# When the Benchmark Moves with the Policy

## Measuring Access in Allocation Simulations

Pablo Loschi · Independent researcher, Berlin, Germany

Technical Note · Version 7.12 · 19 November 2026 · ORCID: 0009-0004-9455-4713

**Abstract**

Allocation policies alter who is served, who continues waiting, and who exits unserved, so evaluation ratios defined over those endogenous groups can move inversely with aggregate outcomes even when underlying behaviour is fixed. This Technical Note formalises moving-denominator and finite-horizon open-cohort distortions in discrete-event and agent-based allocation simulations and packages executable checks for routine use. First, Theorem 1 establishes exact necessary and sufficient conditions for both sign reversals against a baseline and pairwise ranking reversals between competing policies—proving that a lower-value policy can falsely overtake a higher-value policy even when both beat baseline—while Proposition 1 derives an exact finite-horizon censoring bias identity decomposing open-run unserved rates into completed-cohort access plus a late-arrival truncation residual. Two exact counterexamples (a two-period allocation and a bounded-delay deadline queue verified by exact dynamic programming across all 5,832 three-job configurations and independent transversal-matroid matching across 200 six-slot cases) verify reversals in both directions. Second, in a published liver-allocation simulator (LivSim, five synthetic replications), full regional sharing widens geographic dispersion in transplant MELD scores (15 of 15 replication-years) and new-listing transplant chances (5 of 5) with no robust cross-cohort reversal, yet holding status-quo regional denominators fixed shifts third-year mortality disparity by +5.5 percentage points (paired SE 1.5) despite national denominator stability below 0.2%. Third, in a canonical two-class priority-triage discrete-event benchmark across 50 matched common-random-number replications, non-blocking earliest-deadline/value dispatch (Policy B) beats an admission-gated urgent-priority rule (Policy A) by +3.11 percentage points (SE 0.03; 50/50) on a policy-invariant benchmark, yet scores 11.29 percentage points lower (SE 0.20; 50/50) on the moving benchmark; simultaneously, Policy A achieves a 3.17 percentage-point lower completed-cohort unserved rate (SE 0.09; 50/50) yet appears 3.31 percentage points worse (SE 0.13; 50/50) on open full-run accounting—inverting both axes of the policy frontier in all 50 pairs. Fourth, a controlled housing diagnostic confirms small negative oracle shifts (−0.049 to −0.005 pp) alongside moving-benchmark sign flips, 0.75×–2.5× cohort distortions, and discrete quota steps. A five-rule reporting contract and standard-library audit runner verify all 256 exported comparison records.

**Keywords:** simulation output analysis; performance measurement; discrete-event simulation; agent-based simulation; cohort accounting; simulation verification

**JEL classification:** C63; D47; D61; H23; R31

---

## 1. Introduction

Simulations of allocation policies typically evaluate interventions by reporting who was served and how efficiently resources were matched. In repeated allocation systems, however, both quantities are computed over pools that the policy itself reshapes. A rule that serves a high-valuation agent early removes that agent from the residual pool against which subsequent decisions are benchmarked; a rule that changes queueing duration alters which entrants resolve within a finite run horizon and which denominators appear in annual rate summaries. Consequently, a selection ratio computed against an endogenous pool can rise while total realised value falls, and competing instruments can change rank solely because access is counted over an open flow rather than a fixed entry cohort.

The central diagnostic is simple: whenever a policy alters the reference pool against which it is scored, a ratio can move inversely with aggregate allocation value. Isolating that distortion requires pairing aggregate realised value with explicit denominators, a policy-invariant reference benchmark, and a pre-specified entry cohort whose membership does not shift endogenously across policy arms. Following Sargent (2013) and Axtell and Farmer (2025, Section 3.8), this Technical Note combines exact necessary-and-sufficient reversal conditions and a finite-horizon censoring identity with a comparison-specific operationalisation of existing

reporting principles—an executable paired-comparison contract and verification suite—rather than proposing a new scheduling algorithm or replacement reporting standard.

The note assigns five distinct evidential roles to its components. Section 2 establishes Theorem 1 (necessary and sufficient conditions for baseline sign reversal and pairwise policy ranking reversal), Proposition 1 (the exact finite-horizon open-cohort censoring bias identity), two exact counterexamples verified by dynamic programming and transversal-matroid matching, and the executable audit contract. Section 3 pairs an empirical audit of LivSim (Kilambi, Bui and Mehrotra 2018) on its shipped synthetic inputs (Section 3.1, noting why an exploratory PolicySpace2 replay is quarantined in Supplement S1) with a high-power, self-contained 50-replication canonical two-class priority-triage discrete-event benchmark (`triage_benchmark.py`, Section 3.2) that demonstrates large, unmistakable (>3–11 pp, 50/50) pairwise reversals on both Theorem 1(ii) and Proposition 1 simultaneously. Section 4 uses a compact housing agent-based model strictly as a controlled diagnostic where the ground-truth fixed upper bound and completed entry cohorts are known, rather than as an empirical forecast. Section 5 distils a five-rule reporting standard, states scope and limitations, and concludes. Full proofs, the triage offered-load sensitivity grid, extended housing tables, ODD specifications, New York City proxies, and the quarantined PolicySpace2 record are preserved in the Online Supplement.

### 1.1 Worked audit from saved outputs

At the 6% offer discount in the housing stress-test sweep (Supplement S4, Table S4.2), the contention-linked duty raises the moving-reference ratio by a mean of +0.218 percentage points across eight matched replications. Pooled realised allocation value nevertheless falls by 0.352%, while the pooled moving reference falls by 0.573%. Measured against the policy-invariant upper bound, the mean fixed-reference ratio falls by 0.351 percentage points. The moving ratio rises while realised value falls in all eight saved pairs: the residual queue provides a smaller contemporaneous denominator, not a larger total allocation value.

Auditing saved outputs therefore requires reporting realised value and both references beside the ratios, while labelling pooled percentage changes separately from mean per-pair ratio changes. Running `python worked_audit.py` reconstructs all four quantities and the 8-of-8 sign-reversal count directly from `audit/benchmark_runs.json` without invoking a simulator. This calculation documents an accounting distortion in saved simulation outputs; it does not estimate an empirical housing-policy effect.

### 1.2 Related literature

The measurement problem links simulation reporting standards, online-benchmarking theory, and dynamic matching design. STRESS requires explicit definitions and calculation timing for quantitative outputs (item 1.2), initialisation details (4.1), and common random-number stream allocation in discrete-event models (5.2; Monks et al. 2019). ADEMP separates estimands from performance measures and recommends evaluating competing methods on identical simulated datasets (Morris, White and Crowther 2019, Sections 3.3–3.5 and 4.3), while ODD documents entity structure and scheduling order in agent-based models (Grimm et al. 2020). Following Sargent (2013) and Windrum, Fagiolo and Moneta (2007), the checks developed here address internal program and accounting verification rather than empirical validation for an intended policy use.

Outside simulation methodology, Kesselman et al. (2004) benchmark online buffer-management rules against an offline optimum on identical arrival sequences, Arora, Dekel and Tewari (2012) formalise policy regret against a fixed counterfactual history, Perdomo et al. (2020) study performative decision rules that reshape their own evaluation distribution, and Bertsimas, Farias and Trichakis (2011) anchor fairness losses to a common feasible optimum. In applied allocation settings, dynamic queueing and matching studies show how queue discipline, delay, and pool accumulation reshape waiting-list composition and intertemporal match quality (Bloch and Cantala 2017; Leshno 2022), while kidney-exchange and liver-allocation studies show that fragmentation, priority rules, and recipient-only metrics can yield metric-dependent policy verdicts (Gentry et al. 2013; Agarwal et al. 2019; Sönmez, Ünver and Yenmez 2020). This Technical Note translates those principles into a concrete paired-comparison record and executable verification suite.

## 2. Formal Reversal Conditions, Exact Counterexamples, and Executable Audit

### 2.1 Necessary-and-sufficient reversal conditions and censoring identity

For any allocation policy $p$ evaluated on a fixed exogenous problem instance (common arrivals, deadlines, capacities, and feasible assignment set), let $V_p > 0$ denote aggregate realised allocation value, $B_p > 0$ the sum of policy-dependent local references (such as the highest valuation present in the active pool at each decision), $E_p = V_p/B_p$ the moving-benchmark ratio, and $\Phi_p = V_p/F$ the fixed-benchmark ratio against any strictly positive policy-invariant reference $F > 0$ held equal across arms ($F_p = F$). Comparing any two policies $p, q$ obeys the exact log-ratio identity

$$\frac{E_p}{E_q} = \frac{V_p/V_q}{B_p/B_q}, \qquad \log\frac{E_p}{E_q} = \log\frac{V_p}{V_q} - \log\frac{B_p}{B_q}. \tag{1}$$

**Theorem 1 (Necessary and sufficient conditions for sign reversal and pairwise ranking reversal).** *Let policies be evaluated on a shared exogenous instance with $V_p, B_p, F > 0$.*

*(i) Sign reversal against baseline. Comparing policy 1 against baseline 0, a value-reducing policy ($V_1 < V_0$, equivalently $\Phi_1 < \Phi_0$) appears efficiency-improving under the moving benchmark ($E_1 > E_0$) if and only if*

$$\log\frac{B_0}{B_1} > \log\frac{V_0}{V_1} > 0.$$

*Symmetrically, a value-improving policy ($V_1 > V_0$) appears efficiency-reducing ($E_1 < E_0$) if and only if* $\log(B_1/B_0) > \log(V_1/V_0) > 0$.

*(ii) Pairwise policy ranking reversal. Comparing two competing policies $A$ and $B$ relative to a shared baseline 0, moving-benchmark evaluation strictly reverses the fixed-benchmark ranking—placing $A$ ahead of $B$ ($E_A > E_B$, equivalently $\Delta E_A > \Delta E_B$) while $B$ achieves strictly higher realised value ($V_A < V_B$, equivalently $\Phi_A < \Phi_B$)—if and only if*

$$\log\frac{B_B}{B_A} > \log\frac{V_B}{V_A} > 0.$$

*In particular, this pairwise reversal can hold even when both policies strictly outperform the baseline on realised value ($V_B > V_A > V_0$).*

*Intuition and proof.* Because $F$ cancels in $\Phi_p - \Phi_q = (V_p - V_q)/F$, fixed-reference rankings are always sign-faithful to realised value, whereas moving-benchmark ratios reward denominator shrinkage identically to value creation; whenever Policy $A$ compresses the contemporaneous reference pool more aggressively than Policy $B$ by a factor exceeding $V_B/V_A$, $A$ falsely overtakes $B$. Full step-by-step proofs (including matched-replication aggregation) appear in Supplement S3.1.

An analogous distortion governs finite-horizon access rates. Consider a simulation over horizon $t \in [0, T]$ with $N$ exogenous arrivals whose maximum patience or deadline horizon is $H < T$, so every entrant arriving at $t \le T - H$ resolves (served or unserved) by $T$. Partition all $N$ entrants into the completed entry cohort $\mathcal{C} = \{i : t_i \le T - H\}$ of size $N^{\mathrm{comp}}$ and the right-censored late-arrival window $\mathcal{L} = \{i : t_i > T - H\}$ with arrival-time weights $\omega_t = N_t/N$ and total late share $\Omega = \sum_{t>T-H} \omega_t = (N - N^{\mathrm{comp}})/N \in (0, 1)$, both policy-invariant under matched arrivals. Let $U_p^{\mathrm{open}}$ be the full-run unserved rate over all $N$ entrants, $U_p^{\mathrm{comp}}$ the completed-cohort unserved rate over $\mathcal{C}$, and $U_{p,t}^{\mathrm{cens}}$ the truncated unserved rate observed by $T$ among entrants arriving at $t > T - H$ (with weighted mean $\bar{U}_p^{\mathrm{cens}} = \Omega^{-1} \sum_{t>T-H} \omega_t U_{p,t}^{\mathrm{cens}}$).

**Proposition 1 (Finite-horizon open-cohort censoring bias identity).** *For every policy $p$ on matched exogenous arrivals,*

$$U_p^{\mathrm{open}} = U_p^{\mathrm{comp}} + \sum_{t>T-H} \omega_t \big(U_{p,t}^{\mathrm{cens}} - U_p^{\mathrm{comp}}\big) = U_p^{\mathrm{comp}} + \Omega\big(\bar{U}_p^{\mathrm{cens}} - U_p^{\mathrm{comp}}\big). \tag{2}$$

*Consequently, even when two policies deliver identical completed-cohort access ($U_A^{\mathrm{comp}} = U_B^{\mathrm{comp}}$), any difference in exit hazard timing within $(T - H, T]$ shifts $U_A^{\mathrm{open}} - U_B^{\mathrm{open}}$ by $\sum_{t>T-H} \omega_t (U_{A,t}^{\mathrm{cens}} - U_{B,t}^{\mathrm{cens}})$. Moreover, when Policy $A$ strictly lowers completed-cohort unserved exits ($U_A^{\mathrm{comp}} < U_B^{\mathrm{comp}}$), open full-run accounting reverses the ranking ($U_A^{\mathrm{open}} > U_B^{\mathrm{open}}$) if and only if*

$$\sum_{t>T-H} \omega_t \left( U_{A,t}^{\mathrm{cens}} - U_{B,t}^{\mathrm{cens}} \right) > (1-\Omega)(U_B^{\mathrm{comp}} - U_A^{\mathrm{comp}}) > 0, \qquad (3)$$

*equivalently* $\frac{\Omega}{1-\Omega}(\bar{U}_A^{\mathrm{cens}} - \bar{U}_B^{\mathrm{cens}}) > U_B^{\mathrm{comp}} - U_A^{\mathrm{comp}} > 0$.

*Intuition and proof.* Entrants still waiting at horizon $T$ remain in the denominator of $U^{\mathrm{open}}$ while contributing zero to its numerator. Policies that reject excess arrivals early (such as an admission gate) record late-window unserved exits immediately ($\bar{U}_A^{\mathrm{cens}} \approx U_A^{\mathrm{comp}}$), whereas non-blocking policies leave doomed late-window entrants right-censored in the end-of-horizon backlog ($\bar{U}_B^{\mathrm{cens}} \approx 0$), flipping the open-run ranking whenever the censoring wedge exceeds the true completed-cohort gap (full proof and backlog conservation identity in Supplement S3.1).

### 2.2 Exact counterexamples: two-period allocation and deadline queue

A minimal two-period allocation illustrates the downward-value reversal of Theorem 1(i) without strategic bidding, stochastic failures, or parameter uncertainty. One listing arrives in each of periods 1 and 2 and cannot be stored; each buyer can purchase at most once, and sellers select the highest effective offer (valuation multiplied by a certainty factor). Buyer I (an institution) is present only in period 1 with valuation 6 and certainty factor 1. Buyer O (a household) has valuation 10 and certainty factor 0.5 and remains present until served. Buyer L arrives in period 2 with valuation 1 and certainty factor 0.5.

Without intervention, I's effective offer of 6 beats O's offer of 5 in period 1, and O then beats L in period 2, yielding winners (I, O) and realised value $V_0 = 6 + 10 = 16$. Because O remains waiting after losing period 1, the highest valuation present is 10 in both periods, so $B_0 = 20$ and $E_0 = 16/20 = 0.80$ (double-counting O even though O can buy only once). Applying a 20% charge to I lowers I's effective offer to 4.8: O wins period 1, I expires, and L wins period 2, yielding winners (O, L), $V_1 = 10 + 1 = 11$, $B_1 = 10 + 1 = 11$, and $E_1 = 1.00$ (Table 1; $\log(B_0/B_1) = 0.598 > \log(V_0/V_1) = 0.375 > 0$). Enumerating all feasible schedules gives (I, O) = 16, (I, L) = 7, and (O, L) = 11, confirming that $F = 16$ is the exact feasible dynamic optimum.

| Evaluation | No charge | 20% charge | Change |
|---|---|---|---|
| Winners, periods 1 and 2 | I, O | O, L | Different allocation path |
| Realised value $V$ | 16 | 11 | −31.25% |
| Sum of highest present valuations $B$ | 20 | 11 | −45.00% |
| Ratio to highest present valuations $V/B$ | 0.80 | 1.00 | +20 percentage points |
| Ratio to feasible optimum ($F = 16$) | 1.00 | 0.6875 | −31.25 percentage points |

*Table 1. Exact two-period counterexample with fixed arrivals and two immediate listings. The local reference includes all present buyers before the charge. The feasible optimum respects arrival, expiry, and one purchase per buyer.*

The moving ratio still answers a well-defined local question: how winners' valuations compare with the highest valuation present at each decision along a policy's own path. That local selection statistic is informative when labelled as such, but misleads if read as aggregate value or proximity to a fixed optimum. A fixed upper bound $F$ avoids sign reversals under matched exogenous inputs and supplies a conservative efficiency floor (since $V \le V^* \le F$, reaching 99.5% of $F$ guarantees at least 99.5% of the feasible optimum $V^*$).

The opposite reversal in Theorem 1(i)—realised value rising while $V/B$ falls—arises in the bounded-delay unit-service scheduling problem of Kesselman et al. (2004). Jobs arrive with release slots, inclusive deadlines, and fixed values, and each slot serves at most one available job. Highest-value-first (HV) serves the largest currently available job; earliest-deadline-first (EDF) serves the earliest deadline, breaking ties by value and identifier. In a constructed three-slot example, job A has value 10 in slots 0–2, B has value 9 in slot 0 alone, C has value 8 in slot 1 alone, and D has value 1 in slot 2 alone. HV serves A, C, and D for $V = 19$ and $B = 19$ ($V/B = 1.00$), whereas EDF defers A to slot 2 and serves B, C, and A for $V = 27$, achieving the exact offline optimum while retaining A in the pool across all three slots so that $B = 30$ and $V/B = 0.90$ (Table 2; $\log(B_1/B_0) = 0.457 > \log(V_1/V_0) = 0.351 > 0$).

| Policy | Service order | Realised value $V$ | Moving reference $B$ | $V/B$ | $V$/feasible optimum |
|---|---|---|---|---|---|
| Highest value first (HV) | A, C, D | 19 | 19 | 1.00 | 0.704 |
| Earliest deadline first (EDF) | B, C, A | 27 | 30 | 0.90 | 1.000 |

*Table 2. Constructed three-slot deadline queue example. The exact offline optimum is 27. Relative to HV, EDF raises value by 42.11% and the moving reference by 57.89%, lowering $V/B$ by 10 percentage points.*

Across 200 generated six-slot cases (two jobs per slot, integer values uniform on $\{1, ..., 20\}$, deadline extensions uniform on $\{0, 1, 2\}$ truncated at slot 5; seeds `20261006`–`20261205`), EDF raises realised value while lowering $V/B$ in 15 cases, agrees in nonzero sign in 162 cases, ties in 23 cases, and exhibits zero falling-value/rising-ratio cases (since HV always achieves $V/B = 1$ in this design). Mean value-to-optimum ratios are 0.9869 for HV and 0.8940 for EDF, showing that EDF's success in Table 2 is an accounting counterexample rather than a claim of general scheduler superiority. Two independent verification checks confirm the offline optimum: exact dynamic programming matches complete schedule enumeration across all **5,832** three-job, three-slot configurations in $\{1, 2, 3\}$, and descending-weight greedy selection with augmenting-path bipartite matching on the transversal matroid of schedulable subsets (Edmonds 1971; `deadline_matching_check.py`) confirms exact equality on all 200 saved cases (Supplement S3.5; largest Eq. (1) log residual $2.05 \times 10^{-16}$).

### 2.3 Compact executable audit contract

To operationalise these diagnostics across simulators, the accompanying package separates three tasks: (i) reconstructing model allocations and traces, (ii) establishing the comparison population, and (iii) auditing the resulting scalar totals via `compare_benchmarks.py` and `audit_contract.py`. The scalar runner inspects only declared metadata and totals; model-specific exporters must establish that both arms share the same exogenous problem and population (Table 3).

| Record component | Required content | Verification responsibility |
|---|---|---|
| Pair identity | Scenario, policy, and replication identifiers; horizon | Exporter labels the comparison |
| Common problem | Input hash; hash of exogenous feasibility constraints (excluding policy eligibility and decision rules); objective and units | Model-specific evidence establishes meaning; runner rejects unequal declared metadata |
| References | $B_0$ and $B_1$ definitions; equal $F_0 = F_1 > 0$; label as upper bound or feasible optimum | Exporter justifies the reference; comparator checks finite nonnegative totals |
| Population | Entry rule, cohort identifier, follow-up endpoint, and unresolved counts (if a completed cohort is claimed) | Event-level exporter verifies membership; runner rejects declared incomplete or mismatched cohorts |
| Aggregation | Per-pair ratios, pooled totals, or explicitly weighted summary | Analyst selects estimand before combining runs |
| Output | Value and reference changes, both ratio changes, reversal flag, and log residual | Scalar comparator; undefined ratios remain null |

*Table 3. Paired comparison record and verification responsibilities enforced by `audit_contract.py`. Matching declared metadata is a consistency condition; model-specific trace and membership checks establish its meaning.*

Running the executable audit and unit-test suite requires only the Python standard library:

```
python worked_audit.py
python audit_contract.py example_pair.json
python triage_benchmark.py
python -m unittest -v test_comparison_contract.py
```

On `example_pair.json`, the runner confirms the exact Table 1 changes (−31.25% realised value, +20 pp moving ratio, −31.25 pp fixed-reference ratio, opposite-sign flag true) and checks that any defined log residual lies within $10^{-12}$. The unit-test suite in `test_comparison_contract.py` exercises mismatched input/

constraint hashes, objective/unit/horizon/population mismatches, unequal or zero fixed references, negative/NaN/infinite totals, undefined moving ratios, incomplete or mismatched completed-cohort claims, and the canonical triage DES benchmark's exact Theorem 1(ii) and Proposition 1 checks. Using `export_records.py` and `triage_benchmark.py` with canonical UTF-8 sorted-key JSON payloads (Supplement S3.5), the unchanged `audit_contract.py` runner passes all **256** exported records—the exact two-period pair, all 200 saved deadline pairs, all five fixed-horizon LivSim incident-cohort pairs, and all 50 canonical triage DES pairs (which also exercise `audit_contract.py`'s completed-cohort check with `unresolved = 0`)—with zero failures.

**Exporter responsibilities and interpretation.** Because `audit_contract.py` inspects only declared metadata and scalar totals, the exporter carries three substantive obligations. First, the exogenous-feasibility fingerprint must cover common arrivals, expiry, capacity, and one-purchase or one-service limits while excluding policy eligibility, sharing, and decision rules; matching random-seed integers alone is insufficient whenever policy-dependent failures or relistings consume RNG draws differently across arms. Second, population claims must match the observation design: a completed-exit cohort requires every entrant to buy or depart by the endpoint (zero unresolved), whereas a fixed-duration clinical cohort can retain still-waiting patients at its horizon as complete fixed-duration observations without asserting completed exits. Third, because equation (1) holds within each matched pair, mean-of-ratios summaries and ratios-of-pooled-totals use different weights and must be declared separately alongside raw totals so any sign reversal can be checked without rerunning the simulator. Matching declared hashes never validates model realism; it enforces accounting transparency on saved outputs.

---

# 3. External Empirical Audit and Canonical Triage DES Benchmark

## 3.1 External empirical audit: LivSim

LivSim (Kilambi, Bui and Mehrotra 2018) is an open-source Python discrete-event simulator of U.S. liver allocation distributed with randomly generated synthetic inputs across 58 donor service areas (DSAs). It reports geographic disparity using two standard summaries: the cross-DSA standard deviation of median allocation MELD scores among each year's transplant recipients, and the cross-DSA standard deviation of annual mortality rates whose denominator equals candidates waiting at year start plus new arrivals during that year. Neither summary tracks a fixed patient cohort by construction.

I compare the shipped status quo (Share 35 and Share 15 with MELD-Na) against full regional sharing across three simulated years and all five paired synthetic replications, using identical synthetic patient inputs and exogenous disease histories with only input-path edits, a one-line NumPy 2 compatibility update, and per-patient event logging. No human-subject data were collected; all records are synthetic. Event-level checks confirm exact equality of patient identifiers, service areas, and entry times across paired policies, supporting two fixed one-year follow-up cohorts: the 15,406 candidates waiting at simulation start (49 DSAs with $\geq 20$ members) and the 11,571 patients listed during year one (50 DSAs with $\geq 20$ members). Table 4 summarises the resulting comparisons.

| Spread across donor service areas | Population counted | Change under regional sharing |
|---|---|---|
| Median MELD at transplant, years 1, 2, and 3 | Patients transplanted that year | +5%, +17%, +20% (wider in 15 of 15) |
| Mortality rate, year 1 | Candidates at start plus year's arrivals | −7% (narrower in 5 of 5) |
| Mortality rate, years 2 and 3 | The same, recounted each year | +8%, +1% (wider in 3 of 5, 3 of 5) |
| One-year transplant chance | First-year listings (11,571) | +10% (wider in 5 of 5) |
| One-year death risk | First-year listings | +4% (wider in 3 of 5) |
| One-year transplant chance | Waiting at start (15,406) | −1% (narrower in 3 of 5) |
| One-year death risk | Waiting at start | −7% (narrower in 4 of 5) |

*Table 4. LivSim, full regional sharing against status quo: mean percentage change in the standard deviation across donor service areas over five paired synthetic replications, with counts of replications or replication-years sharing the sign. Rows 1–3 are LivSim's published outputs; cohort rows include DSAs with at least 20 cohort members.*

Two findings emerge, both requiring explicit qualification. First, population choice and replication stability interact sharply. Regional sharing consistently widens MELD-at-transplant dispersion (15 of 15 replication-years; means +4.8%, +16.9%, +20.3%) and one-year transplant-chance dispersion among first-year listings (5 of 5 replications; mean +10.4%, paired SE 1.6, leave-one-out range [+9.1%, +11.5%]). Among initial waiters, however, the mean narrowing in transplant-chance dispersion is small (−0.8%, paired SE 2.6) and crosses zero under leave-one-out omission ([−2.7%, +1.2%]), as does third-year mortality disparity (+1.1%, paired SE 7.8, leave-one-out range [−5.5%, +5.5%]; full leave-one-out table in Supplement S2, Table S2.1). These five synthetic replications therefore establish **no robust cross-cohort sign reversal**; rather, they show that recipient-only metrics, annual candidate flows, and adjacent entry cohorts answer distinct questions with differing Monte Carlo stability. Behind the recipient metrics, served-pool membership turns over substantially: in replication 1, 909 of the 5,106 first-year listings transplanted within one year under the status quo are not transplanted within one year under regional sharing, while 922 others are (4,197 shared).

Second, holding area-level mortality denominators common alters annual disparity magnitudes even when the national denominator barely moves (changing by less than 0.2% in every pair-year). Reconstructing each DSA's annual death count and candidate-flow denominator reproduces all 30 published mortality-disparity figures exactly, allowing both policies' death counts to be normalised by a single common denominator (Table 5).

| Year | Own denominators (%) | Common status-quo denominators (%) | Common sharing denominators (%) | DSA denominator change: min to max (%) |
|---|---|---|---|---|
| 1 | −7.5 (2.3) | −7.5 (2.3) | −7.5 (2.3) | [+0.0, +0.0] |
| 2 | +7.9 (9.5) | +10.8 (11.0) | +7.6 (10.2) | [−8.8, +13.7] |
| 3 | +1.1 (7.8) | +6.6 (8.3) | +1.3 (7.0) | [−13.3, +17.5] |

*Table 5. LivSim annual mortality-disparity accounting sensitivity over five paired replications. Columns 2–4 report mean percentage changes (paired SEs) in the cross-DSA standard deviation of normalised death counts under sharing vs. status quo. Area ranges span all pair-years; national denominators shift by <0.2% throughout. Common-denominator normalisations are accounting sensitivities, not fixed-cohort mortality risks.*

Because year-one denominators are identical at initialisation, all three columns agree in year 1. By year 3, individual DSA denominators diverge by −13.3% to +17.5%. Applying common status-quo denominators raises the mean third-year disparity increase from +1.1% to +6.6%, a within-pair shift of **+5.5 percentage points** (paired SE 1.5; positive in 5 of 5 pairs), whereas common sharing denominators yield +1.3% (a within-pair shift of +0.2 percentage points, paired SE 1.0). Near-constant national totals thus cannot rule out substantial sub-regional denominator drift. In the accompanying contract export (`export_records.py`), the five LivSim pairs are additionally audited as one-year incident-cohort transplant counts with event-level equality of patient IDs, service areas, and entry times and an explicit upper-bound count reference equal to the cohort size (ignoring organ capacity, not a feasible clinical optimum); cross-DSA dispersion and mortality disparities require the separate model-specific scripts above. Because inputs are synthetic and span three years across five pairs, these recounts carry no clinical policy inference.

**Quarantined PolicySpace2 extension.** An exploratory audit of PolicySpace2 (Furtado 2022) was excluded from the main evidence and quarantined in Supplement S1. Although archived summaries suggested large discrepancies between monthly flow counts, unique assisted families, and entry-cohort poverty persistence, exhaustive archive checks found that the 40 original monthly log files are unavailable, and controlled container replays of unmodified upstream commit `ced77dd` demonstrated three structural sources of non-determinism (Python `PYTHONHASHSEED` set-iteration order, unseeded `.agents` cache loading, and out-of-stream `uuid.uuid4()` family IDs). Accordingly, PolicySpace2 plays no evidential role in this paper's conclusions.

### 3.2 Canonical multi-class priority-triage DES benchmark

Whereas LivSim supplies five synthetic replications with no robust cross-cohort reversal and the housing diagnostic in Section 4 produces small oracle shifts, Theorem 1(ii) and Proposition 1 predict that large, unmistakable pairwise policy ranking reversals can arise simultaneously on both efficiency and access in standard congested queueing systems. To test both predictions at high statistical power, `triage_benchmark.py` implements a self-contained, deterministic Python standard-library discrete-event simulation of a canonical

**two-class capacity-constrained priority-triage / waiting-list queue with deadlines** (explicitly a synthetic canonical DES benchmark rather than third-party software).

In each discrete period $t \in \{0, ..., 119\}$ ($T = 120$), $c = 10$ service slots allocate capacity among Poisson arrivals at offered load $\rho = \lambda/c = 1.35$ ($\lambda = 13.5$ requests per period), split into **Urgent** requests ($p_U = 0.60$, valuation $v_U \sim U[60, 140]$, tight deadline $d_U \sim U\{1, 2, 3, 4\}$ periods) and **Routine** requests ($1 - p_U = 0.40$, valuation $v_R \sim U[20, 60]$, patience $d_R = 30$ periods). The maximum deadline $H = 30$ partitions all arrivals (mean $N = 1,632.1$) into an exact completed entry cohort $\mathcal{C} = \{t \leq 90\}$ (mean $N^{\text{comp}} = 1,242.4$, zero right-censoring) and a late-arrival window $\mathcal{L} = \{t > 90\}$ ($\Omega = 0.2388$). Using common random numbers across **50 matched replications** (seeds `20261118`–`20261167`), we compare three standard operational rules:

- **Baseline (FCFS):** Non-blocking first-come, first-served dispatch.
- **Policy A (Admission Gate + Urgent HV / Routine FIFO):** Protects urgent service and caps routine backlog at $K = 60$, immediately rejecting incoming routine arrivals at the gate whenever 60 routine requests are already waiting; dispatches urgent requests by highest valuation first (HV) and admitted routine requests in FIFO order.
- **Policy B (Non-Blocking Urgent EDF / Routine HV):** Never blocks arrivals at the door; dispatches urgent requests by earliest deadline first (EDF, value tie-break) and assigns remaining slots to routine requests by highest valuation first (HV).

All three policies share the identical exogenous policy-invariant upper bound $F$ (sum of the $T \cdot c = 1,200$ highest arrival valuations; mean $109,572.9$) and the identical moving-reference definition $B_p$ (sum over allocated slots of the highest valuation present in the active pool). Table 6 reports the 50-replication results.

| Policy / Comparison | Fixed $\Phi = V/F$ (%) | Moving $E = V/B$ (%) | Completed $U^{\text{comp}}$ (%) | Open $U^{\text{open}}$ (%) | Late $\bar{U}^{\text{cens}}$ (%) |
|---|---|---|---|---|---|
| Baseline (FCFS) | 74.32 (0.24) | 50.56 (0.16) | 25.38 (0.24) | 25.29 (0.22) | 24.91 (0.46) |
| Policy A (Gate + Urgent HV / R-FIFO) | 96.71 (0.04) | 95.83 (0.07) | 22.59 (0.28) | 22.92 (0.24) | 23.87 (0.43) |
| Policy B (Non-blocking U-EDF / R-HV) | 99.82 (0.02) | 84.54 (0.15) | 25.76 (0.24) | 19.61 (0.20) | 0.00 (0.00) |
| **Policy B − Policy A (pp; SE)** | **+3.11 (0.03)** | **−11.29 (0.20)** | **+3.17 (0.09)** | **−3.31 (0.13)** | **−23.87 (0.43)** |
| *50-rep CRN sign agreement* | *50/50 (B leads)* | *50/50 (A leads)* | *50/50 (A leads)* | *50/50 (B leads)* | *50/50* |

*Table 6. Canonical two-class priority-triage DES benchmark at offered load $\rho = 1.35$ across 50 matched CRN replications (mean and paired SE in parentheses; `triage_benchmark.py`). Lower U denotes higher access ($1 - U$). Both Policy A and Policy B outperform FCFS on fixed value by +22.39 and +25.51 pp (SE 0.24; 50/50).*

Both formal mechanisms fire at full statistical power (**50 of 50** matched replications) and invert **both axes of the policy frontier simultaneously**:

1. **Efficiency axis — Theorem 1(ii) pairwise ranking reversal.** Although both policies substantially outperform FCFS on fixed value (+22.39 and +25.51 pp; 50/50), Policy B delivers **+3.11 pp** higher fixed-benchmark value than Policy A (SE 0.03; 50/50) by rescuing imminent-deadline urgent requests and selecting higher-value routine requests from the unblocked queue. On the moving benchmark, however, Policy A scores **+11.29 pp** higher than Policy B (SE 0.20; 50/50) because clearing highest-value urgent requests immediately shrinks $B_A$ (mean $110,578$) relative to $B_B$ (mean $129,424$), where deferred flexible urgent requests stay in the pool across slots. The exact condition of Theorem 1(ii), $\log(B_B/B_A) = 0.15714 > \log(V_B/V_A) = 0.03170 > 0$ holds in **50/50** pairs (max Eq. (1) residual $2.50 \times 10^{-16}$), and all 50 exported records (`contract_records/triage_000.json`–`triage_049.json`) pass `audit_contract.py` with `opposite_signs = true` and `unresolved = 0`.

2. **Access axis — Proposition 1 cohort-vs-open ranking reversal.** On the uncorrupted completed entry cohort ($\mathcal{C}$, $t \leq 90$), Policy A leaves **3.17 pp fewer** entrants unserved than Policy B (22.59% vs. 25.76%, SE 0.09; 50/50) because its finite FIFO routine buffer drains completed-cohort entrants during $t > 90$ whereas Policy B skims newly arriving high-value routine requests. On the open full-run rate, however, Policy B appears to leave **3.31 pp fewer** entrants unserved (19.61% vs. 22.92%, SE 0.13; 50/50). Proposition 1 explains the flip algebraically: Policy A's admission gate immediately logs rejected late-window arrivals as unserved ($\bar{U}_A^{\text{cens}} = 23.87\%$, end-of-horizon backlog 57.8), whereas Policy B holds unserved late-window routine arrivals right-censored at horizon $T$ ($\bar{U}_B^{\text{cens}} = 0.00\%$, backlog 112.0). The resulting censoring wedge $\frac{\Omega}{1-\Omega}(\bar{U}_A^{\text{cens}} - \bar{U}_B^{\text{cens}}) = 7.53$ pp (SE 0.18) exceeds the true completed-cohort gap 3.17 pp in **50/50** pairs (max Eq. (2) residual $5.55 \times 10^{-17}$).

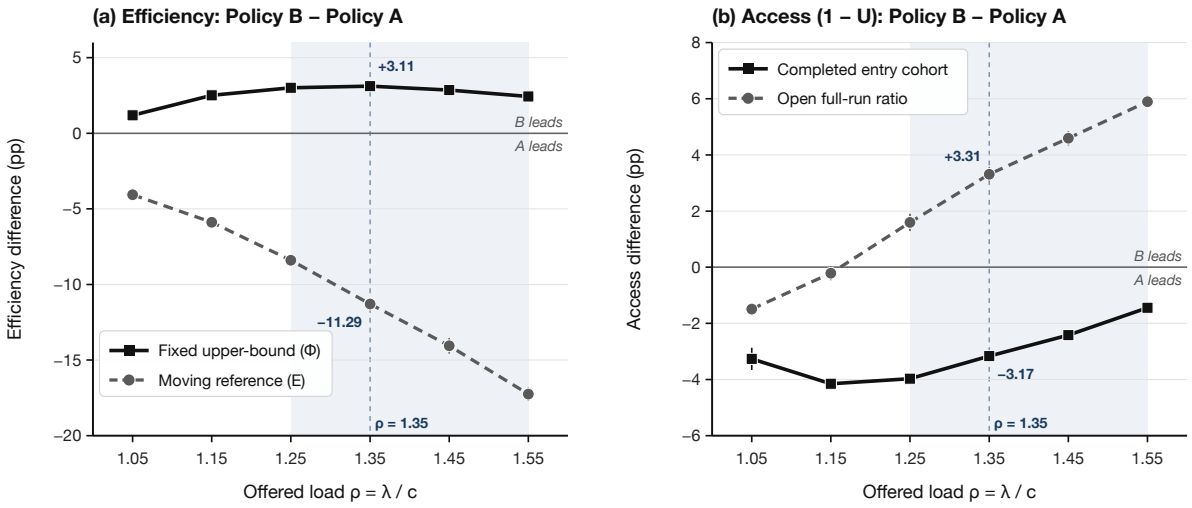


Figure 1. Dual policy-ranking reversals in the canonical two-class triage DES benchmark across offered loads $\rho \in [1.05, 1.55]$ (50 matched CRN replications per node; error bars show $\pm 1.96$ paired SEs). Both panels plot Policy B minus Policy A so values above zero favour B and values below zero favour A. Panel (a) verifies Theorem 1(ii) at every load (50/50 throughout); Panel (b) verifies Proposition 1 as congestion binds ($\rho \geq 1.25$; shaded dual-reversal region, 50/50 at canonical $\rho = 1.35$).

Figure 1 and Supplement Table S3.1 confirm that the dual reversal is not a knife-edge artefact: the efficiency reversal of Theorem 1(ii) holds in **50/50** replications at every tested load from $\rho = 1.05$ to $\rho = 1.55$ (+1.19 to +3.11 pp fixed vs. −4.06 to −17.25 pp moving), while Proposition 1′s access reversal emerges monotonically once the admission gate binds under congestion (0/50 at $\rho = 1.05$, 17/50 at $\rho = 1.15$, 46/50 at $\rho = 1.25$, 50/50 at $\rho = 1.35$ and 1.45, and 49/50 at $\rho = 1.55$). Under sound accounting, the two policies span a genuine efficiency–access trade-off (Policy B leads on value; Policy A leads on cohort access); naive accounting flips both axes at once.

## 4. Controlled Housing ABM Diagnostic

To examine benchmark drift, open-cohort truncation, and discrete quota accounting simultaneously in a setting where the exact exogenous arrivals, fixed upper-bound oracle, and completed entry cohorts are fully observable, I use a compact housing agent-based model strictly as a **controlled diagnostic** (in the sense of Edmonds et al. 2019). Offers, buyer valuations, patience, and supply are held fixed without equilibrium bidding or strategic entry, so the model produces neither a forecast nor a policy ranking for New York City.

The main-market specification simulates five independent districts over 120 periods with Poisson listings (mean 20 per district-period) and civilian arrivals (17 owner-occupiers, 4 small investors), lognormal valuations, and finite civilian patience (36 periods for owner-occupiers, 24 for small investors; full ODD specification and parameters in Supplements S3 and S5). Sellers select the highest eligible offer weighted by certainty factor 1 for three institutional entities, $c_{\text{OO}} = 1 - w$ for owner-occupiers, and $c_{\text{SI}} = \min(1, c_{\text{OO}} + 0.07)$ for small investors. One institutional bid per entity per district-period yields an unconstrained institutional purchase share of 11.0% alongside a baseline offer discount $w = 1.5\%$, chosen solely as unit-scale proxies from linked New York City house transactions (19,019 of 166,227 eligible house sales, or 11.4%, involve a company cash buyer,

with a company-cash price gap of 1.53 log points; Loschi 2026a, 2026d; Supplement S5). Five stylized rules generate contrasting allocation paths on identical arrivals: a static per-entity acquisition cap, a work-conserving period-reset fair share, a carry-forward fair share, a contention-triggered duty scaling institutional offers by $(1 - \tau_{\max} s)$ using trailing six-period resolved unserved exits ($\tau_{\max} = 0.22$ baseline), and a constant institutional charge.

### 4.1 Fixed versus moving benchmark divergence

Three value summaries are compared under matched random inputs: the realised-value index (unconstrained = 100), **oracle efficiency** $V/F$ dividing realised value by the policy-invariant sum $F$ of the $L$ highest generated valuations across all $L$ listings ($V \leq V^* \leq F$), and **contemporaneous efficiency** $V/B$ dividing $V$ by the sum of highest valuations present in the pool at each listing. In the main market, $B$ averages roughly 444,000 per unit under the unconstrained baseline and contention duty versus 463,000 under the static cap, mixing allocation-value shifts with residual-queue shifts.

Table 7 reports the offer-discount sweep for the contention duty across $w \in [0\%, 20\%]$. **Explicitly, the oracle value shifts are very small and strictly negative at every tested discount**, ranging from **−0.049 to −0.005 percentage points** (value index 99.95 to 99.99; remaining negative in 3,943 of 4,000 bootstrap resamples of the eight replications). Yet under the moving contemporaneous benchmark, the same duty appears to generate a positive efficiency gain above $w \approx 2.2\%$ (95% interval [2.1, 2.3]), reaching +0.342 percentage points at $w = 20\%$. Thus the two benchmarks disagree on sign across the entire $4\%$ – $20\%$ interval—including the 5.3% upper-bound and 8.8% headline New York proxy discounts (Supplement S5)—purely because removing high-valuation bidders compresses subsequent contemporaneous maxima.

| Offer discount $w$ | Inst. share | Gain, oracle (pp) | Gain, contemporaneous (pp) | Value index, duty |
|---|---|---|---|---|
| 0% | 10.7% | −0.049 | −0.088 | 99.95 |
| 2% | 11.0% | −0.045 | −0.006 | 99.95 |
| 4% | 11.3% | −0.040 | +0.069 | 99.96 |
| 6% | 11.6% | −0.036 | +0.128 | 99.96 |
| 8% | 11.9% | −0.031 | +0.174 | 99.97 |
| 10% | 12.3% | −0.028 | +0.223 | 99.97 |
| 12% | 12.6% | −0.025 | +0.257 | 99.98 |
| 16% | 13.2% | −0.015 | +0.305 | 99.98 |
| 20% | 13.8% | −0.005 | +0.342 | 99.99 |

*Table 7. Main-market contention duty ($\tau_{\max} = 0.22$) versus unconstrained baseline across owner-occupier offer discounts (8 matched replications per row; SEs of gains $< 0.008$ pp). Oracle efficiency uses the policy-invariant upper bound $F$; contemporaneous efficiency uses the moving reference $B$.*

Figure 2 illustrates the same numerator–denominator mechanism in the high-participation stress-test sweep (58.1% unconstrained institutional share; full sweep and break-even schedule in Supplement S4, Tables S4.1–S4.2). Panel (a) contrasts the fixed upper-bound ratio against the moving ratio, while Panel (b) separates pooled realised value $V$ from the pooled moving reference $B$. Across the shaded 6%–10% region—including the 6% worked audit of Section 1.1—all eight matched seeds per node show opposite signs (33 of 72 paired scenarios overall), shifting the apparent break-even discount from 11.7% (95% CI [11.5, 11.8]) on the fixed oracle down to 5.4% ([5.4, 5.5]) on the moving benchmark.

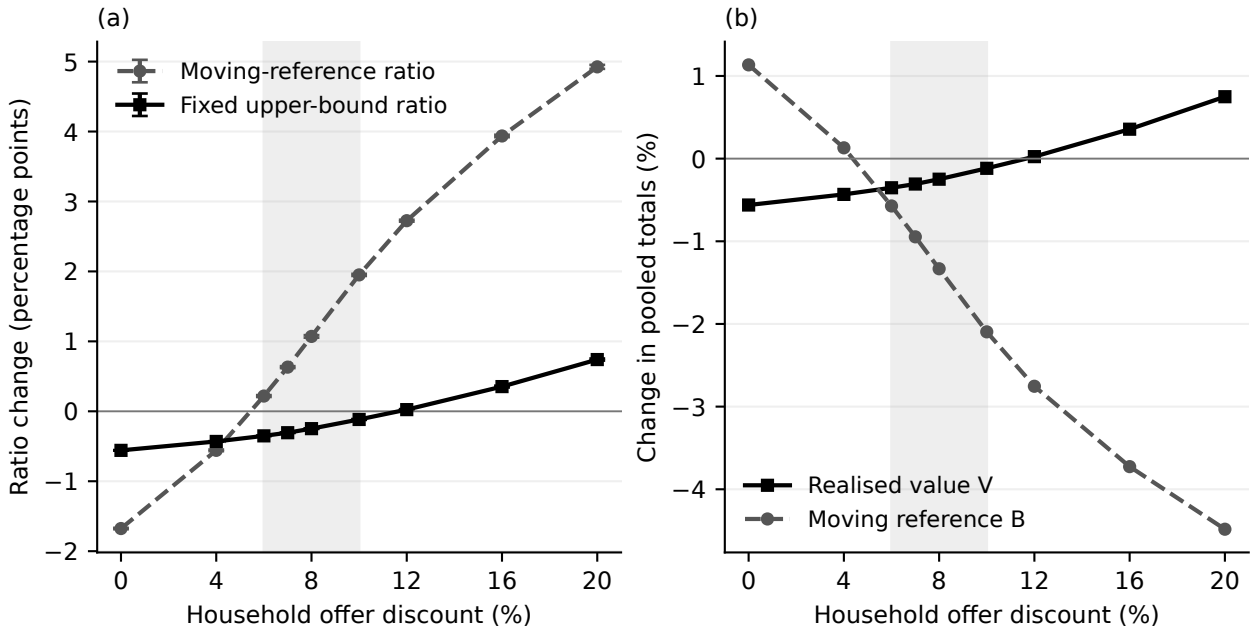


Figure 2. Decomposition of moving-benchmark sign reversals in the stress-test offer-discount sweep (8 matched seeds per node). Panel (a) plots matched-seed mean ratio changes (±1 paired Monte Carlo SE); Panel (b) plots percentage changes in pooled realised value $V$ and pooled moving reference $B$. Across the shaded 6%–10% band, $B$ falls faster than $V$, turning modest value losses into apparent moving-ratio gains.

### 4.2 Open cohorts, completed entry cohorts, and quota accounting

Access comparisons exhibit an analogous denominator distortion governed by Proposition 1. **Full-run starvation** divides unserved owner-occupier departures by all owner-occupier arrivals over periods 0–119, keeping buyers still searching at period 120 in the denominator while omitting their eventual outcomes from the numerator. A **completed entry cohort** instead tracks every owner-occupier entering in periods 36–83 until purchase or patience expiry by period 119, resolving 100% of cohort members. Without intervention on the held-out evaluation seeds, 15.1% of the completed cohort leaves unserved compared with a 10.1% full-run rate (14.87% vs. 9.83% on the 12 baseline seeds).

Table 8 evaluates all five instruments tuned on four training seeds to common full-run access gains of 0.25, 0.5, and 1 percentage point and scored on 16 held-out seeds. Across these settings, completed-cohort gains range from **0.75**× the full-run gain (carry-forward fair share at 0.25 pp) up to **2.5**× (static cap across all three targets) and reorder instruments. At the 0.5 pp target, the static cap and constant charge share the same 0.52 pp full-run gain, yet deliver 1.29 pp versus 0.75 pp completed-cohort gain. Between the two price rules at the 0.25 and 0.5 pp targets, the constant charge retains +0.021 (paired SE 0.0017) and +0.037 (paired SE 0.0044) more value-index points than the duty, but delivers 0.29 and 0.48 pp less cohort access (paired SE 0.028 and 0.039). Retuning directly to completed-cohort targets on 16 further disjoint seeds reverses the charge value comparison: the duty retains +0.018 (paired SE 0.0020) and +0.027 (paired SE 0.0052) more value-index points at 0.5 and 1 pp cohort targets while the constant charge yields +0.13 and +0.23 pp more full-run access (Supplement S4, Table S4.3). Each rule retains slightly more value on the population for which it was tuned, confirming that rankings do not transfer across denominators.

| Target (pp) | Instrument | Setting | Full-run gain (pp; SE) | Cohort gain (pp; SE) | Value index (SE) |
|---|---|---|---|---|---|
| 0.25 | Static cap | 73 | 0.28 (0.023) | 0.71 (0.057) | 99.50 (0.017) |
| 0.25 | Fair share, period reset* | 0.1172 | 0.19 (0.014) | 0.25 (0.021) | 99.94 (0.004) |
| 0.25 | Contention duty | 0.3006 | 0.28 (0.010) | 0.64 (0.029) | 99.93 (0.002) |
| 0.25 | Constant charge | 0.0213 | 0.25 (0.008) | 0.34 (0.016) | 99.95 (0.002) |
| 0.25 | Fair share, carry-forward | 0.1113 | 0.21 (0.016) | 0.15 (0.025) | 99.95 (0.004) |
| 0.5 | Static cap | 68 | 0.52 (0.035) | 1.29 (0.087) | 99.32 (0.018) |
| 0.5 | Fair share, period reset* | 0.0977 | 0.55 (0.022) | 0.74 (0.037) | 99.85 (0.004) |
| 0.5 | Contention duty | 0.5641 | 0.56 (0.017) | 1.24 (0.039) | 99.85 (0.003) |
| 0.5 | Constant charge | 0.0424 | 0.52 (0.013) | 0.75 (0.026) | 99.89 (0.003) |
| 0.5 | Fair share, carry-forward | 0.1036 | 0.52 (0.024) | 0.61 (0.045) | 99.85 (0.006) |
| 1 | Static cap | 60 | 1.03 (0.042) | 2.56 (0.105) | 99.01 (0.021) |
| 1 | Fair share, period reset | 0.0859 | 1.00 (0.017) | 1.37 (0.029) | 99.74 (0.005) |
| 1 | Contention duty | – | not reachable (at most 0.88) | – | – |
| 1 | Constant charge | 0.0802 | 1.04 (0.019) | 1.56 (0.036) | 99.76 (0.004) |
| 1 | Fair share, carry-forward | 0.0946 | 1.05 (0.032) | 1.38 (0.064) | 99.70 (0.007) |

*Table 8. Main-market instruments tuned on 4 seeds to common full-run access targets and evaluated on 16 held-out seeds. Gains are reductions in full-run and completed-cohort unserved rates. Asterisks mark period-reset fair-share rules that could not meet the tuning tolerance (*$\max(0.02 \text{ pp}, 10\%)$*) because of integer rounding.*

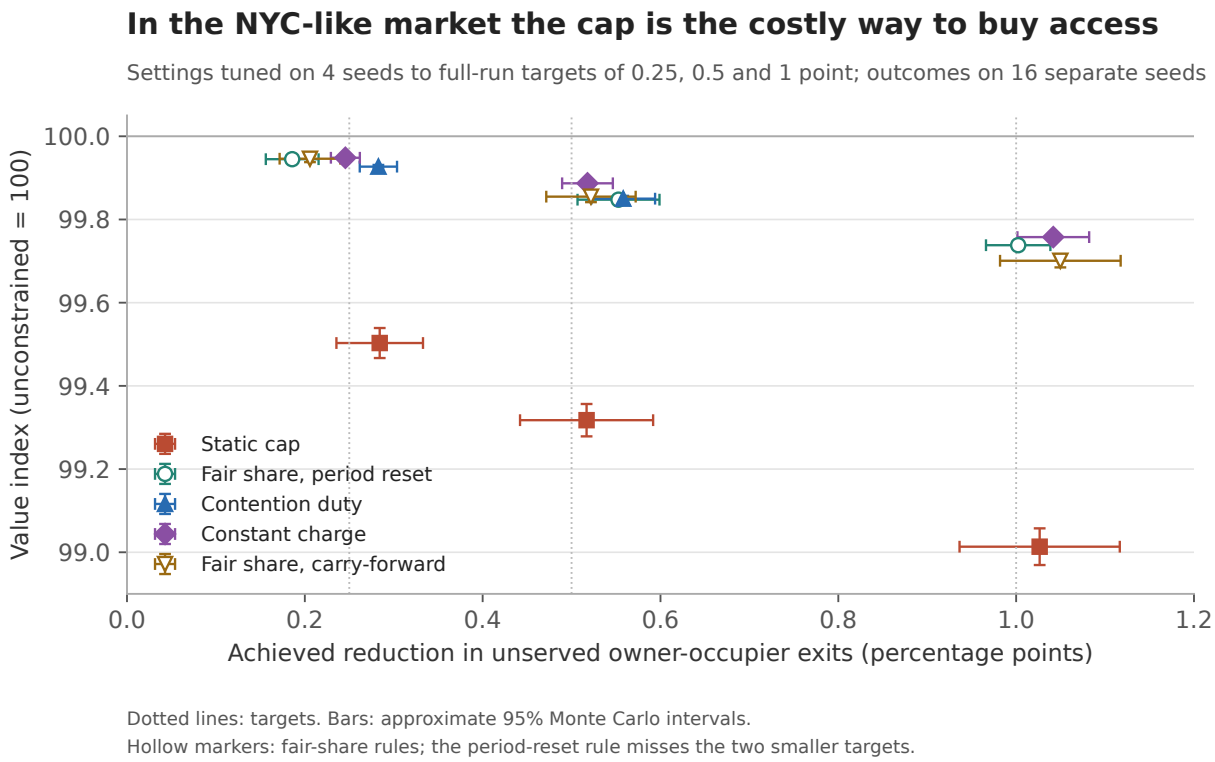


Figure 3. Main-market instruments tuned to common full-run access targets (0.25, 0.5, and 1.0 pp) and evaluated on 16 held-out replications (corresponding to Table 8). Bars show approximate 95% Monte Carlo intervals conditional on tuned settings; points are not joined into a policy frontier.

**Discrete quota accounting.** With three institutional bids and roughly 20 listings per district-period, admitting institutional purchases while prior wins are below $s \cdot L$ permits up to $\lceil s \cdot L \rceil$ wins per period. Full-run gains therefore step discontinuously from 0.21 to 0.31 pp near the 0.25 target and from 0.40 to 0.55 pp near the 0.5 target (asterisks in Table 8; Figure 3). Replacing period-reset limits with fractional carry-forward credits removes the integer gap and hits all three tuning targets (yielding held-out gains of 0.21, 0.52, and 1.05 pp). Under a single institutional bid per district-period (4.47% unconstrained share), period-reset shares degenerate into a binary all-or-nothing switch, whereas carry-forward credits still achieve 0.48 and 0.96 pp gains (value indices 99.737 and 99.477). Finally, removing the 1.5% offer discount shifts baseline full-run exits by only 0.3 pp (9.8% to 9.5%), whereas excluding institutions shifts them by 6.6 pp (Supplement S5, Table S5.3), underscoring that these paths reflect mechanical capacity assumptions rather than an empirical evaluation of cash-buyer taxation.

---

# 5. Practical Reporting Standard, Scope, and Conclusion

## 5.1 Practical five-rule reporting contract

Extending STRESS, ADEMP, and ODD to endogenous allocation pools requires five concrete reporting rules, each directly verifiable from saved simulation outputs:

1. **Report aggregate realised value $V$ directly.** Never present a policy-dependent ratio $V/B$ as a standalone efficiency outcome; pair every ratio comparison with the raw change in $V$ under matched exogenous inputs.
2. **Print every ratio's denominator beside the ratio and run the Eq. (1) check.** Report $B_0$ and $B_1$ (or regional denominators in multi-area models like LivSim) alongside the ratio, flag any pair where $\Delta V$ and $\Delta(V/B)$ carry opposite signs, and verify the log identity within floating-point tolerance ($10^{-12}$).
3. **Anchor ratios to a declared policy-invariant reference $F$.** Hold $F_0 = F_1 > 0$ fixed across paired arms on common input and exogenous-feasibility hashes (excluding policy eligibility and decision rules), and label explicitly whether $F$ is a feasible dynamic optimum or a relaxed upper bound.
4. **Fix and declare the access cohort before tuning instruments.** State the entry rule, follow-up endpoint, and unresolved count explicitly; separate completed-exit cohorts (where every entrant resolves) from fixed-duration clinical cohorts (where still-waiting patients are complete fixed-horizon observations) and from open full-run ratios.
5. **Declare discrete quota/banking rules and replication aggregation weights.** Specify integer rounding or carry-forward rules whenever discrete capacity limits bind, and distinguish mean-of-ratios summaries from ratios-of-pooled-totals before combining replications.

## 5.2 Scope and limitations

These exercises prove that denominator-driven sign and ranking reversals (Theorem 1) and finite-horizon cohort reorderings (Proposition 1) occur under transparent constraints; they do not estimate how frequently published simulations exhibit them. The LivSim audit relies on shipped synthetic inputs, a three-year horizon, and five paired replications, supporting no clinical policy inference; moreover, its cross-cohort differences show Monte Carlo sensitivity rather than an invariant ranking reversal. The canonical triage DES benchmark (`triage_benchmark.py`) is an intentionally stylised, self-contained two-class queue designed to test Theorem 1(ii) and Proposition 1 at high power (50 CRN replications), not an empirical hospital or shelter forecast. The housing diagnostic holds offers, supply, and institutional entity structure fixed (sensitivity checks for spatial reallocation, entity splitting, supply elasticity, and administrative registries are documented in Supplements S4–S5) and uses New York City summaries solely as scale proxies. A relaxed upper bound $F$ guarantees sign fidelity on matched inputs without measuring implementable welfare, and passing `audit_contract.py` checks verifies declared metadata and scalar consistency rather than model validity or exporter correctness. PolicySpace2 remains quarantined in Supplement S1 due to missing original logs and structural non-determinism.

## 5.3 Conclusion

Whenever an allocation rule alters who remains in the queue or who enters a rate denominator, evaluation measures defined over that endogenous population can invert true policy rankings. Theorem 1 and Proposition 1 establish the exact necessary-and-sufficient log-ratio and censoring-wedge conditions under which moving benchmarks and open-run access rates reverse fixed-reference and completed-cohort rankings. Across two exact

counterexamples, a five-pair LivSim audit, a 50-replication canonical priority-triage DES benchmark where both axes of the efficiency–access frontier flip simultaneously in 50 of 50 pairs (+3.11 vs. −11.29 pp on value; −3.17 vs. +3.31 pp on access), and a controlled housing diagnostic, accounting choices systematically alter sign, magnitude, and instrument order. Pairing aggregate value reporting with explicit denominators, fixed references, declared entry cohorts, and standard-library contract verification across all 256 exported records makes these accounting channels transparent before policy conclusions are drawn.

## Declarations

**Funding.** None.

**Competing interests.** The author founded HolaCasaClara, a free comparator of Argentine mortgage loans, which has no connection to New York housing records or this study. No party funded, reviewed, or approved the manuscript.

**Acknowledgments.** None.

**Data and code availability.** Manuscript source, saved model outputs, diagnostic scripts (`worked_audit.py`, `audit_contract.py`, `triage_benchmark.py`, `build_triage_figure.py`, `test_comparison_contract.py`, `deadline_matching_check.py`, `export_records.py`, `reproduce_review.py`), all 256 exported contract records, and synthetic LivSim event logs accompany this release. Saved audit evidence is archived at https://doi.org/10.5281/zenodo.24196689 and https://doi.org/10.5281/zenodo.23257315; no new housing, LivSim, or PolicySpace2 simulations were executed for this revision (only the self-contained canonical triage DES benchmark in `triage_benchmark.py` was added). Original PolicySpace2 monthly histories remain unavailable; the quarantined record is preserved in Supplement S1. Re-estimating the underlying New York City scale proxies requires the separately deposited transaction-linkage dataset (Loschi 2026d), which is not required to reproduce any table or audit check here.

**Declaration of generative AI and AI-assisted technologies.** The author takes full responsibility for all methodology, proofs, analysis, citations, and text. During preparation, generative AI tools (OpenAI ChatGPT/Codex, Google Gemini, and Anthropic Claude) were used only for coding, formatting, and copy-editing assistance.

## References

Agarwal, N., Ashlagi, I., Azevedo, E., Featherstone, C. R., & Karaduman, Ö. (2019). Market failure in kidney exchange. *American Economic Review*, 109(11), 4026–4070.

Arora, R., Dekel, O., & Tewari, A. (2012). Online bandit learning against an adaptive adversary: From regret to policy regret. *Proceedings of the 29th International Conference on Machine Learning*.

Axtell, R. L., & Farmer, J. D. (2025). Agent-based modeling in economics and finance: Past, present, and future. *Journal of Economic Literature*, 63(1), 197–287.

Bertsimas, D., Farias, V. F., & Trichakis, N. (2011). The price of fairness. *Operations Research*, 59(1), 17–31.

Bloch, F., & Cantala, D. (2017). Dynamic assignment of objects to queuing agents. *American Economic Journal: Microeconomics*, 9(1), 88–122.

Edmonds, B., Le Page, C., Bithell, M., Chattoe-Brown, E., Grimm, V., Meyer, R., Montañola-Sales, C., Ormerod, P., Root, H., & Squazzoni, F. (2019). Different modelling purposes. *Journal of Artificial Societies and Social Simulation*, 22(3), 6. https://doi.org/10.18564/jasss.3993

Edmonds, J. (1971). Matroids and the greedy algorithm. *Mathematical Programming*, 1, 127–136. https://doi.org/10.1007/BF01584082

Furtado, B. A. (2022). PolicySpace2: Modeling markets and endogenous public policies. *Journal of Artificial Societies and Social Simulation*, 25(1), 8. https://doi.org/10.18564/jasss.4742

Gentry, S. E., Massie, A. B., Cheek, S. W., Lentine, K. L., Chow, E. H., Wickliffe, C. E., Dzebashvili, N., Salvalaggio, P. R., Schnitzler, M. A., Axelrod, D. A., & Segev, D. L. (2013). Addressing geographic disparities

in liver transplantation through redistricting. *American Journal of Transplantation*, 13(8), 2052–2058. https://doi.org/10.1111/ajt.12301

Grimm, V., Railsback, S. F., Vincenot, C. E., Berger, U., Gallagher, C., DeAngelis, D. L., Edmonds, B., Ge, J., Giske, J., Groeneveld, J., Johnston, A. S. A., Milles, A., Nabe-Nielsen, J., Polhill, J. G., Radchuk, V., Rohwäder, M.-S., Stillman, R. A., Thiele, J. C., & Ayllón, D. (2020). The ODD protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism. *Journal of Artificial Societies and Social Simulation*, 23(2), 7. https://doi.org/10.18564/jasss.4259

Kesselman, A., Lotker, Z., Mansour, Y., Patt-Shamir, B., Schieber, B., & Sviridenko, M. (2004). Buffer overflow management in QoS switches. *SIAM Journal on Computing*, 33(3), 563–583. https://doi.org/10.1137/S0097539701399666

Kilambi, V., Bui, K., & Mehrotra, S. (2018). LivSim: An open-source simulation software platform for community research and development for liver allocation policies. *Transplantation*, 102(2), e47–e48. https://doi.org/10.1097/TP.0000000000002000

Leshno, J. D. (2022). Dynamic matching in overloaded waiting lists. *American Economic Review*, 112(12), 3876–3910.

Loschi, P. (2026a). *Who Gets the House? Related-Party Transfers and the Cash–Mortgage Price Gap in New York City* (Version 5.5) [working paper and replication package]. Zenodo. https://doi.org/10.5281/zenodo.24343848

Loschi, P. (2026d). *NYC residential sales linked to ACRIS deeds and mortgages, 2016–2025* (Version 1.2.0) [dataset]. Zenodo. https://doi.org/10.5281/zenodo.22977532

Monks, T., Currie, C. S. M., Onggo, B. S., Robinson, S., Kunc, M., & Taylor, S. J. E. (2019). Strengthening the reporting of empirical simulation studies: Introducing the STRESS guidelines. *Journal of Simulation*, 13(1), 55–67. https://doi.org/10.1080/17477778.2018.1442155

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102. https://doi.org/10.1002/sim.8086

Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., & Hardt, M. (2020). Performative prediction. *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119, 7599–7609.

Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation*, 7(1), 12–24. https://doi.org/10.1057/jos.2012.20

Sönmez, T., Ünver, M. U., & Yenmez, M. B. (2020). Incentivized kidney exchange. *American Economic Review*, 110(7), 2198–2224.

Windrum, P., Fagiolo, G., & Moneta, A. (2007). Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation*, 10(2), 8. https://jasss.soc.surrey.ac.uk/10/2/8.html