When the Benchmark Moves with the Policy

Measuring Housing-Access Instruments in Allocation Simulations

Pablo Loschi · Independent researcher

Working paper, version 7.2 · 29 October 2026 · Merges The Execution-Certainty Wedge (v6.5) and Auditing Policy-Dependent Benchmarks (v1.0.1)

Abstract

An allocation policy changes which agents remain available for later decisions. A benchmark built from that changing pool can therefore move with the policy. I show how this and a second measurement choice, the population over which access is counted, change the ranking of housing-access instruments in an agent-based market calibrated to New York proxies. A ratio identity and an exact two-period example give the mechanism, in which the moving-benchmark ratio rises from 0.80 to 1.00 while realised value falls from 16 to 11. In the calibrated market a contention-linked duty appears to add value above a 2.2 percent offer discount under the moving benchmark, while a fixed benchmark shows a loss at every discount up to 20 percent. A deadline queue produces the opposite error. Following a fixed entry cohort changes measured access gains by factors of 0.7 to 2.5 and reorders the instruments. In a published model of Brazilian poverty policies, counting distinct families instead of monthly payments cuts the advantage of cash transfers over house purchases from about 170-fold to 2.3-fold, and in a published liver-allocation simulator broader sharing widens geographic differences among transplanted patients and new listings while narrowing them among patients already waiting. The certainty discount itself explains little of the access shortfall.

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

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

---

1. Introduction

A seller choosing between two offers for the same home may take the lower one because it is more likely to close, as cash offers usually are. When the buyers who can pay cash are mainly investors and those who need a mortgage are mainly households who would live in the home, the seller's preference can decide who gets housing. In May 2026 New York legislators considered a 1% tax on all-cash residential purchases of $1 million or more. It was reported as likely to be dropped a week later and was left out of the budget passed on 27 May (Bloomberg, 14, 21 and 27 May 2026). The proposal raises a question that does not depend on its fate. If a charge on one class of buyer is meant to change who wins, how should its effects be measured?

Simulation is the natural tool, because the counterfactual allocations cannot be observed. It is also where the measurement choices are easiest to make without noticing. The housing simulation studied here reports an efficiency ratio against the best agent available at each decision and an access rate over arrivals during a finite run. The first comparison uses a policy-dependent reference; the second includes buyers still waiting when observation ends. A policy that serves a high-value buyer early removes that buyer from the queue on which later benchmarks are computed, so its ratio can improve while total value falls. A policy can leave more arrivals still searching at the horizon, so the full-run rate combines completed outcomes with buyers whose outcomes remain unresolved. Online scheduling avoids the problem by scoring policies against an offline optimum on the same input (Kesselman et al. 2004), and policy regret fixes the history a policy is judged against (Arora, Dekel and Tewari 2012). I bring that discipline to agent-based housing-policy evaluation.

I measure how much those choices matter, starting in Section 5 from an identity: 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 falls. An exact two-period example shows the moving-benchmark ratio rising from 0.80 to 1.00 while value falls from 16 to 11. I then apply the diagnostic to a housing market with owner-occupiers, small investors and institutional buyers, in which sellers rank offers by valuation times a certainty factor and four instruments are compared with an unconstrained market. They are a cap on acquisitions per legal entity, a work-conserving fair-share quota, a constant charge on institutional offers, and a

duty whose rate rises with local contention. The main market is set to New York City proxies, with institutions winning about 11% of sales and financed offers carrying a 1.5% discount, and a market in which institutions win 58% of sales serves as a stress test.

In the main market a fixed benchmark shows the contention duty losing value at every offer discount up to 20%, while the moving benchmark shows it gaining from a discount of about 2.2%. In a deadline queue with an exact offline optimum the error runs the other way, as earliest-deadline-first raises value from 19 to 27 while its moving ratio falls from 1.00 to 0.90. Counting access over a cohort followed until purchase or exit changes measured gains by factors of 0.7 to 2.5 and reorders the instruments, so that the duty and a constant charge swap places depending on which population they are tuned to. A fair-share quota that resets each period cannot reach small access targets because it admits purchases in whole numbers, while allowances that carry forward can.

On the substance, the certainty wedge matters little at the New York proxy values, since removing the discount entirely lowers the share of owner-occupiers who leave without buying from 9.8% to 9.5%, while excluding institutions lowers it by 6.6 points. The per-entity cap buys access at four to nine times the realised-value cost of the charges.

Section 2 places the paper in the literature, Section 3 sets out the model and Section 4 the seller's problem. Sections 5 to 7 take the three measurement choices in turn, Section 8 reports what the instruments do, Section 9 states how to use the audit, Section 10 sets out what the model leaves out, Section 11 what each instrument would need in practice, and Section 12 concludes.

2. Related work

This paper sits between housing finance, market design and public finance. Reher and Valkanov (2024) and Han and Hong (2024) supply the empirical motivation and disagree on how much of the cash premium is excess over risk, and that disagreement is why I sweep the offer discount instead of transporting one estimate. Che and Gale (1998) show that auctions with budget-constrained bidders can allocate inefficiently. The mechanism here runs through the seller's preference for certainty instead, and it operates even when the buyer's willingness to pay is unconstrained. Budish, Cramton and Shim (2015) study how market design can reward an advantage other than valuation, in their case speed. Completion certainty differs because it can save real costs. Piazzesi, Schneider and Stroebel (2020) show that housing search is segmented across clienteles, and the model's districts are a crude version of that segmentation. Kopczuk and Munroe (2015) show that New York's mansion tax, a notch in a transfer tax, moves prices and volume around its threshold, so a charge on transactions changes the transactions on which it is later measured.

The policy question is a second-best one in the sense of Lipsey and Lancaster (1956): a charge in an already distorted market can raise or lower welfare, and nothing general follows. Comparing a quantity rule (a cap or quota) with a price rule (a charge) follows Weitzman (1974), and the carry-forward quota borrows the logic of banking allowances analysed in Weitzman (2018). The treatment of distribution follows the practice of stating welfare weights explicitly (Atkinson 1970) and of reporting the trade-off between efficiency and fairness instead of assuming one objective (Bertsimas, Farias and Trichakis 2012). The method is agent-based in the sense of Tesfatsion (2006) and Axtell and Farmer (2025), who survey its use in economics, and Geanakoplos et al. (2012) use an agent-based model of the housing market to study systemic risk. Recent housing models evaluate loan-to-value and debt-to-income limits against a baseline run, either the historical settings (Yun and Moon 2020) or a pre-shock steady state (Gamal et al. 2024), and report prices, volumes, rents or ownership rates; the access measures I study add a reference group that itself depends on the policy. Bidding follows a fixed rule with no learning or equilibrium strategy, which is the model's largest limitation. In the terms of Edmonds et al. (2019), the housing model serves theoretical exposition with illustrative parameters rather than prediction, and the New York proxies set its scale without validating it.

The measurement findings have close relatives outside housing. Arora, Dekel and Tewari (2012) show why judging a policy against the history that another policy produced can mislead, and define policy regret to correct for it. Perdomo et al. (2020) study decision rules that change the distribution on which they are evaluated. Bertsimas, Farias and Trichakis (2011) measure the cost of fairness against a common feasible optimum, the role the fixed benchmark plays here. In queueing, a waiting-time statistic computed only over served requests

hides the ones that were dropped. Allocation through queues has its own literature in housing: Waldinger (2021) uses public-housing wait lists in Cambridge, Massachusetts to show that allocation rules trade targeting against tenant welfare, Leshno (2022) shows that welfare in overloaded waiting lists depends on how expected waits ration units, and Thakral (2019) studies efficiency and fairness when units arrive at random. Those papers allocate public units by rule, whereas here sellers choose among private offers, and the queue matters through who is still waiting when a listing arrives. I do not apply these results directly; my contribution is to show how much these choices matter for ranking housing-access instruments.

The deadline queue of Section 5.4 is a bounded-delay scheduling problem: Kesselman et al. (2004) analyse jobs with values and deadlines, compare online policies against an offline optimum on the same input and give efficient offline algorithms, so an optimum that respects timing and capacity is the standard reference there. The small dynamic program I use is a verification tool and makes no claim to be a new scheduler. Windrum, Fagiolo and Moneta (2007) separate the empirical validation of agent-based models from program verification, and the checks in this paper are of the second kind. Furtado's (2022) PolicySpace2 shows the broader modelling a policy application would need, with interacting labour, credit and housing markets on empirical spatial inputs, and Appendix A describes the housing program following the ODD protocol (Grimm et al. 2020).

3. The model

3.1 Market

The simulation is a discrete-time market with five districts over 120 periods, with illustrative supply, arrivals and valuations. Each district lists a Poisson-distributed number of units averaging 20 per period. Three classes of buyer bid:

- **Owner-occupiers** arrive at 17 per district per period, draw valuations from a lognormal distribution centred near 400,000 ($\sigma = 0.34$), and leave unserved after 36 periods.
- **Small investors** arrive at 4 per district per period, centred near 372,000 ($\sigma = 0.30$), with 24 periods of patience.
- **Institutional buyers** are three standing entities with valuations centred near 366,000 ($\sigma = 0.18$). In the main market each entity places one bid per district per period, while in the stress test the three together place 42.

Sellers rank offers by effective value, the bid multiplied by a certainty factor: 1.00 for institutional cash offers, c_{OO} for financed owner-occupiers and $c_{SI} = \min(1, c_{OO} + 0.07)$ for small investors. The owner-occupier offer discount is $w = 1 - c_{OO}$: 1.5% in the main market and 12% in the stress test, and Section 5 sweeps it. A financed price premium p over the cash price corresponds to $w = p/(1 + p)$, and a log-price gap β to $w = 1 - \exp(-\beta)$. Units go to the highest effective offer in the eligible pool, and I solve no bargaining equilibrium.

3.2 Two markets

The main market is set to New York City proxies from the companion study and dataset (Table 1). In the 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). One bid per entity per district-period gives an unconstrained institutional share of 11.0% in the model. The company-cash gap of 1.53 log points sets the offer discount to 1.5%, and its upper confidence endpoint gives 5.3%. The price-gap interval includes zero, and Section 8.3 varies both inputs; the comparison is a calibration check and tests no tax.

Both proxies are imperfect. Company buyers include LLCs used by individuals, the absence of a matched mortgage does not prove cash financing, and transaction-price gaps do not separate a certainty premium from composition or completion costs. Neither proxy identifies certain institutional buyers. The main market therefore gets the order of magnitude right without being an 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. Institutional purchases run at a few per cent of sales nationally, so this market is far from observed ones. It shows how the measurement results behave when institutions dominate and every instrument has large effects. Of the house groups in Table 1, only cash buyers other than companies show a discount

near 12%, at 12.9%, and excluding same-surname deeds reduces it to 6.9%. Mortgage denials, which precede accepted offers, 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 1. Financing-related price gaps, New York City 2016–2025, in log points; positive values mean higher adjusted prices for financed purchases. Cash discounts are 100[1 − exp(−β)]. Source: Loschi (2026a). House intervals are parcel-bootstrap intervals, and the co-operative interval clusters by building. Property types differ in sample construction, so the rows are not a matched comparison across tenures.

3.3 Instruments

Unconstrained. Sellers accept the highest effective offer.

Static per-entity cap. Each legal entity may acquire at most 50 units per district over the run.

Work-conserving fair share. Institutional buyers may take at most 30% of a district's listings in a period while owner-occupiers are waiting, and the limit lapses when none is waiting.

Contention-triggered duty. Institutional effective offers are multiplied by $(1 - \tau_{max} \cdot s)$, with $\tau_{max} = 0.22$. The signal s is measured on exits from the local owner-occupier cohort. Over a trailing window of six periods, let U be the number who left without buying and V the number who bought; then $s = U/(U + V)$, with $s = 0$ when the window is empty, and buyers still searching count in neither term. The signal counts exits because a bid-based version, the share of owner-occupier bids that lose, saturates near one in a market with hundreds of waiting buyers and twenty units per period, and would define a different instrument. In the main market the realised mean of s is 0.094 without intervention, 0.092 under the duty and 0.076 under the cap. In the stress test it runs from 0.078 under the cap to 0.406 without intervention, and is 0.341 under the duty. The multiplier is a reduced-form offer haircut, distinct from a statutory tax rate, and I compute no revenue.

Constant charge. Institutional effective offers are multiplied by $(1 − h)$, whatever the contention. Its rate h is set either to the duty's mean scheduled haircut or by tuning to an access target. It needs no registry of entries and exits, although classifying eligible buyers and enforcing the charge still require administration.

Combined. The fair share and the duty together.

3.4 Measures

The primary outcome is the **realised-value index**: the sum of successful buyers' valuations, normalised to the unconstrained regime under matched seeds. It measures realised willingness to pay, which is narrower than social welfare because it omits renters, the use of any revenue, administrative costs and differences in the marginal utility of income.

Two efficiency ratios put that sum over a benchmark. **Oracle efficiency** divides by the sum of the L highest valuations among all bidders generated in the run, where L is the number of units listed. With full search and exogenous listings, matched seeds give every policy the same denominator. The oracle sum ignores arrival, expiry and geography, so it is an upper bound F on the best feasible realised value V*, with $V \le V^* \le F$. A ratio V/F near one therefore shows that little value is left unrealised under this objective, but it does not identify the exact feasible gap. **Contemporaneous efficiency** divides by the sum, over listings, of the highest valuation among the bidders present at that listing. That denominator depends on who is still waiting, and so on the policy.

Access is measured by the **starvation rate**: owner-occupiers who leave without buying, divided by all owner-occupier arrivals during the run. Buyers still searching at the horizon stay in the denominator, and Section 6.1 contrasts this with a completed-cohort measure. Results are means over twelve independent replications unless stated otherwise.

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

4.1 How much discount risk justifies

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

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

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

$$\pi^* = (qd + \kappa) / (1 - q + qd + \kappa). \quad (2)$$

It depends on the failure probability, the cost of a failed sale and the difference in closing times, and it is positive even when no sale fails (at $q = 0$ it equals $\kappa/(1 + \kappa)$). Only a discount beyond π^* is a candidate distortion, and neither a price gap nor a mortgage denial rate identifies π^* without local evidence on accepted offers that fail. Both costs can 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 and a much lower sale hazard, which raises d for a seller whose financed buyer fails. Time to sell also moves with the cycle (Head, Lloyd-Ellis and Sun 2014), so κ is larger in slow markets.

4.2 Which bids a proportional charge displaces

Consider one listing with fixed offers and eligibility. 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 $-G$, where $G = v_j - v_i$. This local measure ignores real completion costs.

Proposition 1 (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*

$$r = 1 - b/a < \theta. \quad (3)$$

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 is (3). A displacement implies $v_i/v_j > (1 - \theta)c_j/c_i$, which is at least one under the stated condition. ∎

The proposition holds offers fixed. A strategic extension must 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, but it does not by itself supply an equilibrium. With private valuations and asymmetric bidders, first-price shading can change which buyer wins (Maskin and Riley 2000; Hafalir and Krishna 2008). Condition (3) describes displacement of given offers, while the response of those offers to a charge requires a separate bidding model. A duty whose rate follows a published signal adds a further response, choosing where and when to bid, and Section 10 reports a limited version of it.

The charge therefore selects listings by their normalised effective-offer margin, which by itself says little about its average loss relative to a cap's. Take two equally likely listings with targeted and untargeted valuations (10,000, 9,700) and (100, 80), and certainty factors 1 and 0.88. A 20% charge changes the winner only at

the first listing, at a valuation loss of 300. Excluding the targeted buyer with probability one-half at each listing gives the same expected number of displacements and a conditional average loss of 160, so aggregate comparisons have to be computed.

In the stress-test simulation they come out as follows. Removing the current 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 mostly moves units to buyers who value them more, while the cap mostly does the opposite. These local comparisons cannot be summed into a policy effect, because a buyer served at one listing might otherwise have bought later.

5. Policy-dependent benchmarks

Of the two efficiency ratios defined in Section 3.4, the contemporaneous one has a denominator that depends on who is still waiting, and so on the policy. This section shows what that does to comparisons, first exactly and then at scale.

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

$$\frac{E_1}{E_0} = \frac{V_1/V_0}{B_1/B_0}, \qquad \log\frac{E_1}{E_0} = \log\frac{V_1}{V_0} - \log\frac{B_1}{B_0}. \tag{4}$$

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.

The executable file `replication/benchmark_toy.py` constructs both paths and enumerates the feasible schedules using only Python's standard library, and its assertions check the values in Table 2.

5.2 The main market

The two efficiency ratios in Table 11 differ systematically. In a market where buyers wait, the policy decides who is waiting. A regime that serves high-value buyers quickly leaves a thinner queue, and a benchmark computed over that queue falls with it. In the main market the contemporaneous benchmark averages about 444,000 per unit without intervention and under the duty, and 463,000 under the cap. Ratios against it are not comparable across regimes.

Table 3 and Figure 1 show the consequence in the main market. On the fixed benchmark the duty loses value at every offer discount from 0% to 20%. The loss shrinks from 0.049 to 0.005 points as the discount rises, and in 3,943 of 4,000 bootstrap resamples of the eight seeds the gain never turns positive on that grid. On the moving benchmark the duty appears to add value from a discount of about 2.2% (95% interval [2.1, 2.3]), and above that point the two measures disagree on the sign of the policy's effect. At the upper company-cash discount of 5.3% and at the headline 8.8%, the moving benchmark would report a gain that the fixed benchmark does not, while at the company-cash discount of 1.5% both report a small loss.

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, and all eleven grid nodes are in `realistic_market_results.json`.

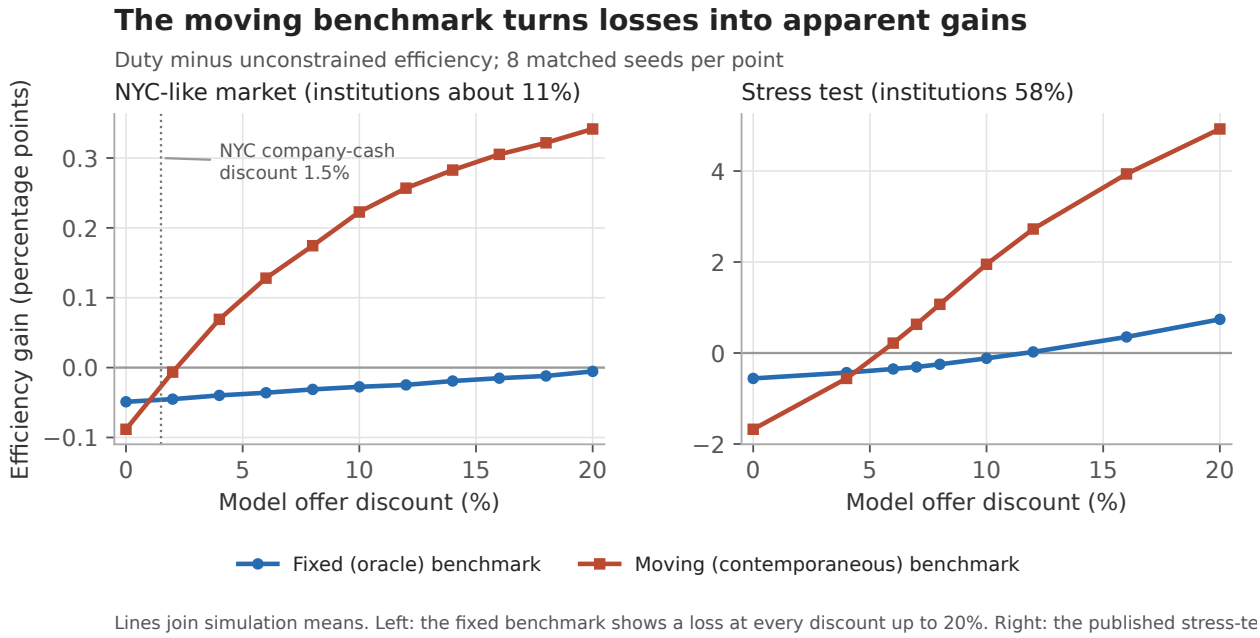


Figure 1. The duty's efficiency gain on the fixed (oracle) and moving (contemporaneous) benchmarks as the offer discount varies, in the main market (Table 3) and in the stress test (Section 5.3). The horizontal axis is the model's offer discount, which differs from an empirical price premium. Lines join simulation means.

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.

5.3 The stress test

The benchmark result is larger in the stress test (Figure 1, 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 starts adding value at an offer discount of about 11.7% (95% interval [11.5, 11.8]). On the moving benchmark it appears to add value from about 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. The moving benchmark reports gains of 0.22 to 1.95 points where the fixed one reports losses of 0.12 to 0.35. Choosing the moving benchmark more than doubles the range of discounts over which the duty seems to pay for itself.

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 4. 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. 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`); they reflect simulation noise and leave out error from the linear interpolation or the model.

Across the stress-test settings in Table 4 the break-even offer discount ranges from 11.7% [11.5, 11.8] to 15.9% [15.6, 16.1], and at the main market's participation it is not reached below 20%. Below the stress test's institutional share, the fewer the institutional purchases, the larger the discount at which the duty starts to pay for itself on the fixed benchmark.

Table 5 decomposes the stress-test comparison into its numerator and denominator, using 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

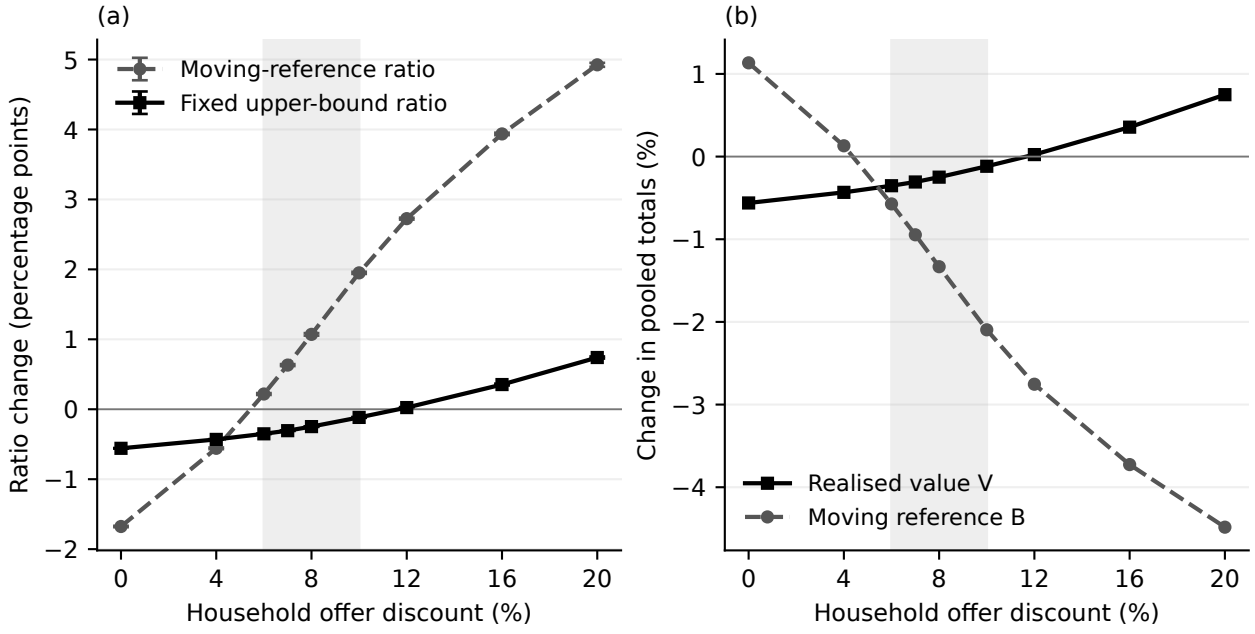
Table 5. Fresh matched-seed sweep, eight seeds per offer discount. Value and reference changes are percentages based on ratios of totals. Ratio gains are percentage-point differences of mean per-seed ratios. F is the fixed top-valuations upper bound, and B is the policy-dependent sum of current maxima. 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%. The moving ratio rises by 0.218 percentage points, while the fixed-reference ratio falls by 0.351 points. At w = 10%, the value loss is smaller, 0.117%, while the moving reference falls by 2.096%. The moving-ratio gain is 1.950 points, so the ratio records a change in the queue as well as a change in allocation value.

All eight seeds at each of the 6%, 7%, 8% and 10% nodes give these opposite signs. At 12%, both mean comparisons are positive, although one seed has a negative value change and a positive moving-ratio change. The grid contains 33 opposite-sign pairs out of 72. That count describes these selected scenarios and seeds and should not be read as a frequency estimate for housing policies or arbitrary model parameters.

At zero discount both mean comparisons are negative, and at 16% and 20% both are positive, so the sign reversal is not universal. The moving reference can make a loss appear as a gain for intermediate inputs, while agreeing in sign elsewhere. The diagnostic reports the agreement cases as well as the reversals, although the offer-discount grid does not establish how frequently either pattern occurs in other models.

Figure 2 separates the ratio changes from the changes in their components.



Changes in allocation ratios and their numerator and denominator

Figure 2. Changes in the two evaluation ratios and their components. Panel a reports matched-seed mean gains, and error bars are 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. 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 exactly. Within each pair, the recovered fixed references are identical. The largest residual in the logarithmic identity is approximately 2.31×10^{-16}. The matched inputs and recovered totals locate the reversal in the policy paths and their moving references.

Pooled and mean-per-seed moving-ratio comparisons have the same signs at all nine nodes. Their largest difference is below 0.0003 percentage points. That agreement is a numerical result for this exercise and does not make the two aggregation methods equivalent. A comparison with heterogeneous run sizes or references should specify which weighting answers the intended question.

5.4 A queue where the error runs the other way

The second model schedules unit-service jobs in discrete slots. Each job has a release slot, an inclusive deadline and a fixed value, and it can be served once. Highest-value-first (HV) selects the largest available value. Earliest-deadline-first (EDF) selects the earliest deadline, breaking ties by value and then identifier. Service takes place after arrivals and before that slot's deadline, and both policies see current jobs only, while the offline reference can see all arrivals and respects every release, deadline and one-job-per-slot constraint.

In a constructed three-slot example, job A has value 10 and is available in slots 0 through 2. B has value 9 and is available only in slot 0; C has value 8 and is available only in slot 1; D has value 1 and is available only in slot 2. HV serves A, C and D. EDF serves B, C and A, retaining A for later service. The moving reference again sums the largest available value before each choice. Table 6 compares the resulting paths.

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 6. 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 reference rises because A remains available for all three decisions under EDF. Local greedy selection gives HV a ratio of one by construction, despite its lower total value. EDF reaches the feasible optimum in this example, although as an online heuristic it need not do so in other queues.

The fixed-seed check uses six slots, two arrivals per slot, values uniform on the integers 1 through 20, and deadline extensions uniform on 0, 1 and 2 slots, truncated at slot 5. All 200 consecutive seeds from 20261006 through 20261205 are retained. EDF raises value while lowering V/B in 15 cases, 162 cases have nonzero changes with matching signs, and 23 have a tie in at least one comparison, and no case has falling value and a rising moving ratio. These counts describe this deliberately small design, with HV's moving ratio constrained to one.

A dynamic program computes the feasible optimum for every case, and a separate full enumeration checks it on all 5,832 three-job configurations with three slots, all feasible release-deadline intervals and values in {1, 2, 3}. Each emitted path is checked for availability, one-use capacity and reconstructed totals. The same generic routine rechecks all 72 housing pairs and 200 service pairs. Undefined zero-denominator and log comparisons return null, and negative or nonfinite totals are rejected. The largest service-suite log-identity residual is below 2.1×10^{-16}. The calculation is therefore portable, and Sections 9.1 and 9.2 apply the audit to published models by other authors.

5.5 What the moving ratio still describes

The moving ratio answers a local question: how winners' valuations compare with the highest valuation encountered at each decision along that policy's realised path. That is useful for describing the selection rule, and misleading when interpreted as a comparison of aggregate realised value or proximity to a common feasible optimum.

A fixed upper bound avoids this particular sign problem under matched exogenous inputs, and its level gives a conservative efficiency bound: an allocation that reaches 99.5% of F must also reach at least 99.5% of the feasible optimum under the same valuation objective. The remaining gap to F is an upper bound on avoidable loss and probably overstates it. Comparing V across matched paths is sufficient to establish the sign of the value change.

I therefore report the realised-value change as the principal aggregate outcome and use the two ratios to illustrate the measurement issue. Neither ratio supplies a complete welfare assessment, and the baseline result gives no basis for an optimal tax calculation. A change in buyer composition can have consequences outside realised valuations, including housing use and distribution, that this program does not measure.

6. Access populations

6.1 The main market

The default settings differ in strength as well as design, so the comparison calibrates each instrument to common access targets. Achieved gains on the evaluation seeds remain visible because they need not match exactly. I tuned each on four simulation seeds to lower full-run starvation by 0.25, 0.5 or 1 percentage point, and the chosen settings were evaluated on sixteen separate seeds (Table 7, Figure 3). The fair share appears twice, with the period-reset allowance of Section 3.3 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 seeds, 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 7. Instruments tuned to common full-run access targets in the main market: settings tuned on four seeds, results on sixteen held-out seeds. 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).

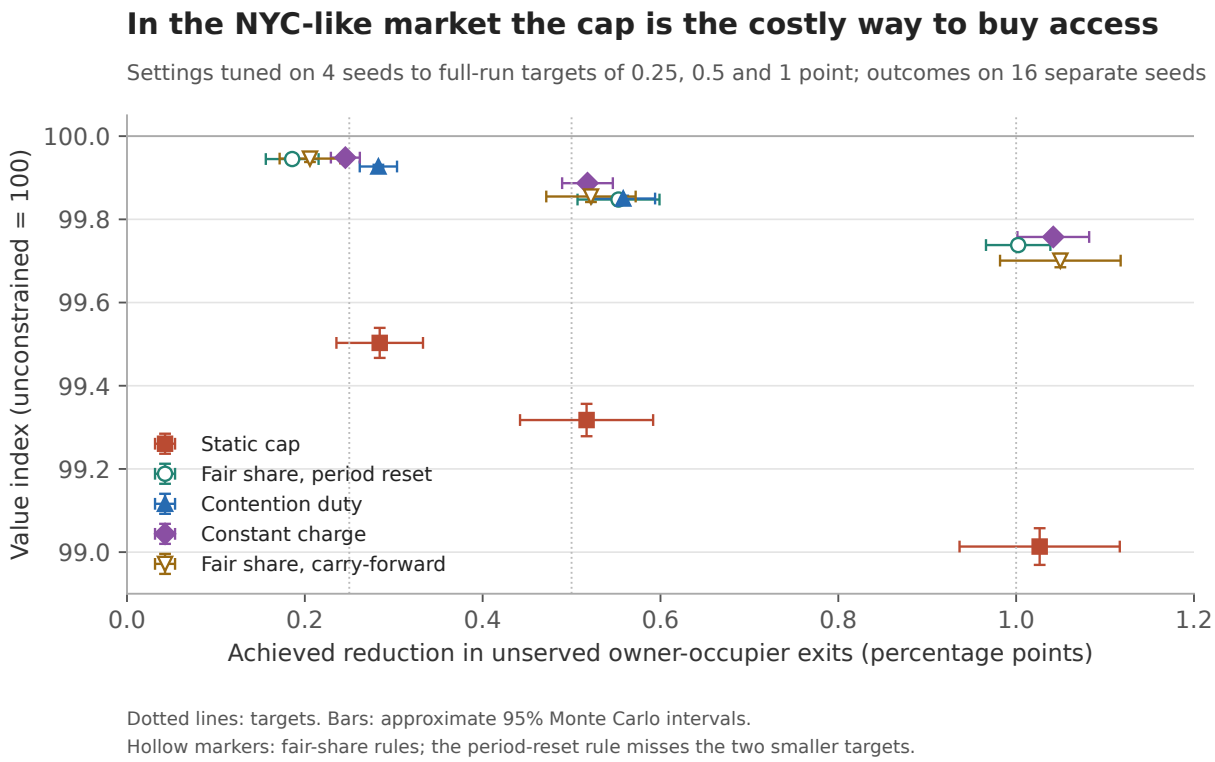


Figure 3. Settings tuned to common full-run access targets in the main market and evaluated on sixteen separate seeds. Bars are approximate 95% Monte Carlo intervals conditional on the settings. Points are not joined into a frontier.

The full-run starvation rate has a blind spot, because buyers still searching when the run ends count as neither served nor unserved, and instruments differ in how many they leave in that state. Following every owner-occupier who enters in periods 36 to 83 until they buy or leave by period 119 resolves the whole cohort. On the evaluation seeds, 15.1% of that cohort leaves unserved without intervention, against a full-run rate of 10.1%.

At the full-run settings of Table 7, cohort gains range from about seven tenths of the full-run gain, for the carry-forward fair share at the smallest target, to two and a half times it, for the cap at every target. The two measures also rank the instruments differently. At the half-point target the cap and the constant charge give the smallest full-run gains, 0.52 points each, while the cap gives the largest cohort gain, 1.29 points, against 0.75 for the constant charge. Matched to the duty on full-run access, the constant charge delivers 0.29 and 0.48 points less cohort access (paired SE 0.028 and 0.039).

Retuning to the cohort measure reverses the comparison between the two charges (Table 8). Tuned on the same four seeds to cohort gains of 0.5 and 1 point and evaluated on sixteen further seeds, the duty keeps 0.018 and 0.027 more index points than the constant charge (paired SE 0.0020 and 0.0052). Their cohort gains are within 0.03 points of each other (paired SE 0.022 and 0.044), while the constant charge delivers 0.13 and 0.23 points more full-run access (paired SE 0.011 and 0.021). Each charge keeps more value on the population it was tuned for, as in the stress test (Section 6.2), and neither result 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 8. Instruments tuned to common completed-cohort access targets in the main market: settings tuned on four seeds, results on sixteen seeds disjoint from those of Table 7.

6.2 The stress test

In the stress test the instruments were tuned in the same way to five- and ten-point full-run targets (Table 9, Figure 4). The cap is again the most expensive way to buy access, giving up about 2.5 index points for five points of access and 4.3 for ten. The fair share and the duty cost almost nothing at five points. At ten points the fair share delivers slightly more access than the duty and keeps more value, so the default ordering between them does not survive tuning.

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 9. The stress test's instruments at common full-run access targets: settings tuned on four seeds, results 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.

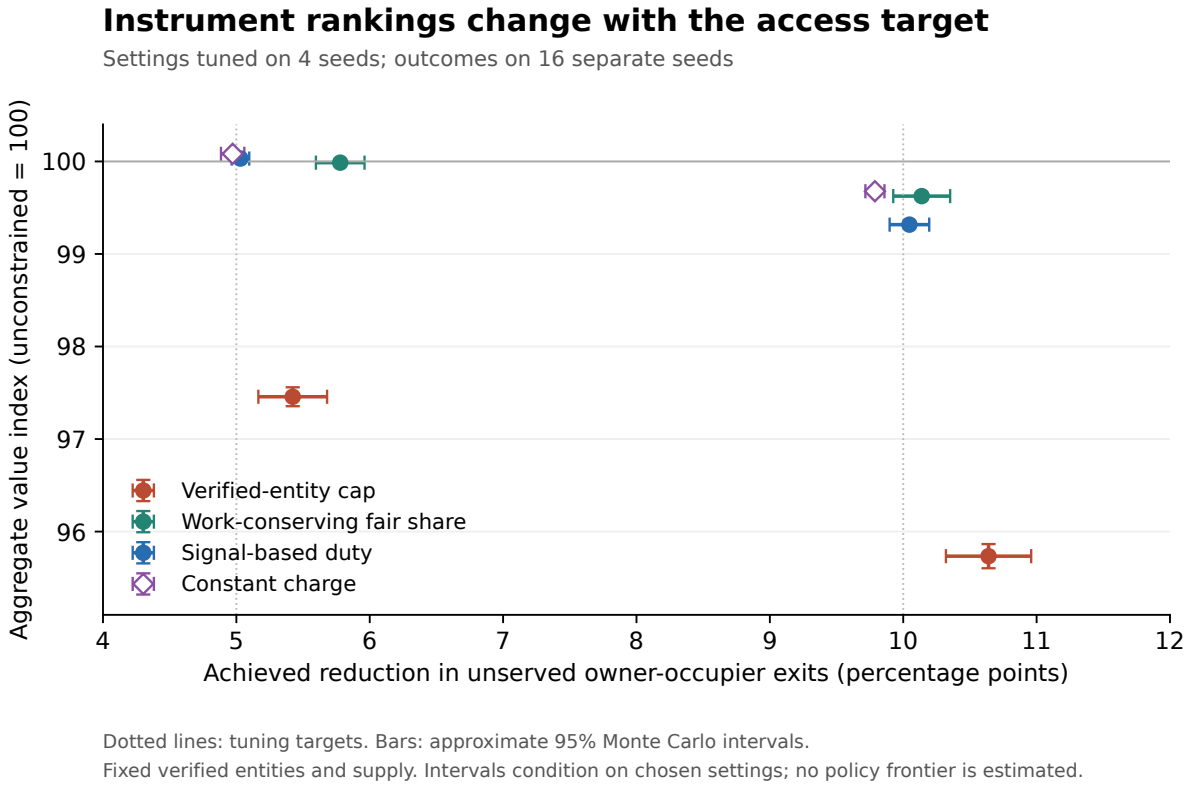


Figure 4. Stress test: settings tuned to common full-run access targets and evaluated on sixteen separate seeds, including the constant charge. Bars are approximate 95% Monte Carlo intervals conditional on the settings. Points are not joined into a frontier.

The constant charge, which ignores contention, keeps more value than the duty at both targets, by 0.05 and 0.36 index points (Table 9, Figure 4). Its access gains fall slightly short of the duty's, so the comparison is close but not 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 10. Stress test: completed-cohort access at the settings of Table 9 and the common-target exercise, without retuning. The cohort averages 4,061 entrants per replication.

The full-run measure understates every instrument's effect on this cohort, by different amounts (Table 10). The unconstrained row shows the size of the gap: 60.1% of the cohort leaves without buying, against a full-run rate of 41.9%, partly because the full-run denominator includes entrants who are still searching when the run ends. For the cap, a 5.42-point full-run gain is a 13.53-point cohort gain, 2.5 times larger, and the fair share's 5.78 becomes 7.91, 1.4 times larger. The rankings also differ. At the five-point settings the fair share leads on full-run access and the duty leads on cohort access. The two measures describe different populations, and a comparison needs to state which one it targets.

Retuning to the cohort measure changes the settings and the comparison. Tuned on the original four seeds to five- and ten-point cohort gains and evaluated on sixteen new seeds, a five-point cohort gain needs a cap of 385 acquisitions per entity instead of 336. At the ten-point cohort target the order of the two charges flips. The

duty keeps 0.032 more index points than the constant charge (paired SE 0.0033), with cohort gains within 0.03 points of each other (paired SE 0.044), while the constant charge still delivers 1.6 points more full-run access (paired SE 0.044). Each charge wins on the population it was tuned for. Near five points the value difference, 0.0012 (paired SE 0.0017), is not resolved.

7. Quota accounting

With three institutional bids per district-period and about 20 listings, a period share s lets institutions make ⌈s·L⌉ purchases in a period of L listings while an owner-occupier waits. The allowance moves in whole purchases, so the full-run gain jumps between steps; on the tuning seeds it moves from 0.21 to 0.31 points near the 0.25 target and from 0.40 to 0.55 near the 0.5 target. No period share reaches either target (asterisks in Table 7). Allowances that carry forward, one credit spent per institutional purchase while an owner-occupier waits, accumulate fractional credit and reach all three targets on the tuning seeds, with held-out gains of 0.21, 0.52 and 1.05 points. The gap is a property of the accounting and says little about fair-share quotas. I do not propose unlimited banking as a design, since it changes the timing of institutional purchases, and limits on accumulation, expiry and borrowing are untested.

The same limit appears in a third configuration, with a single institutional entity placing one bid per district-period and an unconstrained institutional share of 4.47%. Any positive period share then admits the single bid, and a zero share excludes it whenever an owner-occupier is waiting, so the quota can only overshoot a 0.5- or 1-point target or miss it entirely. Carry-forward credits reach gains of 0.48 and 0.96 points at value indices of 99.737 and 99.477. In the stress test itself, carry-forward shares of 0.5103 and 0.4351 reach gains of 5.22 and 10.18 points at value indices of 99.997 and 99.600 (`carry_share_results.json`).

8. What the instruments do

8.1 The main market at default settings

Table 11 compares the instruments at their default settings in the main market. Without intervention, full-run starvation is 9.8%, against 41.9% in the stress test.

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 11. 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.

The default instruments do little here, where the duty and the constant charge each lower full-run starvation by about 0.2 points and cost about 0.05% of realised value. The cap lowers it by 1.77 points at a 1.36% loss, so the cap costs 0.77% of realised value per percentage point of full-run access, against 0.34% at baseline in the stress test. Among entrants followed to purchase or exit, unserved rates are 14.9% without intervention, and the duty and constant charge improve access by 0.45 and 0.37 points. The constant charge does slightly better over the full run (0.24 against 0.20 points) and the duty slightly better in the cohort. Both differences are under 0.1 point, although each exceeds three paired Monte Carlo standard errors.

The certainty wedge seems to explain little of the shortfall: with the offer discount set to zero, full-run starvation remains 9.5%, against 9.8% at the company-cash gap, and the duty still improves access by 0.18 points. Institutional purchases matter far more, since excluding institutions altogether lowers full-run starvation by 6.6 points on the four tuning seeds of Section 6.1. What drives unserved exits in this market is how many units institutions buy, and the charges work mainly by limiting those purchases. Small investors retain the convention $c_{SI} = \min(1, c_{OO} + 0.07)$, making them as certain as institutions at the company-gap inputs. Setting their certainty equal to owner-occupiers' reduces the duty's full-run gain from 0.20 to 0.18 points.

8.2 The stress test at default settings

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 12. Allocation regimes in the stress test, means of 12 replications. OO and Inst. share are the shares of units bought by owner-occupiers and institutions. The value index is the sum of winners' valuations relative to the unconstrained regime under matched seeds.

The instruments deliver very different amounts of access (Table 12). The duty lowers starvation from 41.9% to 36.6% with a value index of 100.03; its gain of about 0.026 index points has a paired Monte Carlo standard error of 0.0054, so it is precisely estimated and economically negligible. Starvation falls to 24.7% under the fair share, at a 1.8% loss of value, and to 8.7% under the cap, at an 11.3% loss. Combining the fair share and the duty adds almost nothing, because the fair share lowers the contention signal and with it the duty's rate. The unconstrained market reaches 99.5% of the oracle benchmark, so the room for aggregate efficiency gains is small, and the differences that are large are distributional.

Whether a larger access gain is worth its cost depends on the welfare weights. Scoring each allocation with $W(\omega) = \omega V_{OO} + \sqrt{\omega}\, V_{SI} + V_I$, where V is the realised valuation of each class, the fair share beats the duty from $\omega \approx 1.16$ and the cap becomes the best of the three original instruments from $\omega \approx 1.66$. The cap breaks even against no intervention at $\omega \approx 1.35$. These weights are normative scenarios rather than estimates, and `replication/welfarist.py` reports the full schedule.

Two exercises in the extended preprint (Loschi 2026b) change the cap's case in particular. When institutional capital is split across legal entities at no cost, the cap's effect decays and disappears at 25 entities, so beneficial-owner aggregation would be needed to enforce it. When listings respond to seller proceeds with unit elasticity and the seller bears the whole charge, the cap's value loss rises to 17.3% and its starvation rate to 16.0%, while the duty loses 2.0 index points. The duty stays the least costly in value at every elasticity tested, but it is not costless.

Waiting times can mislead in the same way. When costless entity splitting erodes a cap, the 99th-percentile wait of owner-occupiers who complete a purchase falls, even as starvation rises. The buyers who would have waited longest leave without buying and drop out of the sample. A waiting-time statistic for completed purchases means little without the share of entrants who never complete.

8.3 Sensitivity to the New York City proxies

Table 13 varies the proxy inputs, holding instrument settings at their defaults. The constant charge equals the duty ceiling times its mean signal across district-periods. This matches an unweighted scheduled haircut, without equating charges on completed purchases or access. Sections 6.1 and 6.2 give the comparisons at common access targets.

Raising the offer discount to the upper endpoint of 5.3% or to the headline gap's 8.8% changes little. Full-run starvation without intervention rises to 10.6% and 11.1%, and the duty still lowers it by about 0.2 points.

Counting all company buyers as institutions raises their simulated share to 25.0% and full-run starvation to 18.8%, every instrument then acts more strongly, and the cap lowers starvation by 11.29 points at a 5.16% loss. In the upper-gap and headline-gap scenarios, the contemporaneous benchmark gives duty gains of 0.11 and 0.19 points, while the fixed oracle gives 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 13. NYC proxy scenarios and the stress test, twelve matched seeds. Each instrument cell is change in full-run starvation (percentage points) / realised value (percent), relative to its scenario's unconstrained market; negative starvation means improved access. The unserved columns give unconstrained full-run and completed-cohort rates. The cohort enters in periods 36 through 83 and is fully resolved by period 119. Instruments retain the baseline settings, and constant charges match the duty's mean scheduled haircut. Shares are simulated purchase shares and only approximate 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 is a model quantity, and the company shares concern a selected set of linked transactions. The two headline inputs cannot establish New York City's policy effects.

New York also charges purchase mortgages. Applying headline borrower-side rates gives averages of 1.41% of price for houses and 1.28% for condominiums before credits, exemptions and consolidation adjustments (Loschi 2026c). The model does not compare replacing that charge with taxing cash purchases.

9. Using the audit

A policy comparison should retain V and its reference for each run. The generic `compare` routine takes V and B for two paths and one fixed F and returns percentage changes, ratio gains and the log decomposition in equation (4). The input contract handles zero totals explicitly but cannot establish that policies face the same exogenous problem, which is left to the model-specific checks.

Common seeds are sufficient only when the random-number schedule remains independent of policy decisions. Full search and zero failures preserve that schedule in the audited housing program. Reduced search and failed transactions require additional care, because choices can change later draws or relistings. The deadline queue instead generates its complete job list before either policy runs, and its offline solver shares that list and the service constraints.

The reference also needs a stated interpretation. Summed current maxima describe opportunities encountered on a policy's path and can count an agent repeatedly. A global top-values reference is an upper bound that omits constraints, and an exact feasible optimum solves a different comparison problem. Keeping these quantities distinct prevents a local selection score from being read as aggregate value or an exact welfare loss.

Aggregation can introduce a further difference, because mean ratios, ratios of pooled totals and averages of percentage changes weight runs differently. Equation (4) applies within each pair, and pooled comparisons require a consistently pooled numerator and denominator. The housing audit reports both calculations and checks their observed agreement in sign.

The access outcome needs its own population contract, since successful buyers' waiting times omit unserved exits. A completed cohort resolves both kinds of outcome for its members, but may differ from other entry cohorts. Instrument settings tuned against one population should be reevaluated or retuned before being described as equal-access settings for another.

The present checks remain narrow: housing bids are fixed, institutions do not choose entry or financing, and the central audit fixes supply. Realised valuations exclude transfers and much of housing welfare. The service queue has unit-length jobs, a short horizon and no behavioural responses, and its disagreement count

is conditional on those choices. Neither exercise measures how often other simulation studies have used a misleading reference.

9.1 An outside application: PolicySpace2

To see whether the access-population problem arises outside my own model, I applied the audit to PolicySpace2 (Furtado 2022), 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), and it describes monetary aid as reaching "a larger number of households" (para. 5.9). The count resets each month. For monetary aid it is everyone on that month's register, and for property acquisition it is the houses handed over that month.

I ran the published code (commit ced77dd) at its default settings (Brasília, 1% of the population, 2010–2020, policy share $\delta = 0.2$, eligibility below the 20th percentile of permanent income) on ten seeds per policy, with a log of who was eligible, who was helped and who was below the eligibility line in each month. The logging changes no model behaviour, and the only other change is a shim that restores Python 3.6′s `random.sample` behaviour on dictionary views, which later Python versions refuse. My runs reproduce the paper's monetary-aid figure, 2,363 a month on average (2,353–2,373 across seeds), and come close to its other two (54 and 14).

Table 14 counts the same runs three ways. On the paper's monthly measure, monetary aid reaches about 170 times as many households as property acquisition in my runs (190 times in the published figures). Counted as distinct families helped at least once in ten years, the ratio is 2.3, because monetary aid pays the same families month after month (75 payments per family on average) while each house goes to a different family. The third measure follows families who first fall below the eligibility line in months 13 to 36 and asks how many are still below it 24 months later. Without any policy that share is 93.6%. Comparing runs that share a seed, monetary aid lowers it by 1.8 points (standard error 0.7), vouchers by 2.7 (0.4) and property acquisition by 6.9 (0.7), so on this measure the ranking reverses.

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 14. PolicySpace2 policies counted three ways (means over ten seeds; ranges across seeds in brackets). Brasília, 1% of the population, 2010–2020, default parameters; policies start after the first year. Newly poor: families whose permanent income first falls below the 20th percentile in months 13–36 (395 a run without policy, 673 under property acquisition). `external_audit/` *holds the instrumentation patch, the analysis script and the results.*

The third column carries the same warning as Section 6. Property acquisition changes who falls below the line: its entry cohort is 673 families a run against about 400 under the other policies, so its lower persistence partly reflects a different population. Across ten seeds the voucher and monetary-aid figures cannot be told apart (paired difference 0.8 points, standard error 0.7). My claim is narrower than a new ranking of Brazilian poverty policies. It is that the published comparison rests on a monthly flow that favours whichever policy pays the same families repeatedly, and that two other defensible populations give ratios of 2.3 and a reversal. A comparison of reach should state whether it counts payments, families or exits from poverty.

9.2 A second application: LivSim

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 example inputs that its authors generated at random from the official simulator's files. It reports geographic disparity across the 58 donor service areas in two ways: the spread of the median allocation MELD score at transplant, computed over the patients who receive

a transplant, and the spread of a mortality rate whose denominator is that year's arrivals plus the candidates waiting at its start. Persistent differences in MELD at transplant between regions were among the reasons for replacing donor-service-area allocation with acuity circles in 2020 (Nagai et al. 2022).

I ran the published code on those inputs, comparing the status quo it ships (Share 35 and Share 15 with MELD-Na) with full regional sharing, for three simulated years on the first three of its five input replications. The only changes are input paths, a one-line update for NumPy 2 and a per-patient event log. LivSim replays arrivals, disease progression and deaths from its inputs, so both policies face the same patients, and two fixed cohorts can be followed for a year: the 15,406 candidates waiting at the start and the 11,571 listed in the first year.

Spread across donor service areas	Population counted	Change under regional sharing
Median MELD at transplant, years 1, 2 and 3	Patients transplanted that year	+6%, +18%, +18% (wider in 9 of 9)
Mortality rate, year 1	Candidates at the start plus the year's arrivals	−11% (narrower in 3 of 3)
Mortality rate, years 2 and 3	The same, recounted each year	+16%, +7% (wider in 2 of 3 each)
One-year transplant chance	First-year listings (11,571)	+12% (wider in 3 of 3)
One-year death risk	First-year listings	+5% (wider in 2 of 3)
One-year transplant chance	Waiting at the start (15,406)	−3% (narrower in 2 of 3)
One-year death risk	Waiting at the start	−8% (narrower in 2 of 3)

Table 15. LivSim, full regional sharing against the status quo: mean percentage change in the standard deviation across donor service areas over three 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. `external_audit_livsim/` *holds the patch, run and analysis scripts and* `comparison.json`*.*

Regional sharing widens the spread of median MELD at transplant in every replication-year and the spread of one-year transplant chances among new listings in every replication. Among patients already waiting it narrows the spread of death risk and, less clearly, of transplant chances, and the published mortality rate, whose first-year population is more than half those patients, narrows in the first year of every replication. Whether broader sharing looks fairer therefore depends on which population is counted. The policy-dependent denominator plays little part here, since it moves by less than 0.2% nationally. The served population does change: in the first replication, 909 of the 5,106 new listings transplanted within a year under the status quo are not transplanted within a year under regional sharing, and 922 others are, so the median at transplant describes partly different patients under the two rules.

The inputs are synthetic, the horizon is three years and the replications are few, so these figures say nothing about actual US policy. The point is narrower: a published allocation model that reports disparity over transplanted patients and over a flow of candidates can give a different verdict from fixed cohorts of the same patients, and fixed cohorts entering at different times can disagree with each other.

10. What the model leaves out

Bidding is not strategic. Valuations are private and fixed, and buyers bid them truthfully, so the simulation tests what the instruments do to allocations and is silent on how buyers would respond to them. In the stress test, a limited exercise lets institutions reallocate a fixed acquisition budget across districts in proportion to $(1 - \tau_{max} s_d)^{\lambda}$. For 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%, so even the sign of its small aggregate effect is not stable once buyers move. Bid shading, entry and manipulation of the contention signal are not modelled, and I regard this as the largest gap.

Outside the two proxy inputs, the market's parameters are illustrative. In the stress test, patience, the number of districts, valuation dispersion, the contention window and search friction leave the full ordering unchanged, but arrival and listing rates change it. With owner-occupier arrivals doubled to 34 per district-period, which lowers the unconstrained institutional share to 38%, the fair share scores 100.20 and overtakes the duty at 100.09, while the duty still lowers starvation more (4.0 against 2.9 points). The cap is the most costly instrument at every point tested in either market, at 97.3 in that case. Its place relative to the contention-linked instruments

holds across the reported exercises, although the ordering within that pair does not. These sensitivity exercises have not been repeated in the main market.

The value index is narrower than welfare, because it counts realised willingness to pay and leaves out renters, the use of any revenue, administrative costs and differences in the marginal utility of income. The supply exercise lets listings respond to seller proceeds, but it has no construction sector, rents or entry. The horizon is finite, so the full-run starvation rate is not a lifetime purchase probability, and entity splitting is costless, so the cap's decay is an upper bound on avoidance without enforcement.

The extended preprint (Loschi 2026b) reports the remaining sensitivity exercises, and the main points are these. Letting 10% of accepted financed sales fail leaves the value indices and ranking similar and the fixed-benchmark break-even almost unchanged (11.8% against 11.7%), but it raises failed transactions under the duty by 21.6% and more than doubles them under the cap. Reduced search, with inspection probabilities of 0.5 and 0.25, preserves the ranking. Sweeping institutional buyers' mean valuation from 340,000 to 410,000 moves the duty's index by at most 0.06 points and raises the cap's cost from 8.6% to 16.0%. The baseline institutional share of 58% is far above observed markets. With the share lowered to 22.5% the duty and fair share barely act, reducing starvation by 0.7 and 0.1 points, while the cap still costs the most value at every share tested. The rate sweep finds the 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% the ceilings 0.11, 0.22 and 0.44 all lose value.

11. Implementation

Each instrument needs information the model takes for granted.

The duty needs an exit registry. Its signal counts owner-occupiers who leave without buying. Recording that requires a register that observes entry into the market and distinguishes purchase, withdrawal and a move to another market, which routine transaction records do not show. The signal can be pushed in either direction. Institutions could depress it by withdrawing bids during the measurement window. Anyone who gains from a higher duty could inflate it with entries or withdrawals staged to register as unserved exits. A workable version would count only verified participation, such as a mortgage pre-approval or a deposit, keep the records with a party that has no stake in the rate, and limit how far the rate can move in one period. Björkegren, Blumenstock and Knight (2026) show how to design decision rules that stay accurate when the people they score change their behaviour to move the score, and a contention signal faces the same problem. The constant charge needs none of this, which is the practical side of Section 6.1: the duty's value advantage is small, depends on the access population, and has to be weighed against the cost of measuring contention.

The cap needs ownership aggregation. A per-entity limit binds a legal person, and in the stress test a single owner who can form entities at no cost defeats it with 25 of them. An enforceable cap would need a definition of common control, disclosure of beneficial owners, verification and sanctions. Federal beneficial-ownership reporting has exempted domestic companies since 2025, so a state rule would have to supply that information itself. A charge on each purchase avoids the gain from multiplying entities, although it still needs an enforceable definition of who pays.

The fair share's release rarely binds. The quota is work-conserving, which means it lifts when no owner-occupier is waiting. In the stress test someone is always waiting, so the simulation tests the quota without testing its release, and in the main market the default quota does not bind at all. Its behaviour in slack markets is untested.

Classification decides who is reached. Payment method, institutional status and owner-occupation are different categories. Some owner-occupiers pay cash and some investors borrow. Whereas the model charges institutions, the 2026 New York proposal would have charged cash purchases of $1 million or more, and a payment-method test can be avoided with a short-term loan repaid after closing. A charge on all non-owner-occupier purchases would also reach small investors and would need its own simulation.

Coverage and geography matter. A tax reaches only what its statute defines. New York's recording tax applies to mortgages on real property, so co-operative share loans fall outside it, while the mansion tax expressly includes co-operative units. Buyers can also move purchases to places the policy does not cover, and the district exercise in Section 10 captures only a limited version of that response.

12. Conclusion

When a policy changes who is still waiting, a benchmark built from the best agent present at each decision moves with the policy, and the ratio against it can rise while realised value falls, or fall while it rises. In the housing model the moving benchmark turns a duty that loses value at every tested discount into one that appears to pay for itself above a 2.2% discount. In a deadline queue it penalises the policy that reaches the feasible optimum. The access population matters in the same way. Counting exits over a finite run leaves buyers still searching out of the reckoning, and following a fixed cohort changes measured gains and reorders the instruments. Quota comparisons depend on whether allowances reset each period. None of these problems is specific to housing. In PolicySpace2, a published model of Brazilian poverty policies, counting families instead of monthly payments shrinks monetary aid's monthly advantage over house purchases from about 170 to 2.3 times, and following policy-specific cohorts of newly poor families reverses the ranking. In LivSim, a published liver-allocation simulator, full regional sharing widens geographic differences among transplanted patients and new listings but narrows them among patients already waiting.

The remedy, in my view, lies in reporting. A comparison should give realised value directly, report the denominator of any ratio alongside the ratio, use a fixed reference whose interpretation is stated (an exact feasible optimum or a labelled upper bound), define the access population before tuning, and state how quota allowances accumulate. The routine supplied with the paper takes saved totals from any simulation and flags reversals. This extends the transparency asked of policy-facing models, documented code and explicit assumptions (Squazzoni et al. 2020), to the reference populations behind each reported outcome.

On the substance, the certainty wedge does little at the New York proxy values. Setting the cash advantage to zero changes unserved owner-occupier exits by three-tenths of a point, while excluding institutions changes them by 6.6 points, so access in this market depends on how many homes institutions buy. The per-entity cap buys access at four to nine times the value cost of the charges and binds only legal entities. These comparisons hold bids fixed, omit supply, rents and construction, and rest on calibration inputs that do not themselves identify a certainty premium.

Declarations

Declaration of generative AI use. The author 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. The computational reviews reported with the replication packages were carried out by AI agents rather than human referees, and a version-by-version record of these uses is in the packages. The author is responsible for the content.

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

Funding. None.

Data availability. This paper merges two earlier preprints, *The Execution-Certainty Wedge* (version 6.5) and *Auditing policy dependent benchmarks in allocation simulations* (version 1.0.1), and their replication packages reproduce every number here, with the housing simulation at https://doi.org/10.5281/zenodo.22164073 and the benchmark audit, deadline queue and portable comparison routine at https://doi.org/10.5281/zenodo.23257315. New York City inputs come from the companion study (Loschi 2026a) and linked dataset (Loschi 2026d). The LivSim audit of Section 9.2 (instrumentation patch, run and analysis scripts and results) is deposited with version 7.1 of this paper.

References

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.

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.

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 3.11) [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 3.4) [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.

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 preprint (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, 7 and 8 and the first row of Table 4) under the plan in `evidence/realistic_market_plan.json`; `realistic_tables.py` builds those tables, `build_figures_v64.py` draws Figures 1 and 3, 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 4 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` (Section 4.2), `replication/welfarist.py` (Section 8.2), `replication/breakeven_schedule.py` (Table 4), `equal_access.py` (common targets), `constant_charge.py` (Table 9), `check_completed_cohorts.py` (Table 10), `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 Section 5.3 and Table 4.

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; the cohort retuning is evaluated on sixteen further seeds disjoint from both. Within a seed every policy sees identical arrivals, valuations and listings, and the oracle denominators match; the reduced-search runs draw search from a separate generator so that this also holds there.

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. All matched the shipped outputs to floating-point tolerance. The New York City estimates in Table 1 come from the companion study (Loschi 2026a). For Tables 11 and 13 run `replication/nyc_anchored.py`, then `replication/nyc_cohort_check.py`. The original model is unchanged; the cohort diagnostic reproduces every original summary exactly. `replication/verify_nyc_inputs.py` recomputes the share proxies from the separately deposited dataset; instructions and input hashes are in the README and `nyc_inputs.json`.

A.2 Model description following the ODD protocol

The model is intended to expose how repeated allocation changes the queue used to evaluate a policy. It is not fitted to the distribution of household search histories in a 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. Institutional instances also store an entity identifier. District state comprises 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. All current civilian buyers observe all local listings under full search. Valuations and patience are fixed at entry. The score does not vary with elapsed waiting time. The model's central interaction is competition among current offers and the removal or expiry of bidders. Institutional entities generate capacity repeatedly but have no endogenous balance sheet or intertemporal objective.

Scheduling and stochastic inputs

The schedule is arrivals, civilian expiry, institutional instances, listings, lagged signal calculation, sequential allocations and exit-history update. An instance can win at most once. Institutional instances disappear at the period boundary. Civilian expiry occurs when current period minus entry period reaches the specified patience, before the period's listings are processed. These scheduling choices are part of the feasible problem, not implementation details that can be ignored by a dynamic optimum.

Poisson arrival and listing draws and lognormal valuation draws use the main random generator. At full search the number of these draws is independent of winners. A separate generator is reserved for transaction failures. The full-search, zero-failure audit does not activate it. The inherited reduced-search diagnostic separates search draws from exogenous inputs; those extensions are outside the reported audit. Model parameter and source hashes accompany the new outputs.

The quota checks institutional wins against quota share times current listings before choosing each winner. Because wins are integers, a period-reset rule can permit a purchase before the quota threshold is crossed. It is not a smooth fractional allocation mechanism. The cap accumulates acquisitions over all periods. These rules help explain why entry cohort and accounting period matter to comparisons.

A.3 Benchmark audit commands

Run `python replication/benchmark_toy.py` for Table 2. Run `python replication/benchmark_audit.py` for Table 5 and the seed-pair file. Run `python replication/cohort_audit.py` for the stress-test cohort check of Table 10. Run `python build_benchmark_figures.py` to draw Figure 2 from the saved outputs. The benchmark-audit package contains the code, run-level results and reproduction instructions.

The housing sweep contains 144 runs; the cohort audit contains 80. Every inherited field selected for those checks reproduced exactly in Python 3.14.7 with NumPy 2.5.3. The cohort checks additionally verified valuation hashes, fresh listings and fixed-reference equality. This is a targeted fresh reproduction, not a rerun of every historical sensitivity or proof of empirical accuracy.

A.4 Service queue and portable comparison

Run `python replication/deadline_queue.py` to regenerate Table 6, all 200 service cases and the independent enumeration check. Run `python replication/portable_audit.py` to apply `compare_benchmarks.py` to the saved housing and service totals. The service queue and portable comparison use Python's standard library; no external data or network access is required. The service example is separately implemented and imports no housing code.

The offline dynamic program maximises value over (slot, served-job set) states. It permits an idle slot or any unused job whose inclusive release-deadline interval contains the current slot. Each transition advances one slot. This recurrence enumerates every feasible next action. Its exponential served-set state space is suitable for the small verification design, not a proposed large-scale scheduler. The independent verifier instead enumerates complete slot assignments, including idling, then rejects 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; all cases appear 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. EDF's success in Table 6 therefore establishes no general superiority over HV. Its policy is included to change queue composition and test the diagnostic.