitude and rank. Two cumulative ballots listing the same projects become indistinguishable even when they distribute points differently among those projects. Table 12 reports each city, seed, interval, Holm-adjusted $p$, exclusion change, and welfare ratio. Table 13 gives every locked condition, threshold, observed value, and verdict.

**Native approval ballots.** The post-hoc native-approval replay uses ballots directly, without projection or refitting. In the complete public repository, Gdynia and Łódź both record approval ballots and have histories long enough for the main temporal protocol. We stage every eligible series from these cities in a separate corpus directory to keep the fitting index fixed. We then replay the frozen Warsaw fit without refitting or policy selection.

We prevent duplicate electorates in the same way as for Warsaw's citywide and Wawer aggregates. When Gdynia runs small-project and large-project ballots in the same neighborhood and edition, we keep only the larger-budget track. We also drop citywide aggregates. Each series is replayed from its first edition and scored from 2023, so its warm-up history remains endogenous. The fit has seen none of these ballots, projects, or cities, and every scored edition postdates fitting. Table 21 reports the contrasts and actuation grid.

The native-approval replay finds no detectable series-level difference. The frozen Warsaw map returns exactly the Equal Shares allocation on 36 of 58 series. These series have fewer projects, with a median of 6.2 per scored election versus 22.2 where the map acts. A post-hoc verification artifact cross-checks this count by authenticating the staged corpus and recording canonical sorted winner IDs for both policies in every election. We count a series as unchanged only if every scored winner set matches exactly. This artifact is a verification record, not a protocol lock. Equal Shares orders purchases by the price at which approvers can cover the cost. The estimated effect stays near zero across the reported minimum-project grid, including the highest threshold, where none of the retained series is inert. These post-hoc subsets differ in size and composition, so they do not establish why the all-series difference is undetectable. We report the full grid without selecting a preferred subset.

## M FULL RESULT TABLES

The following tables use the same result files as the main text. They report the values for each series, seed, partition, and condition that underlie the aggregates.

## N ADDITIONAL ROBUSTNESS AND DIAGNOSTIC TABLES

**Group-partition robustness.** The main analysis partitions voters by age $\times$ sex, and the objective is defined relative to that choice. We therefore rescore the same fitted policies under coarser partitions. The policies remain those trained on the original groups. No model is refitted.

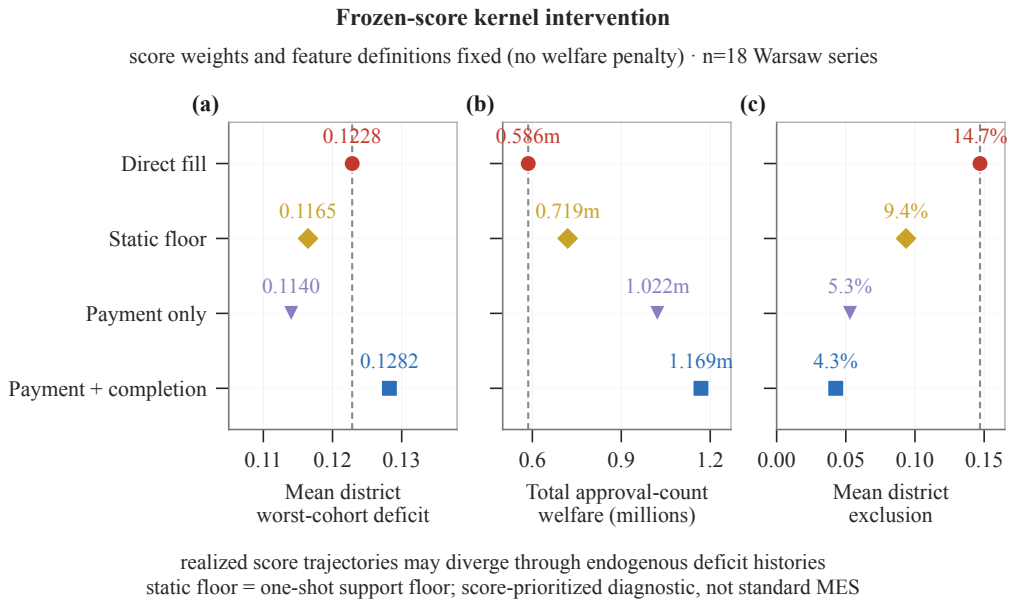



Figure 5: Frozen-score allocation-kernel intervention on $n = 18$ Warsaw district series. The panels compare mean worst-group CSD, total approval-count welfare, and mean exclusion under direct fill, a one-shot support floor, score-prioritized approver-funded payment, and payment plus completion. Score weights and feature definitions remain fixed. Realized score trajectories may still diverge because each rule generates its own deficit history. Payment plus completion lowers exclusion from 0.1470 to 0.0427 and multiplies welfare by 1.99×; the CSD change is not detectable ($p = 0.721$). The diagnostic is not standard MES and provides neither a universal nor an external guarantee.

| Welfare target $\tau$ | Worst CSD | Mean CSD | Welfare | Exclusion |
|---|---|---|---|---|
| *Endowment family (mechanism retained)* | | | | |
| 0.00 | 0.1743 | 0.0374 | 1,274,522 | 0.0418 |
| 0.99 | 0.1736 | 0.0371 | 1,277,125 | 0.0418 |
| 1.00 | 0.1876 | 0.0395 | 1,284,042 | 0.0433 |
| *Outcome family (mechanism discarded)* | | | | |
| 0.00 | 0.1228 | 0.0024 | 586,114 | 0.1470 |
| 0.85 | 0.1153 | 0.0101 | 1,198,311 | 0.0450 |
| 0.95 | 0.1483 | 0.0228 | 1,259,890 | 0.0442 |
| 1.00 | 0.1728 | 0.0320 | 1,303,057 | 0.0447 |
| 1.02 | 0.1828 | 0.0352 | 1,310,536 | 0.0451 |
| 1.05 | 0.1869 | 0.0364 | 1,312,691 | 0.0450 |

Table 8: Sweeps over soft welfare targets for both arms at seed 42 on held-out Warsaw districts. Worst CSD is the equal-weight mean district-level worst-group value; lower CSD and exclusion and higher welfare are favorable. Targets are penalties, not hard constraints. This single-seed descriptive sweep has no intervals; corpus- and budget-specific ranges are not universal reachable sets.

Both learned policies remain ahead of Equal Shares for every tested partition ($p \leq 0.0130$ in exact paired sign-flip tests). These dependent rescores evaluate the same allocations and provide no independent transfer evidence. Absolute deficits fall for coarser groups because averaging over fewer, larger groups removes variation, but the policy differences remain.

Greedy approvals per cost changes rank across partitions: it beats Equal Shares under `age` and loses under `sex`. Rollover Equal Shares has slightly lower mean CSD than Equal Shares under both coarser partitions: 0.1106 versus 0.1127 for age and 0.0558 versus 0.0585 for sex.

**Support-floor separator.**

| Policy | % of funded projects | % of budget |
|---|---|---|
| Greedy (approvals/cost) | 0.06% | 0.05% |
| MES | 0.18% | 0.29% |
| Learned endowment map | 0.31% | 0.64% |
| Learned outcome (soft target 1.0) | 0.41% | 0.43% |
| Learned outcome (unconstrained) | **10.29%** | **25.06%** |

Table 9: Spending on projects below the $\kappa = 1$ payment-phase support floor, $|A(p)|/|N| < c(p)/b$. The unpenalized direct policy spends 25.06% of its budget on such projects. Every tested payment-based rule stays below one percent. For uniform Equal Shares, these projects enter through completion. Nonuniform learned endowments need not satisfy the $\kappa = 1$ floor during payment. This comparison describes the observed policies, not a universal feasible region.

**Outcome-feature ablation.**

| Feature removed | Mean worst-group CSD | $\Delta$ vs full |
|---|---|---|
| Cost share | 0.1791 | +0.0063 |
| Approvals per cost | 0.1781 | +0.0053 |
| Deficit-weighted support | 0.1760 | +0.0032 |
| Approval share | 0.1699 | -0.0029 |
| Cohort concentration | 0.1633 | **-0.0095** |

Table 10: Leave-one-feature-out results for direct scores. Each variant is refitted once with the full model's optimization budget and evaluated on held-out elections. Removing cohort concentration, which identifies projects with demographically balanced support, improves held-out CSD by the largest amount in this single-refit comparison.

**Remaining generated tables.** The appendix tables read the same stored result files and generated values as the main text. Table 11 reports the frozen-score replay at three soft targets, and Table 15 lists every series at the headline operating point, including the losses. Tables 19 and 20 give the disjoint-series contrasts and post hoc coefficient-box sensitivity. Table 17 reports seed sensitivity, while Table 22 gives both learned classes for every held-out district. Table 21 reports the native-approval difference -0.0039 across 58 series and the actuation grid. Table 18 contains the complete group-partition robustness sweep.

| Soft target | Allocation kernel | Worst CSD | $\Delta$ direct | Welfare | W/direct | Exclusion |
|---|---|---|---|---|---|---|
| 0.00 | direct greedy fill | 0.1228 | +0.0000 | 586,114 | 1.00 | 0.1470 |
| | direct + static floor ($\kappa = 1$) | 0.1165 | -0.0064 | 719,168 | 1.23 | 0.0935 |
| | iterative payment only | 0.1140 | -0.0088 | 1,021,566 | 1.74 | 0.0530 |
| | iterative payment + completion | 0.1282 | +0.0054 | 1,168,671 | 1.99 | 0.0427 |
| 0.85 | direct greedy fill | 0.1153 | +0.0000 | 1,198,311 | 1.00 | 0.0450 |
| | direct + static floor ($\kappa = 1$) | 0.1340 | +0.0187 | 1,218,887 | 1.02 | 0.0434 |
| | iterative payment only | 0.1605 | +0.0452 | 1,132,666 | 0.95 | 0.0534 |
| | iterative payment + completion | 0.1836 | +0.0683 | 1,273,668 | 1.06 | 0.0430 |
| 1.00 | direct greedy fill | 0.1728 | +0.0000 | 1,303,057 | 1.00 | 0.0447 |
| | direct + static floor ($\kappa = 1$) | 0.1799 | +0.0071 | 1,302,820 | 1.00 | 0.0449 |
| | iterative payment only | 0.1786 | +0.0058 | 1,130,853 | 0.87 | 0.0570 |
| | iterative payment + completion | 0.1979 | +0.0251 | 1,289,024 | 0.99 | 0.0431 |

Table 11: Allocation-kernel replay over held-out Warsaw district series. Within each target, $\Delta$ is kernel minus direct-greedy CSD; negative values favor the kernel, as do lower exclusion and higher welfare. Values are descriptive without intervals. Score weights and feature definitions are fixed, but history-dependent values and scores may diverge. The static floor uses Proposition 2 at $\kappa = 1$; uniform-balance score-prioritized payment is diagnostic, not standard MES.

| Seed | City | Δ CSD ↓ | 95% paired-bootstrap interval | Holm $p$ | Δ exclusion ↓ | Welfare ratio ↑ |
|---|---|---|---|---|---|---|
| 1 | Katowice | -0.0125 | [-0.0307, 0.0016] | – | +0.0078 | 0.9913 |
| 1 | Krakow | -0.0138 | [-0.0407, 0.0082] | – | +0.0037 | 1.0014 |
| 2 | Katowice | -0.0027 | [-0.0080, 0.0000] | – | +0.0010 | 1.0002 |
| 2 | Krakow | -0.0080 | [-0.0278, 0.0051] | – | -0.0014 | 1.0015 |
| 42 | Katowice | -0.0117 | [-0.0282, 0.0000] | 0.5000 | +0.0085 | 0.9924 |
| 42 | Krakow | -0.0026 | [-0.0259, 0.0166] | 0.8398 | +0.0015 | 1.0041 |

Table 12: Projected-support-set results by city and frozen seed, with series as paired analysis units. Δ is learned endowment minus Equal Shares, so negative values favor the learned map. Intervals use paired-series bootstrap resampling, and Holm-adjusted $p$ values are shown only for the primary seed. Ballot projection limits the transfer interpretation.

| Frozen condition | Observed | Status |
|---|---|---|
| city macro $\leq -0.005$ for every seed | maximum -0.0054 | ✓ |
| negative in each city for every seed | 6/6 negative | ✓ |
| primary paired-bootstrap interval excludes zero in each city | Katowice 0.0000, Krakow 0.0166 | × |
| primary Holm $p \leq 0.05$ in each city | Katowice 0.5000, Krakow 0.8398 | × |
| welfare ratio $\geq 0.98$ in each city and seed | minimum 0.9913 | ✓ |
| exclusion increase $\leq 0.005$ in each city and seed | maximum 0.0085 | × |
| endowment actuation $\geq 8$ and greater than priority | 12 versus 3 series | ✓ |
| city-macro seed spread $\leq 0.005$ | 0.0078 | × |

Table 13: One row per preregistered condition for the projected-support-set external study, judged in the direction printed in the condition. The deterministic joint status is `falsified` because the verdict is conjunctive and not all conditions pass.

| Version | Lock SHA-256 prefix | Stage | Metrics computed | Classification |
|---|---|---|---|---|
| v1 | `064957412947` | approval-only abort | No—no scientific result | protocol_invalid |
| v2 | `1464a34e2274` | Unicode-label technical abort | No—no scientific result | technical_invalid |
| v3 | `a6d7f945b6f6` | valid amended run | Yes | falsified |

Table 14: v1 and v2 ended before metrics were computed. Only v3 contributes results. Matching hashes authenticate the artifacts, not the transfer claim.

| Series | Test years | MES | Learned outcome | Δ | Worst group |
|---|---|---|---|---|---|
| Wesoła | 2023, 2024, 2025 | 0.2727 | 0.1359 | -0.1368 | 60+ (M) |
| Praga-Północ | 2023, 2024, 2025 | 0.1953 | 0.1119 | -0.0833 | 60+ (M) |
| Praga-Południe | 2023, 2024, 2025 | 0.2333 | 0.1808 | -0.0524 | 60+ (M) |
| Włochy | 2023, 2024, 2025 | 0.1178 | 0.0696 | -0.0482 | 25-39 (M) |
| Wola | 2023, 2024, 2025 | 0.1933 | 0.1501 | -0.0432 | 60+ (M) |
| Rembertów | 2023, 2024, 2025 | 0.1665 | 0.1264 | -0.0401 | 60+ (M) |
| Bemowo | 2023, 2024, 2025 | 0.1583 | 0.1185 | -0.0397 | 60+ (M) |
| Bielany | 2023, 2024, 2025 | 0.1056 | 0.0693 | -0.0363 | 40-59 (M) |
| Ursus | 2023, 2024, 2025 | 0.2338 | 0.2010 | -0.0328 | 60+ (F) |
| Mokotów | 2023, 2024, 2025 | 0.2917 | 0.2691 | -0.0226 | 60+ (M) |
| Ursynów | 2023, 2024, 2025 | 0.0802 | 0.0684 | -0.0118 | 40-59 (M) |
| Wawer | 2023, 2024, 2025 | 0.1580 | 0.1517 | -0.0062 | 60+ (M) |
| Śródmieście | 2023, 2024, 2025 | 0.2624 | 0.2578 | -0.0046 | 60+ (M) |
| Targówek | 2023, 2024, 2025 | 0.1697 | 0.1676 | -0.0021 | 60+ (M) |
| Ochota | 2023, 2024, 2025 | 0.2091 | 0.2131 | **+0.0040** | 60+ (M) |
| Żoliborz | 2023, 2024, 2025 | 0.3083 | 0.3153 | **+0.0070** | <25 (M) |
| Wilanów | 2023, 2024, 2025 | 0.2552 | 0.2656 | **+0.0104** | 60+ (M) |
| Białołęka | 2023, 2024, 2025 | 0.1737 | 0.2383 | **+0.0645** | 60+ (M) |

Table 15: Per-series worst-group CSD for the learned outcome policy at soft target 1.00 on held-out Warsaw editions. Δ is learned minus Equal Shares; negative values favor learned and positive losses are bold. Rows are descriptive without intervals. The aggregate is the equal-weight mean district-level worst-group CSD, all from one city.

| Rollover intensity $\lambda$ | Worst CSD | Welfare | Exclusion |
|---|---|---|---|
| 0.00 | 0.1992 | 1,293,610 | 0.0438 |
| 0.25 | 0.1977 | 1,292,741 | 0.0435 |
| 0.50 | 0.1983 | 1,288,507 | 0.0433 |
| 0.75 | 0.1987 | 1,287,880 | 0.0436 |
| 1.00 | 0.1951 | 1,286,793 | 0.0436 |

Table 16: All tested rollover intensities on held-out Warsaw editions. Lower CSD and exclusion and higher welfare are favorable; the former are equal-weight district-series means and welfare is summed. Rows are descriptive without intervals. $\lambda = 0$ is Equal Shares; $\lambda = 1$ is the main-text reference and optimizer initial mean, not selected on held-out results.

| Soft welfare target | seed 1 | seed 2 | seed 3 | seed 42 | spread |
|---|---|---|---|---|---|
| 0.00 | 0.1123 | 0.1217 | 0.1089 | 0.1228 | 0.0139 |
| 0.85 | 0.1126 | 0.1150 | 0.1277 | 0.1153 | 0.0151 |
| 0.95 | 0.1295 | 0.1489 | 0.1171 | 0.1483 | 0.0318 |
| 1.00 | 0.1695 | 0.1735 | 0.1702 | 0.1728 | 0.0040 |
| 1.02 | 0.1823 | 0.1842 | 0.1795 | 0.1828 | 0.0047 |
| 1.05 | 0.1821 | 0.1879 | 0.1821 | 0.1869 | 0.0058 |

Table 17: Outcome-frontier seed sensitivity on held-out Warsaw district series. Cells are equal-weight mean district-level worst-group CSD (lower is favorable); spread is the fitted-seed range and describes optimization sensitivity, not sampling uncertainty.

| Partition | Policy | Worst CSD | vs MES | $p$ | Wins |
|---|---|---|---|---|---|
| age × sex (8) | mes | 0.1992 | – | – | – |
| | res-1.0 | 0.1951 | – | – | – |
| | greedy-cost | 0.1991 | – | – | – |
| | learned-endowment | 0.1736 | -0.0255 | 0.0115 | 11/18 |
| | learned-outcome | 0.1728 | -0.0264 | 0.0130 | 14/18 |
| age (4) | mes | 0.1127 | – | – | – |
| | res-1.0 | 0.1106 | – | – | – |
| | greedy-cost | 0.1081 | – | – | – |
| | learned-endowment | 0.0907 | -0.0220 | 0.0079 | 14/18 |
| | learned-outcome | 0.0815 | -0.0312 | 0.0093 | 13/18 |
| sex (2) | mes | 0.0585 | – | – | – |
| | res-1.0 | 0.0558 | – | – | – |
| | greedy-cost | 0.0759 | – | – | – |
| | learned-endowment | 0.0352 | -0.0233 | $< 0.001$ | 14/18 |
| | learned-outcome | 0.0224 | -0.0361 | $< 0.001$ | 17/18 |

Table 18: Held-out Warsaw district series rescored under each group partition. "vs MES" is policy minus Equal Shares; negative values, paired exact sign-flip $p$ values, and wins use series. Age × sex policies are rescored without refitting, so rows are dependent robustness checks, not transfer evidence.

| Baseline | Learned CSD | Base CSD | $\Delta$ | 95% $\Delta$ interval | $p_{\text{fold}}$ | Series wins | Fold wins |
|---|---|---|---|---|---|---|---|
| greedy-cost | 0.1941 | 0.2161 | -0.0220 | [-0.0385, -0.0056] | 0.0625 | 13/19 | 5/5 |
| greedy-count | 0.1941 | 0.2466 | -0.0525 | [-0.0703, -0.0364] | 0.0625 | 19/19 | 5/5 |
| LLMRule count | 0.1941 | 0.2294 | -0.0353 | [-0.0495, -0.0223] | 0.0625 | 17/19 | 5/5 |
| LLMRule cost | 0.1941 | 0.2375 | -0.0434 | [-0.0593, -0.0293] | 0.0625 | 18/19 | 5/5 |
| MES | 0.1941 | 0.2242 | -0.0301 | [-0.0396, -0.0205] | 0.0625 | 17/19 | 5/5 |
| RES ($\lambda = 1$) | 0.1941 | 0.2225 | -0.0284 | [-0.0376, -0.0194] | 0.0625 | 18/19 | 5/5 |

Table 19: Five-fold out-of-district contrasts over 19 series, each scored by a map fitted on the other districts. This is not a temporal holdout and not cross-city validation: 18 of the 19 series are Warsaw districts. $\Delta$ is learned minus baseline; negative values favor learned. Paired-series bootstrap intervals are descriptive over 19 series. $p_{\text{fold}}$ is an exact sign-flip dependence sensitivity check on fold means (minimum two-sided resolution 0.0625). Overlapping training sets make folds dependent.

| Weight box | Learned | MES | Δ | 95% Δ interval | $p_{\text{fold}}$ | S-win | F-win | W/MES | Excl. |
|---|---|---|---|---|---|---|---|---|---|
| 10 (primary) | 0.1941 | 0.2242 | -0.0301 | [-0.0396, -0.0205] | 0.0625 | 17/19 | 5/5 | 0.984 | 0.0628 |
| 40 (post hoc) | 0.1886 | 0.2242 | -0.0355 | [-0.0477, -0.0235] | 0.0625 | 17/19 | 5/5 | 0.983 | 0.0645 |

Table 20: Post hoc fourfold box-expansion sensitivity on held-out series. Bound-10 is the primary analysis; bound 40 is post hoc. Δ is learned minus Equal Shares; negative favors learned. Intervals use paired series; $p_{\text{fold}}$ tests fold-mean signs under dependence. S-win and F-win count series and fold-mean wins. This is an outcome stability check, not coefficient identification or independent replication.

| Scope | Learned CSD | MES CSD | Δ | 95% Δ interval | $p$ | Wins |
|---|---|---|---|---|---|---|
| Gdynia | 0.1301 | 0.1375 | -0.0074 | [-0.0197, +0.0000] | 0.500 | 2/22 |
| Łódź | 0.1373 | 0.1390 | -0.0017 | [-0.0120, +0.0077] | 0.736 | 10/36 |
| Both cities | 0.1345 | 0.1384 | -0.0039 | [-0.0114, +0.0031] | 0.316 | 12/58 |
| *Actuation by instance granularity (whole prespecified grid)* | | | | | | |
| Min. projects/election | Series | Inert | Δ | 95% Δ interval | $p$ | Wins/losses |
| $\geq 0$ | 58 | 36 | -0.0039 | [-0.0114, +0.0031] | 0.316 | 12/10 |
| $\geq 5$ | 43 | 21 | -0.0052 | [-0.0155, +0.0040] | 0.315 | 12/10 |
| $\geq 10$ | 29 | 10 | -0.0026 | [-0.0155, +0.0089] | 0.682 | 10/9 |
| $\geq 15$ | 20 | 5 | -0.0022 | [-0.0102, +0.0056] | 0.605 | 7/8 |
| $\geq 20$ | 17 | 4 | -0.0024 | [-0.0118, +0.0069] | 0.640 | 6/7 |
| $\geq 25$ | 10 | 1 | -0.0007 | [-0.0117, +0.0101] | 0.969 | 4/5 |
| $\geq 30$ | 5 | 0 | +0.0015 | [-0.0151, +0.0205] | 0.938 | 3/2 |

Table 21: Native-approval Gdynia and Łódź series scored from 2023 with the frozen Warsaw map, without refitting or support-set projection. Δ is learned minus Equal Shares; negative favors learned. Paired-series bootstrap intervals summarize uncertainty. *Inert* means the Equal Shares allocation (median projects: 6.2 inert, 22.2 active). Effects stay near zero across the prespecified grid, so low actuation alone does not explain the null. $p$ uses exact sign flips when feasible and 200,000 draws otherwise. Two cities do not establish wider validity.

| Series | MES | Endowment | Δ | Direct | Δ |
|---|---|---|---|---|---|
| Żoliborz | 0.3083 | 0.1726 | -0.1357 | 0.3153 | **+0.0070** |
| Praga-Północ | 0.1953 | 0.1093 | -0.0860 | 0.1119 | -0.0833 |
| Wesoła | 0.2727 | 0.2080 | -0.0647 | 0.1359 | -0.1368 |
| Praga-Południe | 0.2333 | 0.1773 | -0.0560 | 0.1808 | -0.0524 |
| Wawer | 0.1580 | 0.1080 | -0.0500 | 0.1517 | -0.0062 |
| Ursynów | 0.0802 | 0.0461 | -0.0342 | 0.0684 | -0.0118 |
| Ursus | 0.2338 | 0.2062 | -0.0276 | 0.2010 | -0.0328 |
| Włochy | 0.1178 | 0.0942 | -0.0236 | 0.0696 | -0.0482 |
| Mokotów | 0.2917 | 0.2733 | -0.0184 | 0.2691 | -0.0226 |
| Wilanów | 0.2552 | 0.2410 | -0.0141 | 0.2656 | **+0.0104** |
| Wola | 0.1933 | 0.1813 | -0.0120 | 0.1501 | -0.0432 |
| Rembertów | 0.1665 | 0.1665 | +0.0000 | 0.1264 | -0.0401 |
| Śródmieście | 0.2624 | 0.2624 | +0.0000 | 0.2578 | -0.0046 |
| Białołęka | 0.1737 | 0.1764 | **+0.0027** | 0.2383 | **+0.0645** |
| Bemowo | 0.1583 | 0.1613 | **+0.0030** | 0.1185 | -0.0397 |
| Bielany | 0.1056 | 0.1087 | **+0.0031** | 0.0693 | -0.0363 |
| Ochota | 0.2091 | 0.2257 | **+0.0166** | 0.2131 | **+0.0040** |
| Targówek | 0.1697 | 0.2070 | **+0.0373** | 0.1676 | -0.0021 |

Table 22: Headline learned endowment and direct-outcome policies by held-out Warsaw district, ordered by endowment contrast. Each Δ is policy minus Equal Shares; negative favors learned and positive losses are bold. Rows are descriptive without intervals.