When the Benchmark Moves with the Policy

Measuring Access in Allocation Simulations

Pablo Loschi · Independent researcher

Working paper, version 7.4 · 16 November 2026

Abstract

Allocation policies change who is served, who waits and who leaves, so evaluation measures counted over those endogenous groups can move even when underlying behaviour is fixed. I audit a published liver-allocation simulator and trace the resulting benchmark and cohort mechanisms through exact discrete examples alongside a minimal housing-market counterexample built specifically to illustrate the denominator theorem instead of forecasting policy outcomes. In the liver simulator, broader sharing widens geographic differences among transplanted patients and new listings, whereas the change for patients already waiting varies across replications and can reverse under fixed regional denominators. A ratio identity and an exact two-period example show a moving-benchmark ratio rising from 0.80 to 1.00 while realised value falls from 16 to 11. In a deadline queue, a policy that reaches the feasible optimum receives a lower moving-benchmark score. In the minimal housing counterexample, scaled to New York City proxies, a contention-linked duty appears to add value above a 2.2 percent offer discount under a moving benchmark yet loses value at every discount up to 20 percent under a fixed reference. Following a fixed entry cohort alters measured access gains by factors of 0.7 to 2.5 and reorders the instruments. Those comparisons suggest that policy evaluations in simulation studies may mislead unless realised value, explicit benchmark denominators and cohort definitions are reported side by side.

Keywords: policy evaluation; efficiency measurement; agent-based model; housing access; execution certainty; simulation verification

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

---

1. Introduction

Simulations of allocation policies typically report who was served and how well, yet both counts depend on populations that the policy itself reshapes. A rule that serves a high-value agent early removes that agent from the pool against which later decisions are scored, while a rule altering who enters, waits or exits shifts the group over which access is counted. A ratio computed against such an endogenous pool can rise while total value falls, and one instrument can look stronger than another simply because its beneficiaries are counted over a different denominator. Online scheduling avoids the first distortion by scoring policies against an offline optimum on the same input (Kesselman et al. 2004). Policy regret similarly fixes the counterfactual history a policy is judged against (Arora, Dekel and Tewari 2012), providing the benchmark discipline that I bring to agent-based evaluation here.

LivSim (Kilambi, Bui and Mehrotra 2018), a published liver-allocation simulator, supplies an external application showing that these accounting shifts arise in practice. In its shipped synthetic inputs, broader regional sharing widens geographic differences among transplanted patients and new listings, while changes among patients already waiting vary across replications. Following identical patient cohorts under both policies and holding regional denominators fixed changes those disparities noticeably (Section 3). These comparisons indicate why a measure's population and denominator may need to be fixed before interpreting any simulated policy ranking.

I then isolate the underlying mechanism in a minimal pedagogical housing counterexample where realised value can be computed directly instead of relying on noisy outputs. Sellers rank fixed offers by valuation times an assumed certainty factor, and buyers either purchase or leave the queue, letting the simulation compare a per-entity cap, a work-conserving fair share, a constant charge and a contention-linked duty. A baseline parameterisation adopts New York City orders of magnitude while a higher-institutional setting acts as a stress

test. The model is neither estimated nor offered as a calibrated policy forecast; its sole role is to serve as a minimal counterexample to the denominator and cohort theorems when the true allocation is known.

Section 5 starts from an exact ratio identity showing that the log change in a ratio equals the log change in its numerator minus the log change in its denominator, so a ratio against a policy-dependent benchmark can rise while realised value declines. In an exact two-period example the moving-benchmark ratio rises from 0.80 to 1.00 as value drops from 16 to 11. In the minimal housing counterexample, a fixed benchmark shows the contention duty losing value at every offer discount up to 20%, whereas the moving benchmark shows it gaining from a discount of about 2.2%. In a deadline queue with an exact offline optimum the distortion runs the opposite way, because earliest-deadline-first raises value from 19 to 27 while its moving ratio drops from 1.00 to 0.90. Counting access over an entry cohort followed until purchase or exit shifts measured gains by factors of 0.7 to 2.5 and reverses instrument rankings, so that the duty and constant charge swap places depending on which population they are tuned to. Likewise, a period-reset fair-share quota fails to reach small access targets that carry-forward allowances hit smoothly because period limits admit purchases only in whole integers.

Section 2 situates the paper in the literature and Section 3 reports the LivSim audit, before Section 4 sets out the minimal housing counterexample. Sections 5 to 7 examine policy-dependent benchmarks, access populations and quota accounting. Section 8 summarises the stylized instruments, Section 9 sets out the portable audit contract, Section 10 states the paper's limits and Section 11 concludes, leaving detailed exercises to the appendices.

2. Related work

This paper sits at the intersection of housing finance, market design and public finance. Reher and Valkanov (2024) and Han and Hong (2024) supply the empirical motivation yet disagree on how much of the cash premium exceeds underlying risk, and that disagreement motivates sweeping the offer discount instead of transporting a single point estimate. Whereas Che and Gale (1998) show that auctions with budget-constrained bidders can allocate inefficiently, the distortion here runs through the seller's preference for execution certainty even when buyers' willingness to pay is unconstrained. Budish, Cramton and Shim (2015) similarly examine how market design rewards non-valuation advantages such as speed, though completion certainty differs because it saves genuine transaction costs. Piazzesi, Schneider and Stroebel (2020) document that housing search is segmented across clienteles, which the districts here echo in stylized form. Kopczuk and Munroe (2015) show that New York's mansion-tax notch alters prices and volume around its threshold, meaning that transaction levies may reshape the very transfers over which they are subsequently evaluated.

Any intervention in this setting poses a second-best question in the sense of Lipsey and Lancaster (1956), because a charge inside an already distorted market can either raise or lower welfare with no general ambiguity resolved a priori. Comparing a quantity rule (a cap or quota) against a price rule (a charge) follows Weitzman (1974), while carry-forward allowances borrow the banking logic of Weitzman (2018) and Akbarpour, Dworczak and Kominers (2024) characterize when below-market rationing improves equity over pure price clearing under distributional Pareto weights. The distributional framework states welfare weights explicitly (Atkinson 1970) and reports efficiency-fairness trade-offs directly (Bertsimas, Farias and Trichakis 2012). Although the simulation uses agent-based mechanics in the spirit of Tesfatsion (2006), Axtell and Farmer (2025) and Geanakoplos et al. (2012), recent policy models evaluate credit limits against historical settings (Yun and Moon 2020) or steady states (Gamal et al. 2024), whereas the access metrics studied here add an endogenous reference pool that shifts with the rule itself. In the classification of Edmonds et al. (2019), the short housing ABM acts solely as theoretical exposition and a minimal pedagogical counterexample to the denominator and cohort theorems as opposed to a predictive engine, with New York City proxies fixing unit scales without validating empirical behaviour.

The measurement distortion also has close parallels outside housing. Arora, Dekel and Tewari (2012) show why judging a rule against the history generated by another rule can mislead and introduce policy regret in response. Perdomo et al. (2020) analyse performative decision rules that shift their own evaluation distribution, and Bertsimas, Farias and Trichakis (2011) benchmark fairness losses against a common feasible optimum in the spirit of the fixed reference used here. Just as queueing metrics over served jobs conceal dropped requests,

housing wait-list and dynamic matching studies (Bloch and Cantala 2017; Thakral 2019; Baccara, Lee and Yariv 2020; Waldinger 2021; Leshno 2022) examine how queue discipline, delay, and pool accumulation reshape waiting-list composition and intertemporal match quality, while kidney-exchange analyses show that hospital fragmentation and priority incentives alter participation and pool composition across donor queues (Agarwal et al. 2019; Sönmez, Ünver and Yenmez 2020). Here, private sellers choose among competing offers and the queue matters because who remains waiting changes future denominators. Instead of claiming general empirical incidence, this paper isolates how sensitive instrument comparisons could be to those population choices.

The deadline queue of Section 5.3 implements a bounded-delay scheduling problem from Kesselman et al. (2004), who compare online rules against an offline optimum on identical inputs, so the dynamic program I solve serves strictly as a verification benchmark instead of proposing a new scheduler. Following Windrum, Fagiolo and Moneta's (2007) distinction between empirical validation and internal program verification, the checks throughout this paper establish the second only. This contrasts sharply with full empirical policy frameworks such as Furtado's (2022) PolicySpace2 that couple spatial labour, credit and housing markets on survey inputs; Appendix A records the housing specification under the ODD protocol (Grimm et al. 2020).

3. A published-model audit: LivSim

Before turning to the minimal housing counterexample, I examine a published allocation simulator whose source code and synthetic example inputs are available. Comparing its saved policy runs across served patients, yearly candidate flows and fixed cohorts tests how holding regional denominators common affects mortality-disparity comparisons. This audit does not depend on the housing model of Section 4, and Appendix I retains a provisional archived illustration excluded from the main evidence because the original monthly logs are missing and unmodified upstream code does not reproduce runs identically.

LivSim (Kilambi, Bui and Mehrotra 2018) is an open-source discrete-event simulation of liver allocation in the United States, written in Python and distributed with synthetic inputs generated at random from the official simulator's files. Across the 58 donor service areas, it reports geographic disparity in two ways: the spread of median allocation MELD scores at transplant among recipients, and the spread of a mortality rate divided by that year's arrivals plus waiting candidates. Persistent regional MELD differences helped motivate replacing donor-service-area allocation with acuity circles in 2020 (Nagai et al. 2022).

I ran the published code on those inputs for three simulated years across all five replications, comparing the shipped status quo (Share 35 and Share 15 with MELD-Na) against full regional sharing with only input-path edits, a one-line NumPy 2 update and a per-patient event log. Because LivSim replays identical arrivals, disease progression and deaths across policies, both regimes face the exact same patients. Two fixed cohorts can therefore be followed for one year: the 15,406 candidates waiting at the start and the 11,571 listed during year one.

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 the start plus the 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 the start (15,406)	−1% (narrower in 3 of 5)
One-year death risk	Waiting at the start	−7% (narrower in 4 of 5)

Table 1. LivSim, full regional sharing against the status quo: mean percentage change in the standard deviation across donor service areas over five paired replications, with the number of replications or replication-years that share the sign. The first three rows are LivSim's own outputs. Cohort rows use areas with at least 20 cohort members (50 for first-year listings, 49 for the waiting list). Synthetic inputs, three simulated years.

Regional sharing widens the spread of median MELD at transplant in 15 of 15 replication-years and the spread of one-year transplant chances among new listings in 5 of 5 replications, whereas among waiting patients it narrows death risk in 4 of 5 replications and transplant chances in 3 of 5. Although the published mortality rate narrows in year one across all five replications, the mean narrowing in transplant chances for initial waiters is small and reverses sign when individual replications are omitted (Appendix A.7). The contrasts therefore differ in stability as well as population and establish no robust reversal across entry cohorts. While the national denominator changes very little, that stability does not bound movements inside donor service areas, where holding each area's denominator common alters the mortality-disparity comparison (Appendix A.7). Indeed, the served pool shifts substantially: in replication one, 909 of the 5,106 new listings transplanted within a year under the status quo are not transplanted within a year under regional sharing while 922 others are, showing that the transplanted median compares largely distinct patients across rules.

Because the inputs are synthetic, the horizon spans three years and the replications are few, these figures say nothing about actual US clinical policy. The narrower lesson aligns with dynamic organ-market design (Agarwal et al. 2019; Sönmez, Ünver and Yenmez 2020), where fragmentation and priority incentives alter participation, pool composition, and queue competition. An allocation simulator reporting disparity over transplanted patients and candidate flows may reach a different verdict from fixed cohorts of identical patients, and fixed cohorts entering in adjacent periods could disagree with one another.

4. The illustrative housing market

4.1 Market and inputs

To construct a transparent environment where realised allocation value can be observed directly, the minimal housing counterexample models five districts over 120 periods with Poisson listings and civilian arrivals and lognormal buyer valuations. Owner-occupiers and small investors bid locally and buy once or leave when patience expires, while three institutional entities place fresh offers each period. Sellers accept the highest eligible valuation weighted by a certainty factor equal to 1 for institutions, $c_{OO} = 1 - w$ for owner-occupiers and $c_{SI} = \min(1, c_{OO} + 0.07)$ for small investors, holding offers fixed without equilibrium bidding or supply adjustments.

The baseline setting assigns one institutional bid per entity per district-period, yielding an unconstrained purchase share of 11.0%, and sets the owner-occupier offer discount to 1.5%. Those unit-scale proxies reflect recorded New York City transactions (Loschi 2026a, 2026d), where 19,019 of 166,227 eligible house sales (11.4%) involve a company cash buyer, alongside sensitivity inputs in Appendix F. Company entity status and unmatched mortgage indicators likely misclassify some cash buyers and omit financed investors, underscoring that these inputs merely fix orders of magnitude for a pedagogical counterexample instead of calibrating or forecasting an actual city market. Appendix H lists the remaining parameters and Appendix A.2 the timing schedule.

4.2 Instruments

The unconstrained baseline accepts the highest effective offer each period, while five stylized rules restrict or tax institutional demand solely to generate contrasting allocation paths. A per-entity cap limits each institution to 50 acquisitions per district over the run. A work-conserving fair share permits institutions up to 30% of a district's period listings while owner-occupiers wait, releasing that limit whenever none waits. The contention duty scales institutional offers by $(1 - \tau_{max} s)$ with ceiling $\tau_{max} = 0.22$, where s measures the share of resolved owner-occupier exits that leave unserved over a trailing six-period window (zero when empty), excluding buyers still searching. A constant charge multiplies institutional offers by $(1 - h)$, matched to the duty's average haircut or tuned to a target, and a combined regime applies both rules; none of these adjustments represents a statutory tax or computes fiscal revenue.

Appendix B derives the seller's certainty discount and explains why selecting fixed offers by normalised margin alone cannot determine a charge's aggregate value cost.

4.3 Measures

The realised-value index sums successful buyers' valuations and normalises that total to the unconstrained regime under paired random inputs. Oracle efficiency divides realised value V by the sum F of the L highest

generated valuations across all L listings, ignoring arrival dates, patience expiry and district geography so that V ≤ V* ≤ F for the feasible optimum V*. Contemporaneous efficiency divides V by the sum B of the highest valuation present in the pool at each listing, a moving benchmark whose composition changes endogenously with the policy.

Full-run starvation divides unserved owner-occupier departures by all owner-occupier arrivals during the run, leaving agents still searching at period 120 out of the numerator while keeping them in the denominator. Section 6 contrasts this convention against a completed cohort followed until purchase or exit; unless noted otherwise, figures average twelve paired replications.

5. Policy-dependent benchmarks

Of the ratios in Section 4.3, the contemporaneous one uses a reference that depends on who is still waiting.

5.1 An identity and an exact example

Consider one listing in each of two periods, where each buyer can purchase at most once. Buyer I is an institution present only in period 1, with valuation 6 and certainty factor 1. Buyer O is a household with valuation 10 and certainty factor 0.5, present in either period until served. Buyer L arrives in period 2, with valuation 1 and certainty factor 0.5. Listings are allocated immediately and cannot be held for later sale, and sellers rank offers by valuation multiplied by the certainty factor.

Without a charge, I's effective offer of 6 beats O's offer of 5 in period 1, and O then beats L in period 2, so the realised allocation value is 6 + 10 = 16. The highest valuation in the pool is 10 at both decisions: O remains available after losing the first listing. Adding those two local references gives 20, so the moving-benchmark ratio is 16/20 = 0.80. The denominator uses O twice, although O can purchase only once.

Table 2 compares these allocations. Now reduce institutional effective offers by 20%. I's offer becomes 4.8, so O wins the first listing, and because I expires after period 1, L is the only buyer left for the second listing, giving realised value 10 + 1 = 11. The local reference is 10 followed by 1, and the moving-benchmark ratio is 11/11 = 1.00.

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

Table 2. Exact example 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 feasible schedules can be enumerated. They include (I, O), (I, L) and (O, L), with values 16, 7 and 11, so 16 is the true dynamic optimum here. Giving O the first listing improves that listing's realised value by 4 but sacrifices the chance to serve I, and the later loss of 9 more than offsets the immediate gain. The example needs timing and one-purchase constraints, and it does not rely on strategic bids, stochastic failures or uncertain parameter estimates.

For any policy p, let V_p denote total realised value, B_p the sum of highest present valuations and $E_p = V_p/B_p$. For positive quantities, the comparison satisfies

$$\frac{E_1}{E_0} = \frac{V_1/V_0}{B_1/B_0}, \qquad \log\frac{E_1}{E_0} = \log\frac{V_1}{V_0} - \log\frac{B_1}{B_0}. \tag{1}$$

This is a ratio identity: if $B_1/B_0 < V_1/V_0 < 1$, realised value falls while E rises. A corresponding reversal is possible when both quantities increase and the benchmark grows faster than the numerator. Calling E an efficiency gain without reporting B conceals this possibility.

A fixed positive reference F gives $V_1/F - V_0/F$ the same sign as $V_1 - V_0$ for each run. That sign agreement requires no optimality property of F, although interpreting V/F as proximity to an economically relevant optimum does require a justified feasible benchmark. The small example supplies one, and the larger simulation uses a looser upper bound and labels it accordingly.

Appendix A.3 gives the executable enumeration and its checks.

5.2 The main market

The two efficiency ratios in Table 7 diverge systematically because policies that alter queue composition simultaneously reshape the moving reference itself. A rule serving high-valuation buyers early leaves a thinner residual queue, lowering subsequent maxima. In the main market, the contemporaneous benchmark averages roughly 444,000 per unit without intervention and under the duty against 463,000 under the cap, rendering cross-regime ratios non-comparable.

Table 3 illustrates this divergence across the offer-discount sweep. On the fixed benchmark the duty loses value at every offer discount from 0% to 20%, shrinking from 0.049 to 0.005 points and never turning positive across 3,943 of 4,000 bootstrap resamples of the eight replications. Under the moving benchmark, the duty appears to add value above roughly a 2.2% discount (95% interval [2.1, 2.3]). The two measures thus disagree on sign at both the upper company-cash discount of 5.3% and the headline 8.8% discount, even though both report a modest loss at the 1.5% baseline proxy.

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 3. The duty's efficiency gain over the unconstrained market in the main market as the owner-occupier offer discount varies, 8 replications per row, duty ceiling 0.22. At each row $c_{OO} = 1 - w$ and $c_{SI} = \min(1, c_{OO} + 0.07)$, and institutional certainty stays at one. The institutional share is the unconstrained share at that discount. Standard errors of the gains are below 0.008 points.

The fix is to hold the benchmark independent of the policy: the value index needs only identical arrivals across regimes, and oracle efficiency needs the full arrival stream. A simple diagnostic makes the problem visible, which is to report the benchmark's value per unit alongside each ratio. When this denominator differs across treatment arms, a change in the ratio cannot be attributed solely to total realised value.

The stress test likewise reverses the apparent sign of the duty's value effect at intermediate discounts because its moving reference falls faster than realised value (Appendix C).

5.3 A queue where the error runs the other way

The second counterexample schedules unit-service jobs with release slots, inclusive deadlines and fixed values in discrete slots, where highest-value-first (HV) picks the largest currently available job and earliest-deadline-first (EDF) selects the earliest deadline (breaking ties by value and identifier). Service occurs after arrivals and before each slot's deadline, so both online rules observe only current jobs while the offline optimum uses full arrival information subject to release, deadline and one-job-per-slot capacity limits.

In a constructed three-slot example, job A has value 10 in slots 0 through 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, so HV serves A, C and D whereas EDF retains A for slot 2 and serves B, C and A instead (Table 4).

Policy	Service order	Realised value V	Moving reference B	V/B	V/feasible optimum
Highest value first	A, C, D	19	19	1.00	0.704
Earliest deadline first	B, C, A	27	30	0.90	1.000

Table 4. Constructed service queue example. The exact offline optimum is 27. Relative to HV, EDF raises value by 42.11% and the moving reference by 57.89%, so V/B falls by 10 percentage points. Each job uses one slot and can be served once.

The moving reference rises under EDF because retaining A leaves value 10 present at all three decisions. Local greedy selection guarantees HV a ratio of one despite missing nine units of feasible value, while EDF as an online heuristic reaches the feasible optimum here although it need not do so across arbitrary queues.

Across 200 generated six-slot cases with two arrivals per slot, integer values uniform on 1 through 20 and deadline extensions uniform on 0, 1 and 2 slots truncated at slot 5 (reproduction details in Appendix A.4), EDF raises realised value while lowering V/B in 15 cases. Another 162 cases show matching nonzero signs, 23 display a tie and none exhibits falling value with a rising moving ratio, reflecting this small synthetic design where HV's moving ratio equals one.

A dynamic program computes the exact feasible optimum in every case and is verified against full enumeration over all 5,832 three-job, three-slot configurations with values in {1, 2, 3}. The portable comparison routine checks availability, capacity and value reconstruction across all 72 housing pairs and 200 service pairs while rejecting undefined or negative inputs, leaving a largest service log residual below 2.1×10^{-16}. Section 3 applies the same portable audit logic to an external published simulator.

5.4 What the moving ratio still describes

The moving ratio answers a well-defined local question by measuring how winners' valuations compare with the highest valuation present at each decision. That comparison usefully describes the selection rule yet could seriously mislead if interpreted as aggregate value or proximity to a fixed optimum.

Using a fixed upper bound avoids that sign reversal under matched exogenous inputs and yields a conservative efficiency floor: any path reaching 99.5% of F achieves at least 99.5% of the feasible optimum under the same objective, while the residual gap to F likely overstates avoidable loss. Because comparing V across matched paths establishes the sign of any value change directly, I report realised-value changes as the principal outcome and use moving ratios to illustrate benchmark bias.

6. Access populations

6.1 The main market

The default settings differ in strength as well as design, so the comparison tunes each instrument to common access targets. Achieved gains on the evaluation replications remain visible because they need not match exactly. I tuned each on four simulation replications to lower full-run starvation by 0.25, 0.5 or 1 percentage point, and the chosen settings were evaluated on sixteen separate replications (Table 5, Figure H1). The fair share appears twice, with the period-reset allowance of Section 4.2 and with allowances that carry forward (Section 7). The duty cannot reach the one-point target: at its highest tested ceiling, 0.95, it lowers full-run starvation by 0.88 points on the tuning replications, because its signal stays low when few owner-occupiers leave unserved.

The cap is the expensive instrument at every target, giving up 0.50, 0.68 and 0.99 index points. Per point of access it costs four to nine times as much realised value as the charges, which differ from the fair share by hundredths of an index point. At the two smaller targets the constant charge keeps more value than the duty, by 0.021 (paired SE 0.0017) and 0.037 points (paired SE 0.0044), while achieving 0.04 points less full-run access (paired SE 0.013 and 0.019).

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 5. Instruments tuned to common full-run access targets in the main market: settings tuned on four replications, results on sixteen held-out replications. Gains are reductions in the full-run and completed-cohort unserved rates. The cap setting is acquisitions per entity, the fair-share settings are institutional period shares, and the charges are rates. An asterisk marks a period-reset fair share whose tuning could not come within the larger of 0.02 points and 10% of the target (Section 7).

Because the full-run starvation rate counts agents still searching when the run ends in its denominator without counting their eventual exit status, policies holding buyers in the queue longer can distort unserved comparisons. Following every owner-occupier who enters in periods 36 to 83 until they buy or leave by period 119 resolves the entire cohort, where 15.1% leave unserved without intervention compared with a 10.1% full-run rate.

Across the full-run settings of Table 5, cohort gains range from roughly seven-tenths of the full-run gain for the carry-forward fair share at the smallest target up to two and a half times it for the cap across all targets. The two measures therefore rank the instruments differently. At the half-point target, the cap and constant charge share the smallest full-run gain of 0.52 points, yet the cap delivers a 1.29-point cohort gain against 0.75 points for the constant charge. Even when matched to the duty on full-run access, the constant charge yields 0.29 and 0.48 points less cohort access (paired SE 0.028 and 0.039).

Retuning directly to the cohort target reverses the charge comparison again (Table 6). When tuned to cohort gains of 0.5 and 1 point on the same four seeds and evaluated across sixteen disjoint replications, the duty retains 0.018 and 0.027 more value index points than the constant charge (paired SE 0.0020 and 0.0052). Although both charges achieve cohort gains within 0.03 points of each other (paired SE 0.022 and 0.044), the constant charge delivers 0.13 and 0.23 points more full-run access (paired SE 0.011 and 0.021). Each charge keeps slightly more value on the population for which it was tuned, just as in the stress test (Appendix D), and neither ranking generalises beyond its stated population.

Target (pp)	Instrument	Setting	Cohort gain (pp; SE)	Full-run gain (pp; SE)	Value index (SE)
0.5	Static cap	75	0.44 (0.055)	0.18 (0.022)	99.60 (0.019)
0.5	Fair share, period reset	0.1016	0.56 (0.035)	0.40 (0.018)	99.89 (0.005)
0.5	Contention duty	0.2542	0.50 (0.020)	0.23 (0.006)	99.94 (0.002)
0.5	Constant charge	0.0297	0.52 (0.029)	0.36 (0.012)	99.92 (0.002)
0.5	Fair share, carry-forward	0.1040	0.51 (0.053)	0.44 (0.034)	99.87 (0.010)
1	Static cap	70	0.88 (0.077)	0.35 (0.032)	99.42 (0.020)
1	Fair share, period reset	0.0938	0.95 (0.049)	0.67 (0.025)	99.81 (0.006)
1	Contention duty	0.5307	1.07 (0.040)	0.50 (0.017)	99.86 (0.004)
1	Constant charge	0.0561	1.07 (0.050)	0.72 (0.017)	99.84 (0.004)
1	Fair share, carry-forward	0.0974	1.05 (0.060)	0.80 (0.038)	99.76 (0.011)

Table 6. Instruments tuned to common completed-cohort access targets in the main market: settings tuned on four replications, results on sixteen replications disjoint from those of Table 5.

In the stress test, changing from full-run access to a completed cohort also changes the gains and reverses the comparison between the two charges after retuning (Appendix D).

7. Quota accounting

With three institutional bids per district-period and about 20 listings, a period share s permits institutions $\lceil s \cdot L \rceil$ purchases in a period with L listings while an owner-occupier waits. Full-run gains therefore step discontinuously from 0.21 to 0.31 points near the 0.25 target and from 0.40 to 0.55 near the 0.5 target, leaving no period share able to hit either target (asterisks in Table 5). By contrast, carry-forward allowances spending one credit per constrained institutional purchase accumulate fractional credits and hit all three targets in tuning, reaching held-out gains of 0.21, 0.52 and 1.05 points. That contrast reflects an integer accounting effect instead of an inherent flaw of period quotas, and I do not propose unlimited banking because unrestricted carry-forward may shift purchase timing while accumulation, expiry and borrowing caps remain untested.

The same integer lumpiness appears in a third specification with one institutional entity placing a single bid per district-period (4.47% unconstrained share). Any positive period share then admits that bid and zero shuts it out whenever an owner-occupier waits, so the rule either overshoots a 0.5- or 1-point target or misses it altogether. Carry-forward credits instead deliver gains of 0.48 and 0.96 points at value indices of 99.737 and 99.477, and reach 5.22 and 10.18 points (value indices 99.997 and 99.600) in the stress test at shares of 0.5103 and 0.4351.

8. What the instruments do in the model

Because the housing model serves strictly as a minimal pedagogical counterexample to the denominator and cohort theorems instead of a calibrated market forecast, this section summarises the stylized instruments' baseline behaviour so the measurement contrasts above remain clear. Without intervention, full-run starvation is 9.8%, against 41.9% in the stress test (Table 7).

Regime	OO share	Inst. share	Unserved, full run	Unserved, cohort	Value index	Oracle eff.	Contemp. eff.
Unconstrained	72.3%	11.0%	9.83%	14.87%	100.00	99.54%	99.93%
Static per-entity cap	76.1%	6.3%	8.06%	10.47%	98.64	98.18%	94.53%
Work-conserving fair share	72.3%	11.0%	9.83%	14.87%	100.00	99.54%	99.93%
Contention-triggered duty	72.7%	10.5%	9.63%	14.42%	99.95	99.49%	99.91%
Constant charge	72.6%	10.6%	9.60%	14.50%	99.95	99.49%	99.95%

Table 7. Allocation regimes in the main market at default settings, means of 12 replications. Shares are of units bought. Full-run unserved is the starvation rate, and the cohort follows owner-occupiers entering in periods 36 to 83 until they buy or leave by period 119. The constant charge equals the duty ceiling times the duty's mean signal (h = 0.020). The fair share's default limit of 30% of listings does not bind in any replication, because the three entities place at most three bids per district-period. Its row equals the unconstrained market, and the combined regime equals the duty.

At their default settings, the duty and constant charge each lower full-run starvation by roughly 0.2 points at a 0.05% realised-value cost, whereas the cap lowers it by 1.77 points at a 1.36% value loss; between the two charges, the constant charge edges ahead over the full run and the duty in the cohort (differences below 0.1 point). Because eliminating the offer discount shifts starvation to 9.5% (from 9.8%) while excluding institutions drops it by 6.6 points, quantity share rather than the certainty wedge drives most of the shortfall inside this counterexample.

In the stress test the duty's aggregate value gain stays small while quantity rules buy larger access shifts at steeper value losses, and varying New York proxies changes effect sizes while still separating fixed and moving benchmarks at the 5.3% and 8.8% gaps (Appendices E and F).

9. Using the audit

Every policy comparison should retain V and its reference denominator for each replication so the portable comparison routine can compute percentage changes, ratio gains and equation (1) across V, B and a fixed F. Although the input contract catches undefined zeros explicitly, model-specific checks remain necessary to establish that compared policies share the same exogenous arrival problem.

Paired random inputs remain comparable only when the arrival schedule stays orthogonal to policy choices. Full search and zero transaction failures preserve that schedule in the audited housing counterexample, whereas reduced search or failed deals break comparability whenever choices disturb later draws or relistings. In the deadline queue, pre-generating the full job list before any policy executes ensures that online rules and the offline dynamic program share identical inputs and constraints.

Each reference benchmark also requires an explicit economic interpretation, because summed contemporaneous maxima record opportunities on a policy's realised path while a global top-values sum F supplies an unconstrained upper bound and a dynamic optimum answers a feasibility question. Keeping these denominators separate prevents a path-dependent selection ratio from being misread as an aggregate welfare gain or exact efficiency loss.

Aggregation conventions introduce another discrepancy, since mean ratios, ratios of pooled totals and averages of percentage changes weight simulation replications differently. Equation (1) holds within each matched run

pair, whereas pooled comparisons require matching pooled numerators and denominators, which is why the housing audit prints both aggregations and checks whether they agree in sign.

Finally, any access metric requires its own population contract because waiting times over served buyers omit unserved departures. Tracking a completed entry cohort captures both outcomes, yet cohorts entering at different times still differ and settings tuned on one population do not automatically deliver equal access on another.

10. Limits

Because bids and valuations remain fixed and institutional entities never adjust entry, financing or signal manipulation, missing strategic behaviour represents the model's central simplification. Appendix E sketches a constrained district-reallocation check where even the sign of the duty's modest value effect changes alongside supply and arrival sensitivities. Those checks span limited parameters and take for granted real-world monitoring infrastructure such as tracking unserved exits, linking beneficial owners and classifying buyers (Appendix G).

The New York City proxies set numeric scales without validating the counterexample as an empirical representation of housing markets, and realised willingness to pay omits renters, public revenue uses, administration costs and heterogeneous marginal utility of income. Core runs hold housing supply fixed, while the supply extension omits rental markets and construction lags. Costless entity splitting bounds avoidance without enforcement, and finite-run or single-cohort starvation rates likely differ from lifetime housing access across cohorts.

Likewise, the deadline queue uses unit-length jobs over a short horizon without strategic arrivals. Both its reversal count and the housing comparisons remain conditional on simplified constructions that never estimate how often empirical simulations use policy-dependent denominators or predict actual city outcomes.

11. Conclusion

Whenever an allocation policy changes who stays in the pool or who enters the denominator, evaluation measures built over that endogenous population can move inversely with actual outcomes. In the LivSim audit, regional sharing widens disparities among transplanted patients and new listings while showing less consistent, denominator-sensitive shifts among initial waiters. Both an exact ratio identity and constructed two-period or deadline-queue examples confirm that a moving benchmark can rise as realised value drops or fall when a scheduler reaches the feasible optimum. Similarly, the minimal housing counterexample flips a duty's sign above a 2.2% discount, shifts access gains by factors of 0.7 to 2.5 between full-run and cohort definitions, and separates period-reset from carry-forward quotas through integer rounding.

The practical remedy is transparent accounting alongside realised-value reporting: simulation studies should publish aggregate value V directly, print every ratio's denominator beside the ratio, benchmark against a stated fixed reference, fix the access cohort before tuning, and state quota banking rules explicitly. The portable diagnostic routine packaged with the paper reads saved totals from any simulator and flags denominator-driven sign reversals automatically, extending simulation transparency standards (Squazzoni et al. 2020) to endogenous evaluation pools.

Within the short housing counterexample itself, removing the cash execution wedge shifts full-run exits by only three-tenths of a percentage point while excluding institutions moves them by 6.6 points, showing how heavily the setup depends on mechanical capacity assumptions. Because the housing exercise operates strictly as a minimal pedagogical counterexample to the denominator and cohort theorems instead of a calibrated forecast or welfare test, neither its certainty sensitivity nor the value cost of its per-entity cap should be read as a policy recommendation or instrument ranking for New York City.

Declarations

Declaration of generative AI use. I used Anthropic's Claude and OpenAI's Codex for drafting, coding, literature searches, the service-queue implementation, the LivSim audit, reproduction checks and figure revision.

These systems carried out the computational checks reported with the replication packages, which retain a version-by-version record of their use. I am responsible for the content.

Competing interests. I 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.

Funding. None.

Data availability. Manuscript sources and appendix diagnostics are provided in `methods_v7_4_manuscript_source.zip`, and saved audit evidence is deposited in `Methods_v7_3_audit_evidence.zip` (archived at https://doi.org/10.5281/zenodo.24196689). The housing simulation is retained as `WEDGE_v6_5.zip`. The benchmark audit, deadline queue and portable comparison routine are also available at https://doi.org/10.5281/zenodo.23257315. New York City inputs come from Loschi (2026a) and the linked dataset (Loschi 2026d). The audit-evidence archive contains all five LivSim pairs' saved annual outputs and synthetic patient event logs, the five-pair summary, analysis code and provenance for the later PolicySpace2 replay. The original PolicySpace2 monthly logs remain unavailable in the inspected local and public archives, and Appendix I.2 records the controlled non-determinism checks explaining why exact historical replay does not hold; Table I1 is reconstructed from archived summaries and does not support the main conclusions.

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.

Akbarpour, M., Dworczak, P., & Kominers, S. D. (2024). Redistributive allocation mechanisms. *Journal of Political Economy*, 132(6), 1831–1875.

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*.

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*, 2(3), 244–263.

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.

Baccara, M., Lee, S., & Yariv, L. (2020). Optimal dynamic matching. *Theoretical Economics*, 15(3), 1221–1278.

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

Bertsimas, D., Farias, V. F., & Trichakis, N. (2012). On the efficiency–fairness trade-off. *Management Science*, 58(12), 2234–2250.

Björkegren, D., Blumenstock, J. E., & Knight, S. (2026). Manipulation-robust prediction. *American Economic Review*, 116(9), 3263–3293.

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

Bloomberg News (2026). NY plans new tax on NYC homes over $1 million bought with cash (14 May); NYC cash home purchases tax is likely to be dropped from New York state budget (21 May); New York lawmakers approve second-home tax in state budget (27 May). Bloomberg.

Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *Quarterly Journal of Economics*, 130(4), 1547–1621.

Che, Y.-K., & Gale, I. (1998). Standard auctions with financially constrained bidders. *Review of Economic Studies*, 65(1), 1–21.

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

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.

Gamal, Y., Elsenbroich, C., Gilbert, N., Heppenstall, A., and Zia, K. (2024). A behavioural agent-based model for housing markets: Impact of financial shocks in the UK. *Journal of Artificial Societies and Social Simulation*, 27(4), 5. https://doi.org/10.18564/jasss.5518.

Geanakoplos, J., Axtell, R., Farmer, J. D., Howitt, P., Conlee, B., Goldstein, J., Hendrey, M., Palmer, N. M., & Yang, C.-Y. (2012). Getting at systemic risk via an agent-based model of the housing market. *American Economic Review*, 102(3), 53–58.

Genesove, D., & Mayer, C. (2001). Loss aversion and seller behavior: Evidence from the housing market. *Quarterly Journal of Economics*, 116(4), 1233–1260.

Grimm, V., et al. (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.

Hafalir, I., & Krishna, V. (2008). Asymmetric auctions with resale. *American Economic Review*, 98(1), 87–112.

Han, L., & Hong, S.-H. (2024). Cash is king? Understanding financing risk in housing markets. *Review of Finance*, 28(6), 2083–2118. https://doi.org/10.1093/rof/rfae025

Head, A., Lloyd-Ellis, H., & Sun, H. (2014). Search, liquidity, and the dynamics of house prices and construction. *American Economic Review*, 104(4), 1172–1210.

Kesselman, A., Lotker, Z., Mansour, Y., Patt-Shamir, B., Schieber, B., and 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

Kopczuk, W., & Munroe, D. (2015). Mansion tax: The effect of transfer taxes on the residential real estate market. *American Economic Journal: Economic Policy*, 7(2), 214–257.

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

Lipsey, R. G., & Lancaster, K. (1956). The general theory of second best. *Review of Economic Studies*, 24(1), 11–32.

Loschi, P. (2026a). *Who Gets the House? Related-Party Transfers and the Cash–Mortgage Price Gap in New York City* (Version 3.12) [working paper and replication package]. Zenodo. https://doi.org/10.5281/zenodo.22421850

Loschi, P. (2026b). *The Execution-Certainty Wedge: Allocation, Distribution, and the Design of Housing-Market Instruments* (Version 5.1, extended version) [working paper and replication package]. Zenodo. https://doi.org/10.5281/zenodo.22977667

Loschi, P. (2026c). *The Shape of the Tax: Round Prices and Bunching at New York City's 2019 Transfer-Tax Notches* (Version 3.10) [working paper]. Zenodo. https://doi.org/10.5281/zenodo.22925301

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

Maskin, E., & Riley, J. (2000). Asymmetric auctions. *Review of Economic Studies*, 67(3), 413–438.

Nagai, S., Ivanics, T., Kitajima, T., Shimada, S., Shamaa, T., Collins, K., Rizzari, M., Yoshida, A., Moonka, D., & Abouljoud, M. (2022). Disparities in the effects of acuity circle–based liver allocation on waitlist and transplant practice between centers. *Transplantation Direct*, 8(10), e1356. https://doi.org/10.1097/TXD.0000000000001356

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.

Piazzesi, M., Schneider, M., & Stroebel, J. (2020). Segmented housing search. *American Economic Review*, 110(3), 720–759.

Reher, M., & Valkanov, R. (2024). The mortgage-cash premium puzzle. *The Journal of Finance*, 79(5), 3149–3201.

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

Squazzoni, F., Polhill, J. G., Edmonds, B., Ahrweiler, P., Antosz, P., Scholz, G., Chappin, É., Borit, M., Verhagen, H., Giardini, F., and Gilbert, N. (2020). Computational models that matter during a global pandemic outbreak: A call to action. *Journal of Artificial Societies and Social Simulation*, 23(2), 10. https://doi.org/10.18564/jasss.4298.

Tesfatsion, L. (2006). Agent-based computational economics: A constructive approach to economic theory. In L. Tesfatsion & K. L. Judd (Eds.), *Handbook of Computational Economics*, Vol. 2, 831–880. Elsevier.

Thakral, N. (2019). Matching with stochastic arrival. *AEA Papers and Proceedings*, 109, 209–212.

Waldinger, D. (2021). Targeting in-kind transfers through market design: A revealed preference analysis of public housing allocation. *American Economic Review*, 111(8), 2660–2696.

Weitzman, M. L. (1974). Prices vs. quantities. *Review of Economic Studies*, 41(4), 477–491.

Weitzman, M. L. (2018). *Prices or Quantities Dominate Banking and Borrowing*. NBER Working Paper 24218.

Windrum, P., Fagiolo, G., and 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.

Yun, T.-S., and Moon, I.-C. (2020). Housing market agent-based simulation with loan-to-value and debt-to-income. *Journal of Artificial Societies and Social Simulation*, 23(4), 5. https://doi.org/10.18564/jasss.4410.

Appendix A. Reproducibility

A.1 Housing simulation

The package retains the simulation code of the extended working paper (Loschi 2026b) and adds the main-market and proxy scripts. `realistic_market.py` runs the main-market benchmark sweep, the common full-run and cohort targets and the carry-forward quota (Tables 3, 5 and 6 and the first row of Table C1) under the plan in `evidence/realistic_market_plan.json`, while `realistic_tables.py` builds those tables, `build_figures_v64.py` draws Figures C1 and H1, and `check_realistic_text.py` checks every main-market number in the text against the outputs. The main-market exercises use the same tuning and evaluation seeds as the stress test, and the benchmark sweep uses the first eight baseline seeds. `clean_reproduce.py` regenerates the baseline, the offer-discount and rate sweeps, the supply and entity experiments, the capacity and valuation sweeps, the break-even schedule of Table C1 and the class-weight schedule. The main scripts are `replication/sim_v6.py` (baseline and failed-transaction variants), `replication/sim_v6_supply.py` (listing response), `replication/sim_v6_search_fixed.py` (reduced search), `replication/displacement.py` (Appendix B.2), `replication/welfarist.py` (Appendix E), `replication/breakeven_schedule.py` (Table C1), `equal_access.py` (common targets), `constant_charge.py` (Table D1), `check_completed_cohorts.py` (Table D2), `cohort_targets.py` (cohort retuning) and `carry_share.py` (Section 7). `threshold_intervals.py` reruns seed by seed the grid nodes around each break-even crossing, checks that their means reproduce the published schedule exactly, and bootstraps the matched seeds for the intervals in Appendix C and Table C1.

The baseline uses seed 20260824 and twelve replications, advancing the seed by 100 between runs. The common-target exercises tune on seeds 22260824 + 100i for i = 0, …, 3 and evaluate on 23260824 + 100i for i = 0, …, 15, whereas the cohort retuning is evaluated on sixteen further seeds disjoint from both. Within a seed every policy sees identical arrivals, valuations and listings so that the oracle denominators align, while the reduced-search runs draw search from a separate generator to preserve that alignment.

On 21 September 2026 the package was extracted into a fresh directory with reads from the original workspace blocked, and every result was regenerated, including the 1,020 runs behind the constant-charge, cohort and carry-forward experiments, matching the shipped outputs to floating-point tolerance. The New York City estimates in Table F1 come from Loschi (2026a). Tables 7 and F2 are reproduced by running `replication/nyc_anchored.py` followed by `replication/nyc_cohort_check.py`, which leaves the original model unchanged and verifies every original summary exactly, while `replication/verify_nyc_inputs.py` recomputes the share

proxies from the separately deposited dataset using the instructions and input hashes in the README and `nyc_inputs.json`.

A.2 Model description following the ODD protocol

The model is intended solely to expose how repeated allocation changes the queue used to evaluate a policy and does not attempt to approximate the distribution of household search histories in an observed city. Following the ODD reporting structure, its entities are buyer-instances, institutional entity identifiers and district-period listings. Civilian instances store buyer class, valuation and entry time, whereas institutional instances also store an entity identifier, and district state records waiting buyers, acquisitions by entity and a rolling exit history. The evaluation routines store realised value, the current-pool reference and all generated valuations.

There are no spatial links between districts in the core program, and all current civilian buyers observe all local listings under full search. Valuations and patience are fixed at entry without varying by elapsed waiting time, so the model's central interaction is competition among current offers alongside the removal or expiry of bidders while institutional entities generate capacity repeatedly without an endogenous balance sheet or intertemporal objective.

Scheduling and stochastic inputs

Each period sequences arrivals, civilian expiry, institutional instances, listings, lagged signal calculation, sequential allocations and the exit-history update. Every buyer instance can win at most once, institutional instances disappear at the period boundary, and civilian expiry removes entrants whose elapsed periods since entry reach the specified patience before new listings are processed. These scheduling conventions define part of the feasible dynamic problem instead of minor implementation details.

Poisson arrival and listing draws and lognormal valuation draws use the main random generator, so at full search the number of draws stays independent of earlier winners. A separate generator handles transaction failures, though the full-search, zero-failure audit does not activate it. The inherited reduced-search diagnostic likewise separates search draws from exogenous inputs, leaving those extensions outside the reported audit alongside parameter and source hashes for the updated outputs.

The quota checks institutional wins against quota share times current listings before choosing each winner. Because wins take integer values, a period-reset rule can permit a purchase before the quota threshold is crossed instead of acting as a smooth fractional allocation mechanism, while the cap accumulates acquisitions over all periods. These discrete rules appear to explain why entry cohort and accounting period matter to policy comparisons.

A.3 Benchmark audit commands

Run `python replication/benchmark_toy.py` for Table 2, `python replication/benchmark_audit.py` for Table C2 and the seed-pair file, `python replication/cohort_audit.py` for the stress-test cohort check of Table D2, and `python build_benchmark_figures.py` to draw Figure C2 from the saved outputs in the benchmark-audit package.

The housing sweep comprises 144 runs and the cohort audit comprises 80, and every inherited field selected for those checks reproduced exactly in Python 3.14.7 with NumPy 2.5.3. The cohort checks also verified valuation hashes, fresh listings and fixed-reference equality, providing a targeted fresh reproduction without re-running every historical sensitivity or testing empirical accuracy.

A.4 Service queue and portable comparison

The fixed service-queue check retains all 200 seeds from 20261006 through 20261205.

Run `python replication/deadline_queue.py` to regenerate Table 4, all 200 service cases and the independent enumeration check, and run `python replication/portable_audit.py` to apply `compare_benchmarks.py` to the saved housing and service totals. Both components use Python's standard library without external data or network access, and the service example imports no housing code.

The offline dynamic program maximises value over (slot, served-job set) states by permitting either an idle slot or any unused job whose inclusive release-deadline interval covers the current slot. Because each transition advances one slot, the recurrence enumerates every feasible next action, making its exponential served-set state space suited to a small verification design instead of large-scale scheduling. The independent verifier

instead enumerates complete slot assignments, including idling, before rejecting repeated jobs and interval violations.

I report the constructed four-job example separately from the random suite. The random generation design and full consecutive seed range were fixed before inspecting outcomes, with all cases recorded in `audit/deadline_runs.json` and `deadline_pairs.csv`. Mean value-to-optimum ratios in that suite are 0.9869 for HV and 0.8940 for EDF, so EDF's success in Table 4 establishes no general superiority over HV and merely shifts queue composition to test the diagnostic.

A.5 PolicySpace2 audit provenance

The provisional PolicySpace2 illustration and its archived table appear in Appendix I. That appendix documents the missing historical monthly logs, the local and archive search, and the controlled non-determinism probes explaining why unmodified upstream commit `ced77dd` does not replay identically; PolicySpace2 plays no role in the abstract or main conclusions.

A.6 LivSim audit provenance

The original three paired replications use the shipped synthetic inputs, three simulated years and NumPy seed 7777 separately for each policy and replication. Only paths, the one-element NumPy acceptance comparison and audit logging are patched, while the earlier extension completed the remaining two pairs. The five-pair summary and saved annual area counts used for the diagnostics are retained under evidence/ in the manuscript source archive, and the audit-evidence archive provides all five pairs under livsim/runs/ alongside instrumented source, input hashes and run-time records under livsim/. Only completed saved outputs inform these diagnostics, and I ran no additional simulation for them.

A.7 LivSim replication sensitivity

I recompute the mean percentage change for each disparity measure from the five paired runs underlying Table 1. The standard error is the sample standard deviation of the five within-pair percentage changes divided by the square root of five, while the leave-one-out range drops each pair in turn to test sensitivity to individual synthetic input replications without running another simulation.

Disparity measure	Mean change (%)	Paired SE	Widening runs	Leave-one-out mean range (%)
MELD at transplant, year 1	+4.8	2.3	5/5	[+2.5, +5.6]
MELD at transplant, year 2	+16.9	5.1	5/5	[+12.5, +19.8]
MELD at transplant, year 3	+20.3	2.6	5/5	[+18.4, +22.0]
Mortality rate, year 1	−7.5	2.3	0/5	[−9.0, −5.6]
Mortality rate, year 2	+7.9	9.5	3/5	[+0.9, +13.8]
Mortality rate, year 3	+1.1	7.8	3/5	[−5.5, +5.5]
New listings: transplant chance	+10.4	1.6	5/5	[+9.1, +11.5]
New listings: death risk	+4.1	3.3	3/5	[+2.4, +6.4]
Already waiting: transplant chance	−0.8	2.6	2/5	[−2.7, +1.2]
Already waiting: death risk	−6.6	3.7	1/5	[−8.6, −3.1]

Table A1. Dispersion and sensitivity of the LivSim comparisons in Table 1. A positive change widens the standard deviation across donor service areas, whereas a negative change narrows it. The standard-error and range columns are expressed in the same percentage-change units as the mean. These statistics describe the five shipped synthetic input replications and do not quantify uncertainty about actual patients or clinical policy effects.

The mean change in transplant-chance disparity for patients already waiting crosses zero under leave-one-out analysis, whereas the widening among new listings retains its sign in every individual replication and every leave-one-out mean. Mean mortality disparity in the third year also changes sign under omission, and death-risk cohort comparisons keep their mean signs under omission while varying in sign across individual runs, suggesting that a cohort ranking reversal cannot be treated as invariant. I therefore retain the differing population descriptions and qualify any claim of a ranking reversal. The script livsim_uncertainty.py writes the table and evidence/LIVSIM_UNCERTAINTY.json from the saved five-pair comparison, leaving all existing Table 1 values unchanged.

National denominator changes can conceal offsetting shifts across donor service areas. To check this accounting channel, I reconstruct each area's death count and denominator from the saved annual outputs, recover every published mortality-disparity value, and compare both policies' death counts first under the status-quo denominator for each area and then under the sharing denominator. Empty areas are omitted under the simulator's convention, and the nonempty area set stays identical across policies in each pair and year.

Year	Own denominators, %	Common status-quo denominators, %	Common sharing denominators, %	DSA denominator changes: 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 A2. Mortality-disparity accounting sensitivity over the same five pairs. The first three estimate columns report mean percentage changes, with paired standard errors in parentheses, in the standard deviation of area-level normalized death counts under sharing against the status quo. The common-denominator columns apply the named policy's area-level denominator to both policies' death counts. Area ranges span all pairs in the indicated year, while the national denominator changes by less than 0.2% in every pair-year. These normalizations do not estimate mortality risks for a common patient cohort.

Using common status-quo denominators changes the mean disparity increase in the third year from 1.1% to 6.6%, whereas common sharing denominators yield 1.3%. The choice of reference therefore appears consequential, and the near-constant national total cannot establish that area-level denominator shifts are negligible. The first-year comparison stays unchanged because both policies begin with identical area denominators. The script livsim_fixed_denominators.py saves the 30 reconstructed published values, each accounting comparison and input hashes in evidence/LIVSIM_FIXED_DENOMINATORS.json, serving strictly as an accounting check on existing outputs instead of an additional simulation or clinical finding.

Appendix B. The seller's problem and which bids a charge displaces

B.1 How much discount risk justifies

Suppose a seller holds a unit worth P, while a cash offer C closes at horizon τ_C and a financed offer F closes with probability $1 - q$ at $\tau_F > \tau_C$. If the financed transaction fails, the seller incurs a relisting loss δ and eventually sells at P, leaving a risk-neutral seller with holding cost k per period indifferent when

$$(1 - q)(F - k\tau_F) + q(P - \delta - k\tau_F) = C - k\tau_C. \quad \text{(B1)}$$

Setting $P = C$, $d = \delta/C$ and $\kappa = k(\tau_F - \tau_C)/C$, the justified discount on the financed price, $\pi^* = (F - C)/F$, is

$$\pi^* = (qd + \kappa) / (1 - q + qd + \kappa). \quad \text{(B2)}$$

This threshold depends on the failure probability, the cost of a failed sale and the gap in closing times, and it remains positive even when no sale fails (equalling $\kappa/(1 + \kappa)$ at $q = 0$). Only a discount beyond π^* provides a candidate distortion, and neither a price gap nor a mortgage denial rate identifies π^* without local evidence on accepted offers that fail. Both cost channels can plausibly exceed a mechanical relisting cost because sellers facing a nominal loss ask more and sell more slowly. Among Boston condominium owners in the 1990s, Genesove and Mayer (2001) find asking prices higher by 25 to 35 percent of the expected loss alongside a lower sale hazard, raising d when a financed buyer fails. Time to sell also moves with the cycle (Head, Lloyd-Ellis and Sun 2014), so κ becomes larger in slow markets.

B.2 Which bids a proportional charge displaces

Consider one listing with fixed offers and eligibility where a targeted buyer j has valuation v_j and certainty factor c_j, and an untargeted buyer i has v_i and c_i. Moving the unit from j to i changes realised valuation by $-G$, with $G = v_j - v_i$, although this local calculation leaves aside real completion costs.

Proposition B1 (fixed-offer selection). *Let $a = v_j c_j$ be the highest effective offer among targeted buyers and $b = v_i c_i$ the highest among untargeted buyers. Suppose $a \geq b > 0$, ties go to the targeted group, and offers and eligibility are fixed. Multiplying every targeted offer by $(1 - \theta)$, with $0 < \theta < 1$, changes the winning group exactly when*

$$r = 1 - b/a < \theta. \quad \text{(B3)}$$

If $c_j > c_i$ and $\theta \leq 1 - c_i/c_j$, every such displacement has $v_i > v_j$.

Proof. Scaling all targeted offers by the same factor preserves their ranking, so the best targeted offer after the charge is $(1 - \theta)a$. The untargeted group wins if and only if $b > (1 - \theta)a$, which reduces to (B3), and a displacement implies $v_i/v_j > (1 - \theta)c_j/c_i$, which exceeds or equals one under the stated condition. ∎

Because the proposition holds offers fixed, a strategic extension would need to specify bidders' information, payoffs, the charge's treatment and tie-breaking. Public knowledge of valuations and score weights bounds the offers a buyer can support, yet it does not supply an equilibrium on its own. With private valuations and asymmetric bidders, first-price shading could alter which buyer wins (Maskin and Riley 2000; Hafalir and Krishna 2008). Condition (B3) therefore characterizes displacement only for given offers, leaving equilibrium offer responses to a separate bidding model. A duty linked to a published signal also adds a margin over where and when to bid, explored in a limited spatial exercise in Section 10.

The charge selects listings by their normalised effective-offer margin, which by itself reveals little about its average valuation loss relative to a cap's. Consider two equally likely listings with targeted and untargeted valuations (10,000, 9,700) and (100, 80) paired with certainty factors 1 and 0.88. A 20% charge alters the winner only at the first listing at a valuation loss of 300. By contrast, excluding the targeted buyer with probability one-half at each listing yields the same expected number of displacements at a conditional average loss of 160, showing why aggregate comparisons have to be simulated directly.

In the stress-test simulation these local displacements diverge sharply across instruments. Removing the active instrument from the candidate pool at each listing and comparing winners gives a mean G of +143,789 for the cap, +22,188 for the fair share and −12,905 for the duty, with 9.0%, 32.8% and 80.4% of displacements replacing the winner with a higher-value buyer. The duty tends to steer units toward buyers who value them more, whereas the cap mostly does the opposite, yet these local comparisons cannot simply be summed into a policy effect because a displaced or served buyer may transact in a later period.

Appendix C. Benchmark stress tests

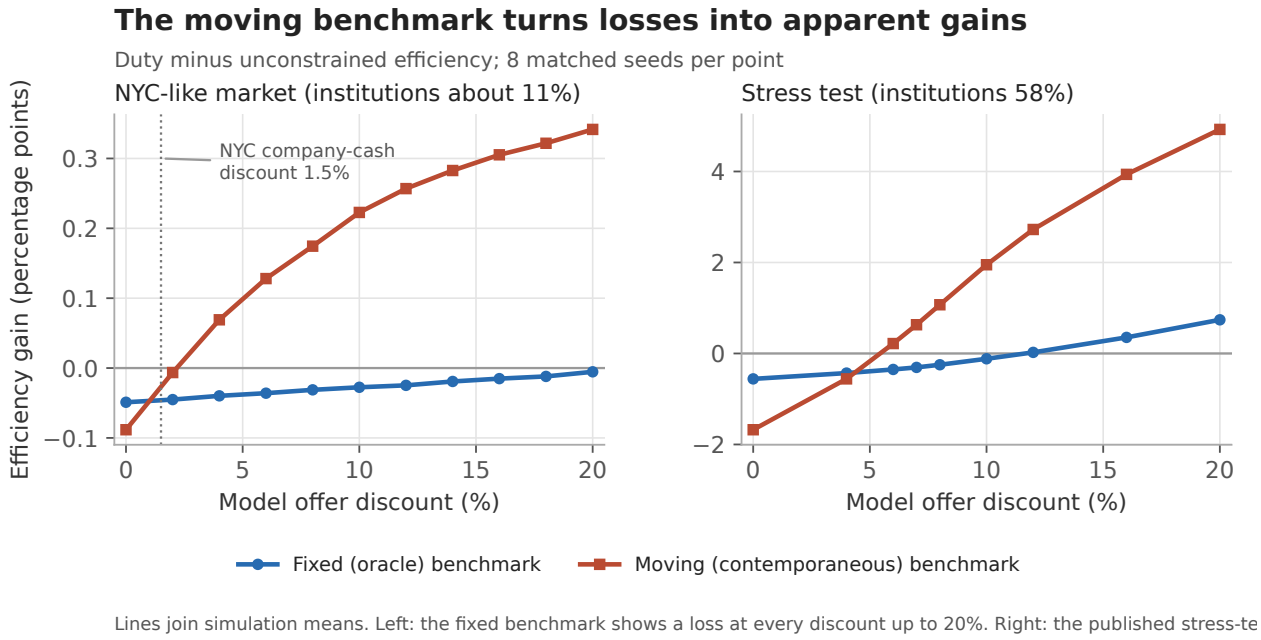


Figure C1. The duty's efficiency gain on the fixed (oracle) and moving (contemporaneous) benchmarks as the offer discount varies, in the main market (Table 3) and in the stress test (Appendix C). The horizontal axis is the model's offer discount, which differs from an empirical price premium. Lines join simulation means.

The benchmark gap widens substantially in the stress test (Figure C1, right panel). Here the contemporaneous benchmark averages about 521,000 per unit without intervention, 562,000 under the cap and 506,000 under the duty, a spread of about 11%. On the fixed benchmark the duty begins adding value at an offer discount of about 11.7% (95% interval [11.5, 11.8]), while on the moving benchmark it appears to add value starting around 5.4% [5.4, 5.5]. At the tested discounts of 6%, 8% and 10% the two measures disagree on the sign of the policy's effect, with the moving benchmark reporting gains of 0.22 to 1.95 points where the fixed benchmark records losses of 0.12 to 0.35. Adopting the moving benchmark therefore more than doubles the interval of offer discounts over which the duty appears self-financing.

Institutional share at w = 12%	Interpolated break-even w	95% interval
12.6%	none up to 20%	–
22.4%	15.9%	[15.6, 16.1]
29.9%	14.2%	[13.8, 14.7]
40.2%	12.5%	[12.2, 12.7]
46.8%	12.1%	[11.8, 12.3]
52.8%	11.9%	[11.8, 12.1]
58.1%	11.7%	[11.5, 11.8]
63.5%	11.9%	[11.7, 12.0]
68.3%	11.8%	[11.7, 11.9]
73.7%	12.0%	[11.8, 12.1]
79.9%	12.3%	[12.2, 12.4]

Table C1. Offer discount at which the duty stops losing value on the fixed benchmark, across eleven institutional-capacity settings, with the duty ceiling fixed at 0.22. The first row is the main market's capacity, where the fixed-benchmark gain stays negative across the whole grid up to 20% (Table 3), and the other ten rows are the stress test's capacity sweep. Institutional shares are unconstrained shares at a 12% discount, and crossings are linearly interpolated between grid nodes two points apart with eight replications per node. Intervals are 95% percentile intervals from 4,000 bootstrap resamples of the eight matched seeds (`threshold_intervals.py`), reflecting simulation noise alone while omitting linear-interpolation or model uncertainty.

Across the stress-test settings in Table C1 the break-even offer discount ranges from 11.7% [11.5, 11.8] to 15.9% [15.6, 16.1], whereas at the main market's lower institutional participation it is not reached below 20%. Below the stress test's institutional share, lower institutional participation generally implies a larger offer discount before the duty begins paying for itself on the fixed benchmark.

Table C2 decomposes the stress-test comparison into its numerator and denominator across a finer grid of offer discounts.

Offer discount	Value change (%)	B change (%)	Gain vs F (pp)	Gain vs B (pp)	Opposite signs
0%	−0.561	+1.134	−0.559	−1.676	0/8
4%	−0.433	+0.131	−0.431	−0.558	0/8
6%	−0.352	−0.573	−0.351	+0.218	8/8
7%	−0.307	−0.946	−0.306	+0.632	8/8
8%	−0.249	−1.332	−0.248	+1.071	8/8
10%	−0.117	−2.096	−0.117	+1.950	8/8
12%	+0.024	−2.754	+0.024	+2.725	1/8
16%	+0.356	−3.726	+0.353	+3.936	0/8
20%	+0.749	−4.483	+0.740	+4.925	0/8

Table C2. Fresh matched-seed sweep with eight seeds per offer discount. Value and reference changes are percentages based on ratios of totals, and ratio gains are percentage-point differences of mean per-seed ratios. F is the fixed top-valuations upper bound, and B is the policy-dependent sum of current maxima. The final column counts opposite signs for realised value and the moving-ratio change within a seed.

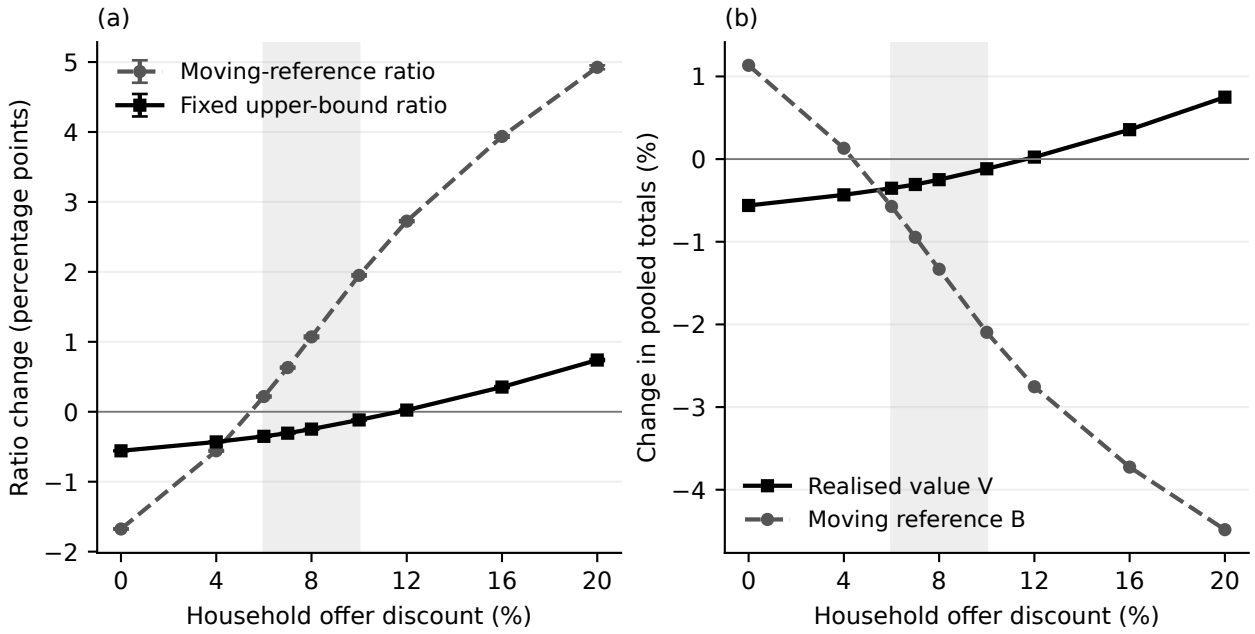
At w = 6%, the charge lowers mean realised value by 0.352% and lowers the pooled moving reference by 0.573%, so the moving ratio rises by 0.218 percentage points even as the fixed-reference ratio falls by 0.351 points. At w = 10%, the value loss shrinks to 0.117% while the moving reference drops by 2.096%, generating a moving-ratio gain of 1.950 points that reflects a compressed queue alongside a small loss in allocation value.

All eight seeds at each of the 6%, 7%, 8% and 10% nodes display these opposite signs, and at 12% one seed still combines a negative value change with a positive moving-ratio change even though both mean comparisons turn positive. Across the grid 33 of 72 paired scenarios exhibit opposite signs, though that count describes

these selected parameters and seeds instead of a general frequency estimate for housing policies or arbitrary model specifications.

At zero discount both mean comparisons are negative, and at 16% and 20% both are positive, indicating that the sign reversal is bounded instead of universal. A moving reference can therefore make a welfare loss appear as a gain over intermediate offer discounts while agreeing in sign elsewhere, although this parameter sweep cannot determine how often either pattern arises in other calibrated models.

Figure C2 separates the ratio changes from the changes in their components.



Changes in allocation ratios and their numerator and denominator

Figure C2. Changes in the two evaluation ratios and their components. Panel a reports matched-seed mean gains, with error bars showing plus or minus one paired Monte Carlo standard error across eight seeds. Panel b shows percentage changes in pooled realised value and pooled moving reference, where the reference drops faster than value at the four interior nodes with consistent sign reversal. Uncertainty covers simulation draws at fixed settings and excludes model or empirical parameter uncertainty.

Rerunning the sweep reproduces every archived field to numerical precision, yielding identical recovered fixed references within each pair and a largest residual in the logarithmic identity of approximately 2.31×10^{-16}. The matched inputs and recovered totals trace the reversal directly to the policy paths and their moving references.

Pooled and mean-per-seed moving-ratio comparisons also agree in sign at all nine nodes, differing by less than 0.0003 percentage points. That close agreement holds for this balanced exercise without making the two aggregation schemes equivalent, so studies with heterogeneous run sizes or disparate references should still state which weighting matches their evaluation question.

Appendix D. Access-population stress tests

In the stress test the instruments were tuned similarly to five- and ten-point full-run targets (Table D1, Figure D1). The cap again proves the most expensive way to purchase access, sacrificing about 2.5 index points for five points of access and 4.3 for ten. The fair share and the duty cost virtually nothing at five points, while at ten points the fair share delivers slightly more access than the duty and preserves more value, reversing their untuned ordering.

Target (pp)	Instrument	Setting	Access gain (pp; SE)	Value index (SE)	Constant − duty, value (SE)
5	Static cap	336	5.42 (0.121)	97.46 (0.048)	–
5	Fair share	0.5000	5.78 (0.086)	99.99 (0.007)	–
5	Signal duty	0.2115	5.03 (0.031)	100.03 (0.005)	–
5	Constant charge	0.0779	4.97 (0.041)	100.08 (0.005)	+0.051 (0.003)
10	Static cap	273	10.64 (0.150)	95.73 (0.061)	–
10	Fair share	0.4121	10.14 (0.100)	99.63 (0.014)	–
10	Signal duty	0.4657	10.05 (0.070)	99.32 (0.009)	–
10	Constant charge	0.1466	9.79 (0.034)	99.68 (0.008)	+0.361 (0.008)

Table D1. The stress test's instruments at common full-run access targets with settings tuned on four seeds and evaluated on sixteen held-out seeds. The cap setting is acquisitions per entity, the fair share is the institutional period share, and the charges are rates. The last column is the paired value difference between the constant charge and the duty using the delta method on matched seeds.

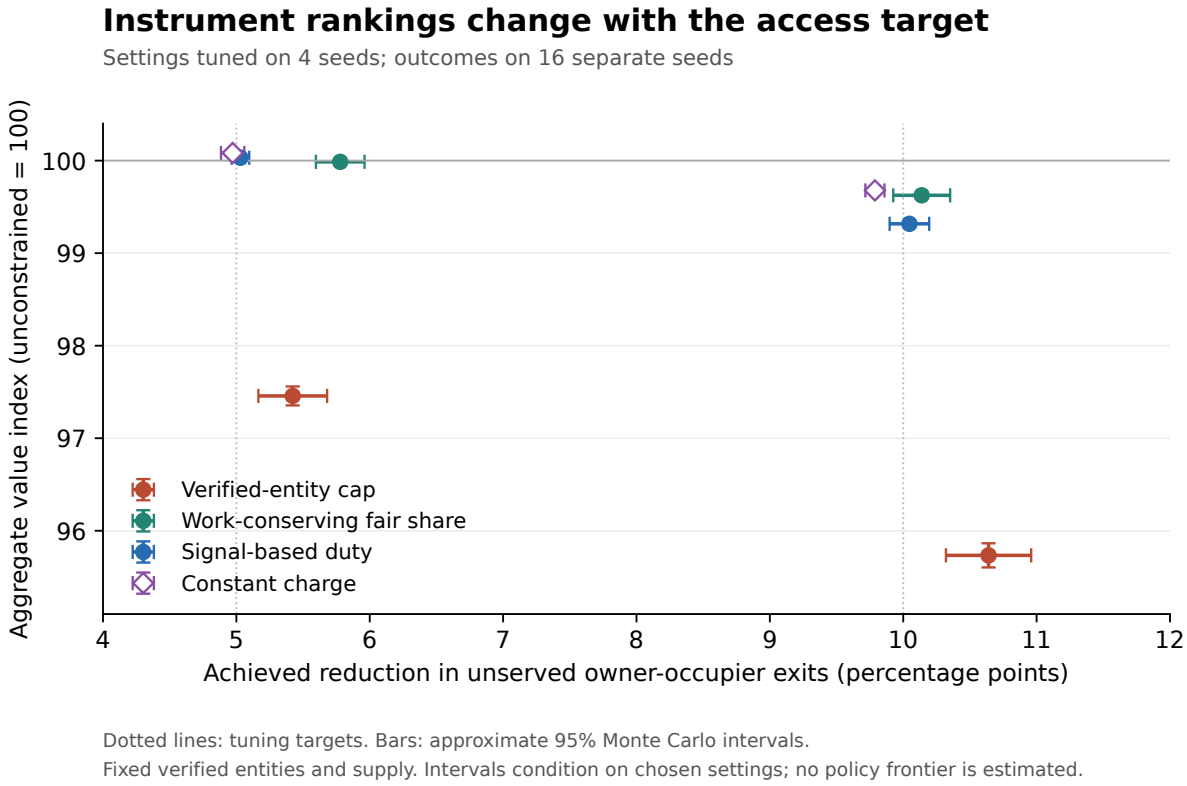


Figure D1. Stress test results with settings tuned to common full-run access targets and evaluated on sixteen separate seeds, including the constant charge. Bars show approximate 95% Monte Carlo intervals conditional on those settings, and points are not joined into a frontier.

The constant charge, which ignores contention, preserves slightly more value than the duty at both targets, by 0.05 and 0.36 index points (Table D1, Figure D1), though its access gains fall just short of the duty's so the comparison remains close without being exact.

Regime	Setting	Full-run gain (pp)	Cohort gain (pp; SE)	Cohort unserved (%)
Unconstrained	–	–	–	60.09
Static cap	336	5.42	13.53 (0.279)	46.56
Fair share	0.5000	5.78	7.91 (0.134)	52.18
Signal duty	0.2115	5.03	9.34 (0.072)	50.75
Constant charge	0.0779	4.97	7.15 (0.065)	52.94
Static cap	273	10.64	26.44 (0.318)	33.65
Fair share	0.4121	10.14	13.97 (0.164)	46.12
Signal duty	0.4657	10.05	18.04 (0.140)	42.05
Constant charge	0.1466	9.79	14.03 (0.091)	46.06

Table D2. Stress-test completed-cohort access at the settings of Table D1 and the common-target exercise, without retuning, where the cohort averages 4,061 entrants per replication.

The full-run measure understates every instrument's effect on this completed cohort by varying degrees (Table D2). Without intervention, 60.1% of the cohort leaves unserved against a full-run rate of 41.9%, partly because the full-run denominator includes entrants still searching when the simulation ends. For the cap, a 5.42-point full-run gain expands to a 13.53-point cohort gain (2.5 times larger), while the fair share's 5.78 rises to 7.91 (1.4 times larger). Rankings shift across definitions as well, with the fair share leading on full-run access at the five-point settings whereas the duty leads on cohort access, confirming that evaluations should specify which population they target.

Retuning to the cohort measure alters both settings and comparisons. Tuned on the original four seeds to five- and ten-point cohort gains and evaluated on sixteen held-out seeds, a five-point cohort gain requires a cap of 385 acquisitions per entity instead of 336. At the ten-point cohort target the ordering of the two charges flips: the duty retains 0.032 more index points than the constant charge (paired SE 0.0033) with virtually identical cohort gains (within 0.03 points, paired SE 0.044), whereas the constant charge still delivers 1.6 points more full-run access (paired SE 0.044). Each charge thus appears favoured when evaluated on the population for which it was tuned, while near five points the small value gap of 0.0012 (paired SE 0.0017) stays within Monte Carlo noise.

Appendix E. Instrument stress tests and sensitivities

Regime	OO share	Inst. share	Starvation	Value index	Oracle eff.	Contemp. eff.
Unconstrained	34.0%	58.1%	41.9%	100.00	99.5%	95.4%
Static per-entity cap	75.8%	6.3%	8.7%	88.72	88.3%	78.5%
Work-conserving fair share	54.5%	32.2%	24.7%	98.22	97.7%	95.2%
Contention-triggered duty	41.1%	49.4%	36.6%	100.03	99.5%	98.1%
Combined (share + duty)	54.7%	32.1%	24.6%	98.20	97.7%	96.3%

Table E1. Allocation regimes in the stress test reported as means across 12 replications, where OO and Inst. shares track purchases by owner-occupiers and institutions and the value index normalises total winner valuations relative to the unconstrained regime under matched seeds.

The instruments generate very different degrees of access expansion (Table E1). The duty lowers starvation from 41.9% to 36.6% alongside a value index of 100.03, and although its gain of about 0.026 index points carries a paired Monte Carlo standard error of 0.0054, the efficiency improvement remains economically minor. Starvation drops further to 24.7% under the fair share at a 1.8% value loss and to 8.7% under the cap at an 11.3% loss, while combining the fair share with the duty adds almost nothing because the fair share lowers the contention signal and mechanically dampens the duty's rate. Because the unconstrained market already

captures 99.5% of the oracle benchmark, aggregate efficiency headroom appears narrow while distributional shifts are large.

Whether a larger access gain justifies its valuation loss depends on welfare weights. Scoring each allocation with $W(\omega) = \omega V_{OO} + \sqrt{\omega}\, V_{SI} + V_I$, where V denotes realised valuation by class, the fair share overtakes the duty above $\omega \approx 1.16$ and the cap becomes the top-ranked of the three original instruments above $\omega \approx 1.66$, with the cap breaking even against no intervention near $\omega \approx 1.35$. These thresholds trace normative weighting schedules instead of empirical parameters, and `replication/welfarist.py` reports the full grid.

Two exercises from the extended working paper (Loschi 2026b) bear especially on the cap. When institutional capital can split across legal entities at zero cost, the cap's effect decays and vanishes at 25 entities, indicating that beneficial-owner aggregation would be essential for enforcement. When listings instead respond to seller proceeds with unit elasticity and sellers bear the full charge, the cap's value loss widens to 17.3% and starvation rises to 16.0%, while the duty incurs a 2.0-point value loss and remains the least costly option across every tested elasticity without being free.

Reported waiting times can mislead through the same selection channel. When costless entity splitting erodes a cap, the 99th-percentile wait among owner-occupiers who complete a purchase declines even as starvation rises, because buyers who would have waited longest exit unserved and drop out of the completed sample, leaving completed waits uninformative without the non-completion rate.

E.1 Additional numerical sensitivities

In the stress test, a limited spatial exercise allows institutions to reallocate a fixed acquisition budget across districts in proportion to $(1 - \tau_{max} s_d)^{\lambda}$. Across λ from zero to eight, the duty's value index ranges from 99.60 to 100.22 and its starvation rate from 36.5% to 36.7%, suggesting that even the sign of its small aggregate effect can flip once buyers arbitrage across locations.

Varying patience, the number of districts, valuation dispersion, the contention window or search friction leaves the full stress-test ordering intact, whereas arrival and listing rates alter it. Doubling owner-occupier arrivals to 34 per district-period reduces the unconstrained institutional share to 38%, allowing the fair share to score 100.20 and overtake the duty at 100.09 while the duty still reduces starvation by 4.0 points versus 2.9 points. The cap remains the most costly instrument at every tested setting (97.3 there), so while its position relative to contention-linked instruments persists across exercises, the ranking between fair share and duty does not, and these sensitivities remain untested in the main market.

The extended working paper (Loschi 2026b) documents the remaining sensitivity runs. Letting 10% of accepted financed sales fail preserves the main value indices and ordering alongside a nearly unchanged fixed-benchmark break-even (11.8% against 11.7%), while raising failed transactions by 21.6% under the duty and more than doubling them under the cap. Reducing search inspection probabilities to 0.5 and 0.25 leaves the instrument ranking intact, while sweeping institutional buyers' mean valuation from 340,000 to 410,000 shifts the duty's index by at most 0.06 points and expands the cap's loss from 8.6% to 16.0%. Because the stress-test institutional share of 58% lies well above observed markets, lowering that share to 22.5% leaves the duty and fair share lowering starvation by just 0.7 and 0.1 points while the cap remains the costliest option across all shares tested. Finally, the rate sweep reaches its highest mean value index at $\tau_{max} = 0.11$ (100.12) above the baseline 0.22 (100.02), and at an offer discount of 5.6% all three ceilings (0.11, 0.22 and 0.44) lose value.

Appendix F. New York City proxies and sensitivity

F.1 Proxy evidence

The main market is calibrated to New York City proxies drawn from a study of recorded sales (Loschi 2026a) and its linked dataset (Table F1). In that linked dataset, company cash buyers account for 19,019 of 166,227 house sales of at least $100,000 with an ORDINARY linked deed, or 11.4%, and all company buyers account for 22.8% (Loschi 2026d), whereas one bid per entity per district-period generates an unconstrained institutional share of 11.0% in the model. The company-cash gap of 1.53 log points sets the baseline offer discount to 1.5%, and its upper confidence endpoint gives 5.3%. Because the estimated price-gap interval includes zero, Appendix F varies both inputs, treating the comparison strictly as an input check without testing any tax.

Both proxies remain imperfect, and neither identifies institutional buyers with certainty. Company buyers include LLCs formed by individuals, the absence of a matched mortgage does not prove cash financing, and transaction-price gaps cannot separate a certainty premium from buyer composition or real completion costs. The main market therefore captures a plausible order of magnitude without offering an empirical estimate of New York City's market, and supply, arrivals, patience and valuations remain illustrative in both markets.

The stress test uses 42 institutional bids per district-period, an unconstrained institutional share of 58.1% and a 12% offer discount. Because institutional purchases run at only a few per cent of sales nationally, this scenario sits far outside observed markets and illustrates how the measurement diagnostics behave when institutions dominate and every instrument exerts large effects. Among the house groups in Table F1, only non-company cash buyers display a discount near 12%, at 12.9%, and excluding same-surname deeds reduces it to 6.9%, while mortgage denials precede accepted offers and cannot identify subsequent failure risk.

Market	Log-price gap, 100β	95% interval	Approx. cash discount
One- to three-family houses, headline	9.25	[7.17, 11.35]	8.8%
Houses, excluding same-surname deeds	4.87	[2.69, 6.76]	4.8%
Houses, cash buyer is a company	1.53	[−1.79, 5.44]	1.5%
Houses, cash buyer is another party	13.77	[10.96, 15.98]	12.9%
Houses, other party, excluding same-surname deeds	7.11	[4.48, 9.12]	6.9%
Condominiums	0.83	[0.06, 1.61]	0.8%
Co-operatives	−1.42	[−3.70, 0.86]	−1.4%

Table F1. Financing-related price gaps in New York City over 2016–2025 (log points, source Loschi 2026a), where positive values indicate higher adjusted prices for financed purchases and cash discounts follow 100[1 − exp(−β)]. House intervals are parcel-bootstrap intervals and the co-operative interval clusters by building, while differing sample constructions across property types prevent a matched comparison across tenures.

F.2 Proxy scenarios

Table F2 varies the proxy inputs while holding instrument settings at their defaults. Here the constant charge equals the duty ceiling multiplied by its mean signal across district-periods, matching an unweighted scheduled haircut without equating charges on completed purchases or access (Sections 6.1 and Appendix D provide comparisons at common access targets).

Raising the offer discount to the upper endpoint of 5.3% or to the headline gap's 8.8% alters the results only modestly: full-run starvation without intervention rises to 10.6% and 11.1%, and the duty still lowers it by about 0.2 points. Treating all company buyers as institutions raises their simulated share to 25.0% and full-run starvation to 18.8%, causing every instrument to bite more strongly and leading the cap to cut starvation by 11.29 points at a 5.16% value loss. Under the upper-gap and headline-gap scenarios, the contemporaneous benchmark yields apparent duty gains of 0.11 and 0.19 points whereas the fixed oracle records losses of 0.04 and 0.03 points.

Scenario	Discount / model share (%)	Unserved: full / cohort (%)	Cap	Fair share	Duty	Constant
Company cash	1.5 / 11.0	9.8 / 14.9	−1.77 / −1.36	0.00 / 0.00	−0.20 / −0.05	−0.24 / −0.05
Upper gap	5.3 / 11.5	10.6 / 16.0	−2.15 / −1.47	0.00 / 0.00	−0.21 / −0.04	−0.26 / −0.04
All companies	1.5 / 25.0	18.8 / 27.2	−11.29 / −5.16	−0.46 / −0.06	−1.09 / −0.18	−1.21 / −0.17
Headline gap	8.8 / 12.1	11.1 / 16.8	−2.53 / −1.56	0.00 / 0.00	−0.20 / −0.03	−0.25 / −0.03
Stress test	12.0 / 58.1	41.9 / 59.8	−33.16 / −11.28	−17.14 / −1.78	−5.24 / +0.03	−4.85 / +0.10

Table F2. NYC proxy scenarios and the stress test across twelve matched seeds. Each instrument cell reports the change in full-run starvation (percentage points) alongside realised value (percent) relative to that scenario's unconstrained market, where negative starvation indicates improved access. The unserved columns report unconstrained full-run and completed-cohort rates for the cohort entering in periods 36 through 83 and resolving by period 119. Instruments retain their baseline settings and constant charges match the duty's mean scheduled haircut, leaving simulated shares as approximations to empirical targets (`nyc_anchored.json` supplies full-run results and `nyc_cohort_check.json` supplies cohort results, seed-level runs and paired Monte Carlo standard errors).

Starvation remains a model construct and the empirical company shares come from a selected set of linked transactions, so these two headline inputs cannot establish policy effects in New York City. New York also levies mortgage recording taxes on financed purchases, averaging 1.41% of price for houses and 1.28% for condominiums at headline borrower rates before credits, exemptions and consolidation adjustments (Loschi 2026c), a margin outside the model's cash-charge comparison.

Appendix G. Implementation requirements

Each instrument presupposes administrative information that the stylized model takes for granted.

The duty needs an exit registry. Its signal counts owner-occupiers who leave without buying, requiring a register that observes market entry and separates purchase, withdrawal and moves to another jurisdiction, outcomes that routine deed records never show. That signal could plausibly be distorted from either side, since institutions might suppress it by withdrawing bids during the measurement window while beneficiaries of a higher rate could stage nominal entries or withdrawals. A workable design would therefore restrict counts to verified participation such as a mortgage pre-approval or deposit, house the register with an independent custodian, and cap period-to-period rate jumps (analogously to the manipulation-resistant decision rules studied by Björkegren, Blumenstock and Knight 2026). The constant charge requires no exit registry, underscoring the practical trade-off behind Section 6.1: the duty's value advantage appears small, shifts with the target population, and must be weighed against measurement and gaming costs.

The cap needs ownership aggregation. Because a per-entity limit binds a legal person, an owner able to form entities costlessly neutralizes the stress-test cap with 25 shell entities. An enforceable ceiling would therefore require statutory common-control definitions, beneficial-owner disclosure, verification and sanctions, which have become harder to obtain since federal beneficial-ownership rules exempted domestic companies in 2025 and shifted the burden onto state law. A per-purchase charge avoids the incentive to split entities, even though it still relies on an enforceable taxpayer definition.

The fair share's release rarely binds. The work-conserving quota lifts only when no owner-occupier is waiting, so because someone is always waiting in the stress test and the default quota never binds in the main market, the simulation leaves its slack-market release largely untested.

Classification decides who is reached. Payment method, institutional status and owner-occupation represent distinct legal categories, as some owner-occupiers pay cash while some small investors borrow. Whereas the model charges institutions specifically, the May 2026 New York proposal (targeting all-cash residential purchases of $1 million or more before being omitted from the budget passed on 27 May; Bloomberg, 14, 21 and 27 May 2026) relied on a payment-method test vulnerable to short-term bridge financing repaid after closing. A broader tax on all non-owner-occupiers would also sweep in small investors and require separate evaluation.

Coverage and geography matter. A tax reaches only the transfers specified in its statute. New York's mortgage recording tax applies to real-property mortgages and thus excludes co-operative share loans, whereas the mansion tax expressly includes co-operative units. Buyers can likewise shift demand across municipal boundaries in ways the limited district exercise of Section 10 only begins to illustrate.

Appendix H. Housing parameter details

Civilian arrivals average 17 owner-occupiers and 4 small investors per district-period while listings average 20, with lognormal valuations centred near 400,000 and 372,000 ($\sigma = 0.34$ and 0.30) and patience set to 36 and 24 periods. Institutional valuations centre near 366,000 with $\sigma = 0.18$, and while the main market simulates one bid per institutional entity per district-period, the stress test submits 42 institutional bids in total alongside an unconstrained institutional share of 58.1% and a 12% offer discount.

A financed price premium p corresponds to offer discount $w = p/(1 + p)$, and a log-price gap β corresponds to $w = 1 - \exp(-\beta)$. The duty signal averages 0.094 without intervention, 0.092 under the duty and 0.076 under the cap in the main market, whereas the stress-test signal spans 0.078 under the cap to 0.406 without intervention (0.341 under the duty), differing from a naive lost-bid share that saturates near one under heavy waiting-buyer competition.

H.1 Main-market access display

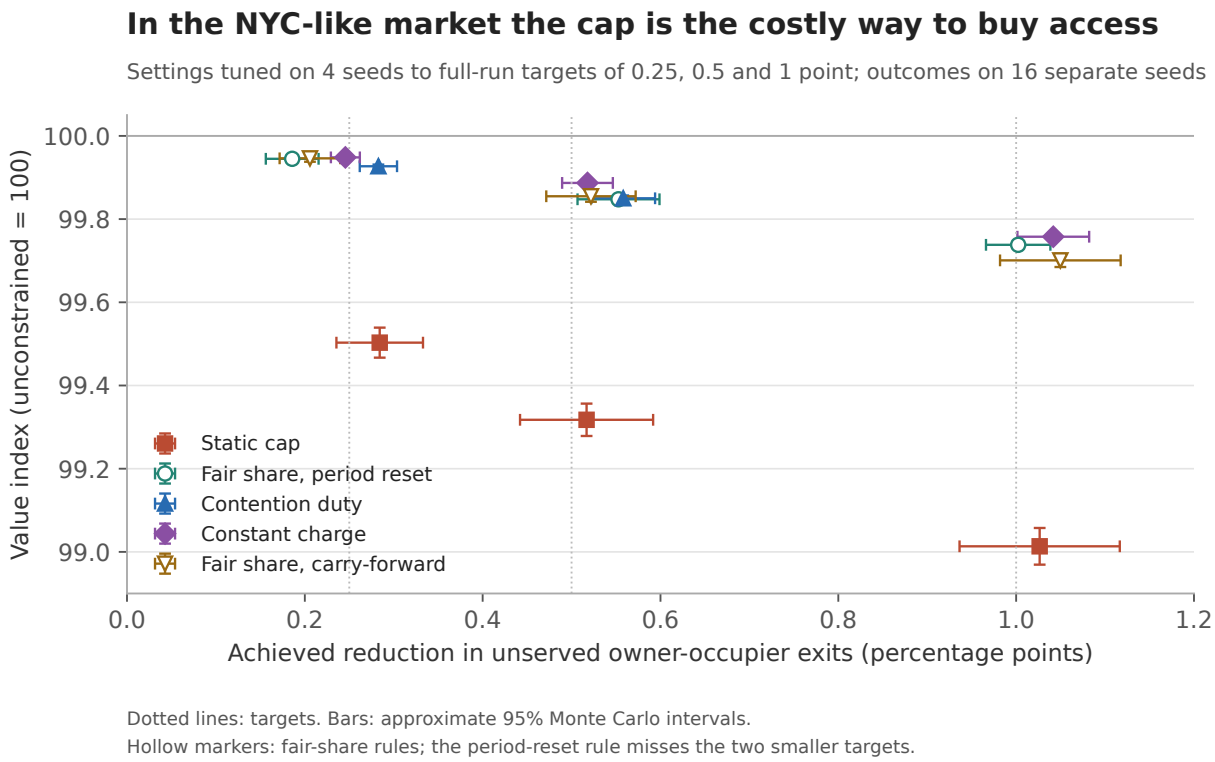


Figure H1. Settings tuned to common full-run access targets in the main market and evaluated on sixteen separate replications. Bars depict approximate 95% Monte Carlo intervals conditional on the tuned settings, and points are not joined into a frontier.

Appendix I. Provisional PolicySpace2 illustration

I retain this archived illustration strictly for transparency, as its estimates have not been reproduced from the original monthly histories and play no role in the abstract, main evidential claims or conclusion.

I.1 Archived comparison

PolicySpace2 (Furtado 2022) is a published agent-based model of a Brazilian metropolitan economy with open code. Furtado compares three poverty policies funded by the same share of municipal revenue: buying houses and giving them to the poorest registered families, 24-month rental vouchers, and a monthly cash transfer shared among all registered families. The paper reports that "on average, Monetary aid policy helps

2,358.05 households monthly per municipality, compared to 47.5 for Rental vouchers and 12.37 for the Property acquisition program" (Furtado 2022, para. 5.3), describing monetary aid as reaching "a larger number of households" (para. 5.9). Because that headcount resets each month, monetary aid counts everyone on each monthly register whereas property acquisition counts only the houses handed over during that month.

The archived audit uses the published code at its default Brasília setting, adding instrumentation intended to log eligibility, assistance and poverty status without altering policy behaviour (Appendix I.2 documents provenance, missing logs and non-determinism). Saved summaries sit close to the paper's monetary-aid figure at 2,363 households a month on average (2,353–2,373 across replications), alongside 54 for vouchers and 14 for property acquisition.

Table I1 recounts those archived runs along three distinct denominators. On the monthly flow count, monetary aid reaches about 170 times as many households as property acquisition in the archived summaries (190 times in Furtado's published figures), whereas counting distinct families assisted at least once over ten years reduces the ratio to 2.3 because monetary aid repeatedly pays the same families (75 payments per family on average) while each house transfers to a new family. A third cohort metric tracks families first falling below the eligibility line in months 13 to 36 and checks how many remain below that threshold 24 months later: without intervention 93.6% remain below, compared with reductions of 1.8 points (SE 0.7) for monetary aid, 2.6 (0.4) for rental vouchers and 6.9 (0.7) for property acquisition across paired seeds, suggesting a potential ranking reversal across denominators.

Policy	Households helped per month (reported)	Distinct families helped, 2011–2019	Newly poor helped within 24 months	Newly poor still below the line after 24 months
No policy	–	–	–	93.6% [91.5, 94.8]
Monetary aid	2,363 [2,353, 2,373]	3,405 [3,354, 3,479]	100%	91.8% [88.5, 93.9]
Rental vouchers	54 [52, 56]	1,712 [1,659, 1,786]	57.5% [55.8, 62.7]	91.0% [89.1, 94.5]
Property acquisition	14 [12, 15]	1,468 [1,320, 1,586]	17.0% [13.5, 19.9]	86.7% [84.3, 88.1]

Table I1. PolicySpace2 policies counted three ways (archived summaries for Brasília, 1% of the population, 2010–2020 at default parameters with policies starting after the first year, reporting means over ten replications and ranges across replications in brackets). Newly poor refers to families whose permanent income first falls below the 20th percentile in months 13–36 (395 per run without policy and 673 under property acquisition).

These PolicySpace2 comparisons remain provisional because the archived simulations have not been reproduced, and the entry-cohort contrast requires the same population caution emphasized in Section 6. Property acquisition alters who falls below the poverty line, raising its entry cohort to 673 families per run against about 400 under the other policies. Its lower persistence therefore partly reflects compositional shifts instead of pure exit effects, while across ten replications voucher and monetary-aid persistence cannot be distinguished statistically (paired difference 0.8 points, SE 0.7). The 2.3 ratio and apparent reversal are unverified properties of archived summaries that cannot serve as evidence about real-world policies without the original monthly histories or a matching replay, underscoring why reach metrics should state whether they count monthly payments, unique families or poverty exits.

I.2 Provenance and structural non-determinism

The original audit uses commit `ced77dd`, Brasília, 1% population, 2010–2020, policy share $\delta = 0.2$ and eligibility below the 20th percentile of permanent income, with seeds 1–10 for each policy and a compatibility shim targeting Python 3.6 `random.sample` behaviour on dictionary views. While the saved per-run summaries reconstruct Table I1, exhaustive checks across local directories, Docker containers and volumes, Git history and all four benchmark-audit Zenodo versions found none of the 40 original monthly logs (`audit_*.jsonl`); the lone saved `audit_wage_1.jsonl` (`12,476,844` bytes) in `Methods_v7_3_audit_evidence.zip` matches the later 11 November local Docker replay byte-for-byte.

Two full 120-month replays of monetary aid (`wage`, seed `1`) in isolated Python 3.6 environments with identical 186-package Linux conda stacks illustrate why Table I1 cannot be reproduced historically. In the 11 November Docker replay, eight of the eleven archived summary fields differed (mean monthly assistance `2360.9` vs. `2357.9`, total assistance counts `254,981` vs. `254,649`, entry cohort `409` vs. `382`), halting any multi-

seed extension; that run is retained under `policyspace2/later_replay/` in the audit-evidence archive. On 15 November, a second independent rebuild from the same package manifest (`PYTHONHASHSEED` unset, `1,403s`) produced a third distinct realization at `256,334` total assistance counts (`2,373.5` monthly, `444` cohort, `0.8986` still below the threshold after 24 months). Neither later run replaces the archived Table I1 summaries.

The repaired analyser, `analyze_ps2_audit.py`, enforces the full declared policy-by-seed file set, consecutive calendar months and complete follow-up instead of returning an empty success on missing logs, includes zero-assistance months in the calendar denominator, and counts any assistance from cohort entry through observation 24 months later. Although that timing rule guards against histories recording assistance prior to the first logged poverty month, synthetic regression cases confirm its effect only in principle because no such case occurs in the retained replay and its incidence in the original runs cannot be verified without their logs. On the saved replay run, the repaired analyser leaves all eleven shared summary fields unchanged and all ten unit-test cases in `test_audit_contract.py` pass.

Inspection of Docker metadata and source files alongside seven controlled container probes (`SEED=1`, `POLICIES="wage"`, same `186`-package image) identifies three structural mechanisms behind the divergence inside unmodified `ced77dd`. First, Python hash randomization (`PYTHONHASHSEED`) alters unordered `set` iteration orders across `simulation.py` (`mun_to_regions`), `agents/family.py` / `world/demographics.py` (`relatives`), and `world/funds.py` (`set(Family)` sorting): loading identical cached `.agents` at `SEED=1` and changing only `PYTHONHASHSEED` from `1` to `2` immediately shifts initial vacant and total houses (`909 / 11,267` vs. `915 / 11,266`) and Months 1–3 poverty headcounts (`[2067, 2084, 2094]` vs. `[2059, 2081, 2092]`). Second, loading an existing `.agents` cache skips `create_regions()` and `create_all()` while keeping the same `self.seed` stream (`simulation.py`), immediately shifting initial states even with identical `PYTHONHASHSEED=1` and `SEED=1` (`909 / 11,267` cached vs. `913 / 11,265` fresh `run_a`), and the cache filename omits `SEED`. Third, `world/generator.py` draws `uuid.uuid4()[:12]` identifiers outside `self.seed`: across two paired 26-month probes (`2010-01-01` through `2012-02-01`, 14 active policy months) pinning both `.agents` and `PYTHONHASHSEED=1`, all five CSV outputs (`temp_banks.csv`, `temp_construction.csv`, `temp_firms.csv`, `temp_regional.csv`, `temp_stats.csv`) match byte-for-byte and all 26 monthly policy counts agree (`rep_sum = 36,726`), yet family IDs in `audit_wage_1.jsonl` still diverge on all 26 days (`diff=132` with `only_a=66, only_b=66` on `2010-01-01`, rising by `2012-02-01` to `1,706` for poverty registers and `1,818` for eligible and helped registers) as new marriages receive fresh random UUIDs. Exact family-level reproduction of Table I1 is therefore blocked structurally in unmodified `ced77dd` rather than by an unrecovered Python 3.12 package lock, leaving PolicySpace2 strictly as a provisional, non-evidential appendix unless upstream code is modified with deterministic identifiers, sorted set iterations and seed-keyed caches for fresh multi-seed runs. Records of these probes are preserved in `evidence/provenance/`.