BUILD-TO-ORDER:
ENDOGENOUS SUPPLY IN CENTRALIZED MECHANISMS*

Andrew Ferdowsian†
Kwok Hao Lee‡
Luther Yap§

This version: November 2, 2024

How should the supply of public housing be optimally designed? Although commonly used queuing mechanisms treat the supply of goods as exogenous, designers often control the inflow of goods in practice. We study a dynamic matching model where the designer minimizes a convex combination of mismatch count and vacancies, based on the Singaporean housing allocation process, Build-To-Order. With endogenous supply, the optimal mechanism overproduces underdemanded housing relative to the proportional benchmark, and competition over housing improves match quality. Batching applications artificially generates competition and is optimal when the planner places a high weight on match quality. Following our dissemination, the Singaporean government increased batching.

Keywords: Dynamic Matching, Market Design, Queueing, Waiting Lists.

JEL: C72, C73, D47, D61, D82.

*We thank Adam Kapor, Alan Wei, Alessandro Lizzeri, Can Urgun, Daniel McGee, Erez Yoeli, Joseph Ruggiero, Kate Ho, Moshe Hoffman, Neil Thakral, Nick Arnosti, Pietro Ortoleva, Sylvain Chassang, Thomas Gresik, Tim Wang and audiences at Princeton University, Stony Brook University, National University of Singapore, University of Texas: Dallas, and the EEA for their helpful suggestions and feedback. We are especially grateful to Leeat Yariv for her continued advice and support. Last, we acknowledge the financial support of the Dietrich Economic Theory Center and the Princeton Economics Department.
†Corresponding Author: 3060 Jenkins Nanovic Hall Notre Dame, IN 46556, aferdows@nd.edu, (650) 575-3855
‡Department of Strategy and Policy, National University of Singapore Business School, kwokhao@nus.edu.sg
§Department of Economics, Princeton University, lyap@princeton.edu

1 Motivation

Because land and funding are limited, public housing is rationed. Much attention has been placed on how to best allocate public housing, but an equally important question is *what type* of housing to build.[1] These decisions are not made arbitrarily: governments can infer which developments are more desirable using past realizations of demand. When deciding what apartments to build and how to allocate them, a public housing authority must contend with two objectives: minimizing vacancies and matching households to the apartments they desire. In this setting, how can a government better build and allocate public housing? To answer this question, we form a dynamic queuing model specialized to the Singaporean public housing system, which houses 80% of the resident population.[2] We characterize the government's optimal strategy when supply is endogenous. We show that building underdemanded apartments is crucial to ensure incoming households report their types truthfully. Furthermore, higher demand can improve allocative efficiency when supply is endogenous. In contrast, when supply is exogenous, allocative efficiency is reduced when demand increases. A key takeaway of our results is that batching multiple applications together is highly desirable in terms of reducing the aggregate uncertainty households face, thereby increasing their willingness to apply sincerely, improving the overall allocation. After the initial dissemination of our results, in 2024, the Singaporean government increased the level of batching, changing from four cycles per year to three.[3]

The first contribution of this paper is to develop a model that explores the link between revealed demand and hidden preferences, when the demand responds to the quantity of supply. Due to the wait times inherent to the housing process, households may choose to take a less desirable apartment today in lieu of their preferred option in the future. We examine the circumstances under which *manipulated* applications are prevalent—applications where households apply to a queue that does not match their type. Under exogenous supply, manipulated applications occur when one queue "overflows" and households of that type apply for the underdemanded apartment. In our solution to the government's dynamic problem under endogenous supply, the optimal mechanism is limited by the reverse issue; the designer's binding constraint comes from households who desire underdemanded apartments. The government prefers to only build apartments in high demand; but if the government does so, it cannot motivate households to apply sincerely. Hence, the government trades off exploiting its knowledge about the current stock of households against learning about

[1] See, e.g., Waldinger (2021), Van Dijk (2019), and Armentano et al. (2024) for empirical work pertaining to public housing allocation in Cambridge, Amsterdam, and Uruguay; and Arnosti and Shi (2020) for a theoretical treatment.

[2] Singapore is an important case study because of the size of its public housing market. Over 80% of Singaporeans live in government-built housing. In 2019, there were over 15,000 apartments transacted; each with a sticker price of at least US$200,000. These figures imply that at least US$3 billion were transacted in apartment value of government-built housing, suggesting large potential gains to improvements in efficiency.

[3] See https://shorturl.at/yeg3h for details.

the preferences of incoming households.

To further motivate why the optimal mechanism prescribes supplying less-demanded apartments, consider an environment where households care both about match quality and about match timing. If a household knows only apartments in high demand will be built, it will be tempted to misreport if it prefers an apartment type that is rarely demanded and so built infrequently. Were that household to enter the queue for its desired apartment, it would need to wait for at least one allocation period before having the chance to receive an apartment. Then, unless the cost of waiting is small, the household prefers to enter the queue for apartments that are more likely to be built.[4]

In Section 3, we propose a dynamic model of public housing allocation in which supply can be adjusted over time. At the beginning of each period, the government observes previous household applications, then decides how to allocate new housing across apartment types, where types reflect both apartment size and location. Simultaneously, a household arrives with a private type, corresponding to the apartment type that the household prefers. Newly arrived households select one type of apartment to apply for.[5] When an apartment type has more applicants than available apartments, the available apartments of that type are randomly allocated. If a household is allocated an apartment, both exit the market. Otherwise, the household pays a waiting cost and applies again in the next period. The government aims to minimize a convex combination of two types of inefficiency: allocation and unassignment. Allocation inefficiency captures the government's desire to assign households to apartments of their type. Unassignment inefficiency reflects the government's aim to minimize the number of households that remain unmatched in a given period.[6]

We focus our attention on the Singaporean mechanism, the Build-To-Order scheme (BTO), which we describe in Section 2. Our model is carefully tailored to fit the Singaporean setting, which is of interest to policymakers and academics alike because the market for public housing is large, and the associated policy problem is highly nontrivial. We show in Section 3.1 that the optimal unconstrained mechanism is a first-in-first-out mechanism, and detail why the government stopped using a first-in-first-out mechanism because of historical and other policy concerns.[7] Accordingly,

[4]Importantly, in Singapore, most applicants' alternatives to receiving an apartment are living with family or renting. While non-trivial, the cost of being unmatched for an additional period can be said to be far less than it would be in, e.g., American public housing, where failures of provision may entail homelessness. We show that, whenever the cost of waiting is sufficiently high, it is optimal for the designer to implement a pooling equilibrium: all applicants receive the same housing. This analysis corroborates results in Arnosti and Shi (2019): many public housing programs ignore household preferences.

[5]Unlike the standard literature on mechanism design where the designer observes types within a truthful mechanism, here, in line with the real-world BTO mechanism, the government can only observe applications to queues.

[6]Allocation inefficiency is standard in the literature, and unassignment generates vacancies which are expensive to maintain. We elaborate on the government's historical incentive to minimize unassignment in Section 2.

[7]In practice, a household can be given priority in a few instances, many of which are no longer in place as of October 2024. For example, prior to October 2024, if a household has previously been rejected twice, then applies for a flat in a "non-mature" neighborhood (a neighborhood with more land available for development), they are more likely to receive an early queue number.

our model differs from standard mechanism design. We restrict the space of policies, only allowing the government to utilize lottery mechanisms, focusing our attention instead on the optimal supply.

Our second contribution is to add to a growing literature on thickness in dynamic markets through characterizing when batching multiple application cycles is optimal. In Section 5 we examine the effect of competition, both when supply is endogenous and when it is exogenous.[8] We show that when the government can control the supply of housing, responsive mechanisms are more effective in oversubscribed environments.[9] While oversubscription increases both market thickness and competition, the increase in thickness enables the government to manipulate the expected wait times between different queues. This policy lever increases the willingness of households to apply sincerely, improving allocative efficiency. In contrast, when the apartment supply is exogenous, the overwhelming impact of competition is to increase expected wait times. Thus, competition exacerbates the inefficiencies of the exogenous setting, striking a contrast with improved outcomes under endogenous supply. We utilize this insight to show that, in the optimal mechanism, thickness is artificially generated in the market by batching applications.

While we focus our attention on the BTO mechanism in this paper, we note there exist several other instances wherein a centralized planner must choose the supply of a good with incomplete preference information. For instance, consider class schedules. Electives are often substitutable for students, and faculty may be reallocated to address demand spikes. Thus, schools face a year-over-year decision regarding which electives to offer and the number of students to cap each course at. The trade-offs in class scheduling are similar to those considered in the BTO mechanism. Budish and Cantillon (2012) explore class selection at the Harvard Business School. They show that students strategically report their preferences in the course allocation mechanism.

Another example can be found in the area of food donations. Prendergast (2016) and Altmann (2024) examine an innovation in the allocation of food donations by Feeding America, the second-largest American charity by revenue. In 2005, a market was established with "Monopoly money" to improve the distribution of food among food banks across the US. Prior to the introduction of the market, food banks had a fixed food need by weight, and received "take it or leave it" offers commensurate with their need levels. These offers made no allowance for the type of food, whether it be produce or pasta. However, these food types featured real differences in storage needs and often individual food banks received food donations outside the Feeding America system. One primary goal of the new market was to ensure that food banks could bid on the types of food they actually wanted when they wanted. Equally important, upon observing the relative pricing of foods, Feeding America was able to then structure its fundraising requests to increase the quantity

[8] We refer to environments where households anticipate low odds of success in the queuing lottery as competitive. An environment is *oversubscribed* if the average ratio of households to available apartments is high.

[9] A responsive mechanism is one where different preferences of incoming households lead to different allocations, i.e., the government responds to preferences.

of highly demanded food types.

One key difference between the Singaporean public housing system and Feeding America is that the Singaporean government does not allocate apartments by an applicant's willingness to pay, because they believe that housing assistance should not necessarily be disbursed to the highest bidder. In this paper, we will take it for granted that a market equilibrium will not achieve the government objective. Instead, we focus on finding the mechanism that minimizes inefficiency subject to the government's outside constraints.

Related Literature

A large literature, focusing on the optimal allocation of scarce resources, has improved the design of markets ranging from kidney exchange to school choice. Standard models in this literature have centered primarily on markets where the supply of the scarce resource is exogenous. For instance, a designer of a kidney allocation scheme cannot choose the blood types of the organs entering the system. Importantly, under traditional allocation mechanisms, the supply remains fixed and independent of agent preferences. However, in many markets, a centralized agency may control both the incoming supply of goods and the allocation of goods to agents. In this paper, the government can control the type of apartments built and the allocation of apartments.

This paper speaks to the theoretical literature in matching, much of which stems from studying the problem of optimal student assignment to schools (Abdulkadiroğlu and Sönmez 2003).[10] Within this literature, our paper is most closely related to papers on optimal dynamic matching. In this context, agents are "born" in sequence and face a trade-off between taking their best option at birth and waiting for a better match (Baccara, Lee, and Yariv 2020). Akbarpour, Li, and Gharan (2020) shows that a mechanism designer may wish to focus on increasing market thickness, over matching agents myopically. We show that this insight carries over even when the good is produced endogenously.

In particular, we offer a new take on the queuing literature. The vast majority of this literature focuses on the allocation of a fixed supply of goods, such as organs. Recent work in this literature include Shi (2022), which examines the optimal priority system to allocate agents to objects; and Agarwal et al. (2019), which develops a new organ allocation mechanism. Thakral (2019) shows that there may never exist an ex-post efficient mechanism when supply is stochastic. In all of these papers, the supply remains exogenous; the mechanism designer cannot control or alter the inflow of goods. In this paper, we consider the impact of relaxing the assumption of exogenous supply and allow the designer to freely control the types of goods that arrive.

Closely related to our work, Leshno (2022) characterizes the optimal mechanism when goods

[10] See Abdulkadiroğlu and Sönmez (2013) for a survey.

arrive according to an exogenous process. We study a different class of markets in which the designer can not only control the allocation procedure, but also the arrival rate of each good. We show that several of his insights are due to this exogenous process, giving rise to different policy prescriptions for social planners with endogenous supply. For instance, while increased household demand decreases allocative efficiency in the exogenous supply setting, it actually improves the government's ability to manipulate wait times in the endogenous setting. We explore these phenomena in Section 5.

Several other recent papers consider dynamic allocation problems with private information when monetary transfers are not permissible. Verdier and Reeling (2022) examine the allocation of bear hunting licenses and show that a dynamic mechanism which repeatedly allocates licenses to the same individuals can improve matching over a static mechanism. In our context, individuals only require one apartment, making this approach impractical. Guo and Hörner (2020) consider repeated good allocation to a single agent whose valuation fluctuates over time. Galichon and Hsieh (2018) shows that as long as money burning is permissible, stability can be achieved in many settings with private information.

Last, in our companion work, Lee et al. (2024), we utilize the same dataset to construct a dynamic choice model over housing lotteries and estimate it. Using this model, we are able to answer a separate question: what is the impact of increasing the *total* supply of housing on wait times, vacancies, and prices on the aftermarket for government housing? We find that simply increasing the housing supply can fail to reduce wait times, because the resulting demand response eclipses the supply increase. Indeed, improving the allocation procedure complements increasing supply. Specifically, when combined with a strategyproof mechanism, increasing supply can keep wait times low and reduce upward pricing pressure on the aftermarket.

2 Policy Background

2.1 The Build To Order Scheme

Over 80% of resident households in Singapore live in government housing, which makes up 80% of the housing stock in Singapore. These apartments, numbering over 1 million, are administered and maintained by the Housing and Development Board (HDB). In Singapore, since 2001, all government apartments have been first introduced into the housing stock via the BTO scheme. Under BTO, every four months, the government announces several new developments to be built. Households may apply to at most one development. Apartments in oversubscribed developments are rationed by lottery.

The BTO scheme superseded the previous Registration for Flats System (RFS). Under RFS,

homebuyers first chose the broad geographical area in which they wanted to live, then were informed of the cost and exact location when their queue number was called. Not only did buyers not know when they could move in to their apartment, but they only had to pay the down payment for any home loan when their apartment was completed; if their apartment was finished early, some of these buyers could not raise enough funds for their down payment. The Asian Financial Crisis in the late 1990s exacerbated this issue. The government suddenly found itself with a surplus of vacant housing, and incurred heavy maintenance fees. Soon after the crisis, the government switched from RFS to BTO. The key difference between the two mechanisms is that, under RFS, households need only apply once to a queue that serves earlier applicants first; whereas under BTO, households must reapply every time the lottery is run. BTO ensured that current applications were an accurate representation of current demand. This history motivates our modelling restriction that precludes the government from using allocation mechanisms that reward seniority.[11]

Introduced in April 2001, the BTO exercise allows potential homeowners to ballot for their preferred neighborhood and apartment size.[12] Each such ballot, termed a "booking", is secured by a down payment due on application. After the application phase, the HDB assigns each applicant a queue number, indicating the number of other applicants ahead of her in the queue.[13] If, when her turn arrives, she does not like any of her choices (or has none), she may withdraw her application, after which she may participate in a future cycle. Upon withdrawing, her position in the queue is lost and not preserved for future applications; they incur severe penalties to their priority in future applications. Given that apartments of a similar size have similar floor plans, we assume in our model that if a household applies for an apartment and is successful, they always accept.

After all applicants have either selected an apartment or withdrawn their application, the HDB begins building the apartments if 70% of all properties in a development have been allocated. In practice, BTO apartments of all sizes are oversubscribed, most by at least 2-3 times. To our knowledge, all BTO exercises have successfully reached the construction stage. These apartments

[11] Currently, the HDB allocates its apartments through a complex system of allocation processes with transfers, from which we abstract. This is because we want to keep our model tractable, and moreover, BTO is the scheme through which most government apartments in Singapore are initially allocated. For instance, in Financial Year 2013/2014, there were 86,298 BTO residential units under construction. Under the next largest comparable scheme "Design, Build and Sell", a scheme targeting households with higher incomes and with more extensive private developer involvement, only 3,893 units were under construction (Housing and Development Board 2014). By Financial Year 2018/2019, all residential units under construction were BTO apartments (Housing and Development Board 2019). See https://www.hdb.gov.sg/cs/infoweb/residential/buying-a-flat/buying-procedure-for-new-flats/modes-of-sale for more details.

[12] For more on the historical context for the BTO, see the government archives: https://www.nlb.gov.sg/main/article-detail?cmsuuid=d33acabb-a341-460c-8fde-99cf0a9270f4.

[13] To enforce social mixing, there are ethnic quotas for each housing development. I.e., there is a cap on the number of Chinese, Malay, and Indian households permitted in each government apartment building. This aspect of the housing system has been extensively studied in previous work (see Wong [2013, 2014]), so we will abstract from these concerns.

are typically ready for homebuyers to move in within 3 years of the corresponding BTO exercise. To prevent immediate arbitrage, apartments may not be sold on the secondary market before 5 years of occupation after the initial move-in.[14]

These apartments are often oversubscribed because they are sold at highly subsidized prices, resulting in selection into BTO from the private market. Buyers with higher incomes default to the private market because they may be ineligible to apply for BTO apartments, or are eligible but receive much lower subsidies than the median household does. Thus, private market prices reflect the preferences of a selected sample of Singaporeans, which do not necessarily represent those of the households the BTO scheme serves.[15]

2.2 Responsive Apartment Supply

Under BTO, each booking made by a household is a signal of housing demand. This allows the HDB to adjust future housing supply to meet expected demand. In Figures 1 and 2 we compare non-mature and mature neighborhood apartment allocations by type. The larger level of volatility in mature neighborhood apartment supply reflects the government's willingness to adjust the supply of housing when it has access to demand data.

Figure 1 displays the relative quantities of housing types supplied in non-mature neighborhoods. We observe that while the total quantity supplied fluctuates over time, the relative proportions of each type do not. That is, in neighborhoods where the government has less information, it opts to avoid adjusting the supply of housing.

In contrast, Figure 2 shows that housing in mature neighborhoods changes sharply over time. While 3- and 5-room apartments initially are built at equal rates, their relative ratios change dramatically. Comparing Figure 2 and Figure 1, there is more variation in proportions of various types built over time in mature neighborhoods than in non-mature neighborhoods. In mature neighborhoods the government has more information about prior demand realizations, while in non-mature neighborhoods the government knows little beyond the overall population-level preferences. We believe that this difference is due to the government accounting for neighborhood-level demand, and using it to determine future offerings within that neighborhood.

[14] From October 2021, apartments in well-located projects (classified as "Plus" or "Prime") may not be resold before 10 years of occupation.

[15] Despite the hefty subsidies given to successful BTO applicants, a government official we spoke with noted that only about 10% of BTO buyers sell their apartments within 5 to 10 years of purchase. While there is an incentive for arbitrage, this opportunity may only be available to households that are not capital constrained.

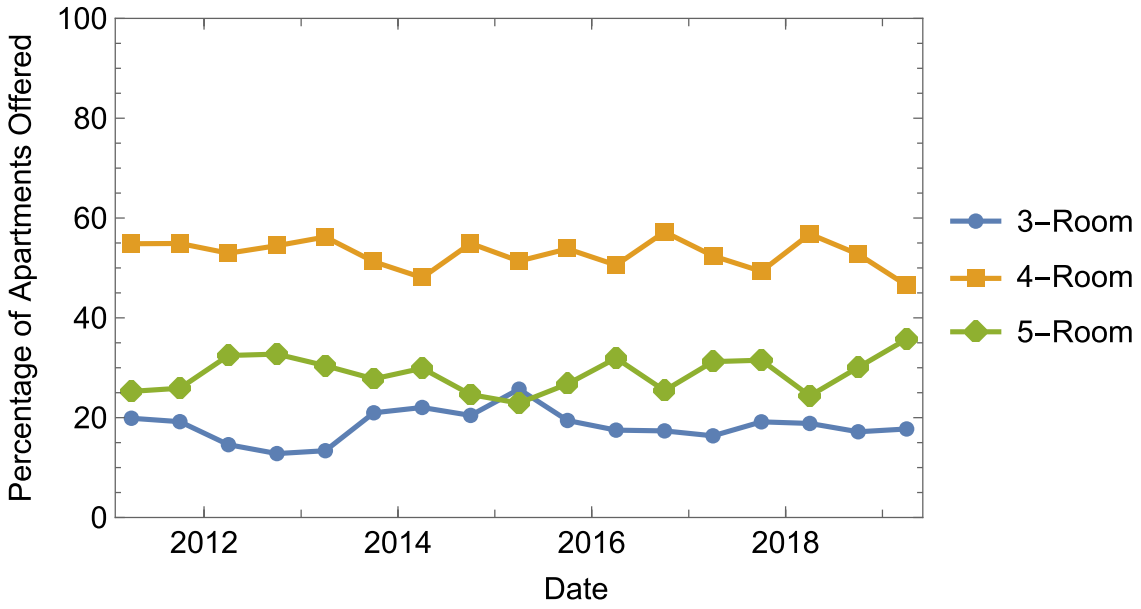

Figure 1: *Percentage of apartments, per half year by type, built in non-mature neighborhoods.*
Notes*: Non-mature neighborhoods have more land for further housing and infrastructure development, relative to mature neighborhoods.*

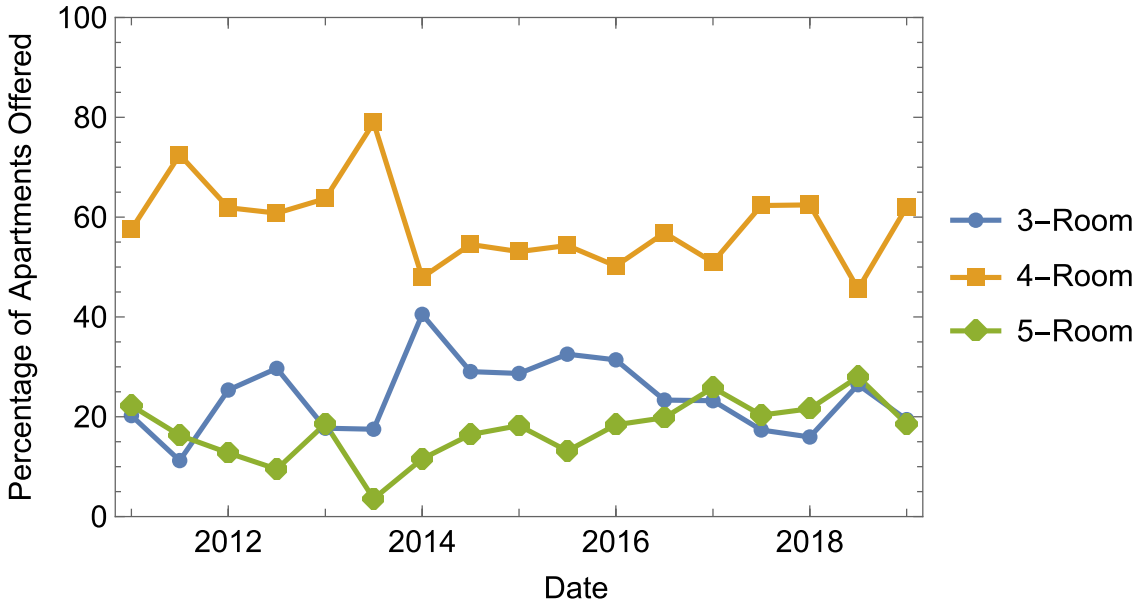

Figure 2: *Percentage of apartments, per half year by type, built in mature neighborhoods.*
Notes*: Mature neighborhoods are limited in land for housing development, but see high demand for housing.*

3 Model

The descriptive facts suggest that the Singaporean government responds to demand, but leave us short on *how* it should design a responsive mechanism when supply is endogenous. We develop a parsimonious model to fill in this gap. We model an allocation mechanism with endogenous

supply of goods, akin to the Singaporean BTO mechanism. Time is discrete with an infinite horizon, $t \in \{0,1,2,\ldots\}$. The agents are young households born with preferences over the goods, apartments. Each period, one household arrives.[16] The government has no apartments available at $t = 0$, but builds one apartment in every subsequent period. Apartments and households can be of $|\Theta| = 2$ types, $\Theta = \{A,B\}$.[17] We use θ_t to denote the type of the arriving household in period t, and ϕ_t to denote the apartment built in that period. Household types are private information, unknown to the government. A household matched to an apartment of the same type receives utility h. A household matched to an apartment with a different type receives utility $l < h$. An unmatched household incurs a per-period flow cost of waiting, c, and remains in its queue.[18]

There is one queue for each type of apartment, queue A and queue B. At the beginning of each period, all agents are informed of the number of households and apartments in each queue. An incoming household chooses the queue it wishes to enter. We denote the queue choice of the period-t household by $d_t \in \{A,B\}$. Before observing the incoming household's choice, the government chooses $\Phi_t^A \in [0,1]$, the probability with which $\phi_t = A$.[19] The timing of the market is summarized in Figure 3.

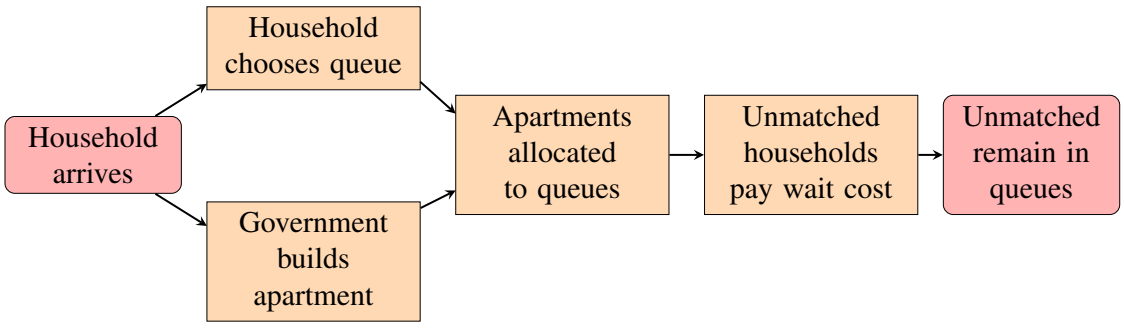

Figure 3: *Market Timing.*
Notes*: This figure describes the timing of the market. One household of uncertain type arrives. Then, simultaneously, the household chooses an apartment queue to join, while the government decides what kind of apartment to build. Apartments are allocated to households in each queue by uniform random lottery. Unmatched households pay a wait cost and remain in their queues till the next period.*

The government's strategy must satisfy *queue-anonymity*: it must treat households within a

[16] When we investigate the effects of competition in Section 5, we relax this assumption, permitting multiple households to arrive at $t = 0$.

[17] In Section B of the online appendix, we show that our results are robust to differing intensities of apartment preferences. For example, some agents may have strong preferences over apartment type, while others prefer to minimize wait times. In Section D, we also show that our results are robust to settings with more than two types of apartments.

[18] In the actual BTO mechanism, households can change the queues they join between periods. In Section C of the online appendix, we evaluate an alternative framework in which queue switching is permitted. We show that the mechanisms presented in the body of the paper remain optimal within this broader class of mechanisms.

[19] The government chooses the location of upcoming projects several cycles in advance, motivating why Φ_t^A is chosen before observing household applications for the current cycle.

single queue identically, without regard to their seniority in the queue. If an apartment is available and at least one household is in the corresponding queue, that apartment is randomly allocated through a uniform distribution to one of the households in the corresponding queue. For instance, if k households are in queue ϕ, and there are $m \leq k$ type-ϕ apartments, each household in queue ϕ has the same m/k probability of receiving an apartment.

We use $s^t = (s_A^t, s_B^t) \in \mathbb{Z}^2$ to denote the net demand of the market in the beginning of period t. For $s_\phi^t > 0$, s_ϕ^t denotes the number of households in queue ϕ. Otherwise, queue ϕ has no households and $-s_\phi^t$ denotes the number of vacant type-ϕ apartments. The public history at the beginning of time t is $H_t = (d_0, d_1, \phi_1, \ldots, d_{t-1}, \phi_{t-1}) \in \{A, B\}^{2t-1}$.

Definition 1. *A **mechanism** μ is a sequence of functions $\{\Phi_t^A\}_{t=1}^\infty$, where each Φ_t^A maps public histories to an assignment probability: $\Phi_t^A : H_t \to [0,1]$.*

We assume the government has commitment power and declares the mechanism μ at the beginning of time. We will focus on Markovian mechanisms that condition only on payoff relevant variables, summarized by the state.[20] Accordingly, where appropriate, we drop time-scripts. The payoff-relevant information in the public history can be summarized by the state s^t. Then, a strategy for the government can be summarized by $\Phi^A : \mathbb{Z}^2 \to [0,1]$.[21]

In state s, $W_\phi^\mu(s)$ will be used to denote the expected wait time for an incoming household that enters queue ϕ. Similarly, $w_\phi^\mu(s)$ denotes the expected wait time at the beginning of the period for a household already in queue ϕ when the state is s. We will drop the reference to the mechanism when the context poses no confusion.

The expected utility in state $s_t = (s_A, s_B)$ for a household that enters queue ϕ is the match benefit from a type-ϕ apartment minus expected waiting costs from queue ϕ:

$$U(d_t, \theta_t, s_t) = \mathbb{E}_\mu \left[l + \mathbb{1}_{\{d_t = \theta_t\}}(h - l) - cW_{d_t}(s_t) \right].$$

When a household enters the queue of its type, we will say it has *applied sincerely*. A household only applies sincerely if doing so maximizes its utility:

$$\theta_t \in \arg\max_{d_t} U(d_t, \theta_t, s_t).$$

[20] We restrict the government's mechanism to be Markovian as motivated by our setting. In informal conversations with Singaporean housing applicants and anonymous officials from HDB, we learned that BTO hews closely to being Markovian to disincentivize the non-needy from gambling for an apartment with high resale potential.

[21] We comment briefly on our approach. Rather than consider the general mechanism design problem, with associated individual rationality and incentive compatibility constraints, we consider the induced market game wherein households take the government's strategy as given and play against one another. In Section 4.3 we show that given the government's choice of mechanism, the government's outcome parameters of interest are well-defined and unique.

The government's objective depends on two elements: the quality of matches and the frequency of unassignment. Since there is always one more household than available apartments, there will always be a minimum of one unassigned household in every period. To normalize the measure of inefficiency, we only consider unassignment above the baseline of one household. In this context, counting the number of vacant apartments is equivalent to counting the number of excess unassigned households. We will use the two measures interchangeably, except when investigating batching in Section 5.2 where the difference between the two measures is relevant.

Because we do not know exactly how the Singaporean government ranks these two sources of inefficiency in practice, we characterize the general solution to the government's problem. The government aims to minimize total inefficiency, with weights α and $1-\alpha$ placed on allocation and unassignment inefficiency respectively. If the period-t household and the period-τ apartment are matched, the match generates 1 unit of allocation inefficiency if $\theta_t \neq \phi_\tau$. Then, m_t, the level of allocation inefficiency in period t, is simply the number of households that did not apply sincerely in some period that were matched in period t. Similarly, we define the unassignment inefficiency in state $s^t = (s_A, s_B)$ as $v_t = \max\{s_A, s_B\} - 1$, the normalized number of unassigned apartments. A mechanism is evaluated by the *average inefficiency* it creates:

$$V(\mu) = \lim_{T\to\infty} \sup \frac{1}{T} \mathbb{E}_\mu \left[\sum_{t=1}^{T} \alpha m_t(\mu) + (1-\alpha) v_t(\mu) \right]. \tag{1}$$

Definition 2. *A mechanism μ^* is **optimal** if it minimizes average inefficiency, $V(\mu^*) = \inf_\mu V(\mu)$.*

In Section 4.3, we show that in the limit as $T \to \infty$, every mechanism generates at least one steady state. Furthermore, even when an optimal mechanism generates multiple steady states, those steady states feature equivalent values of V. This allows us to simplify Equation 1. Let $m(\mu)$ and $v(\mu)$ refer to allocation and unassignment inefficiency in some steady state of μ. Then, the government's problem can be rewritten as:

$$\min_\mu \ \alpha m(\mu) + (1-\alpha) v(\mu).$$

We emphasize that this framework encapsulates utilitarianism, but can also address more general environments. A government wishing to maximize household utility would place a weight of $h-l$ on allocation inefficiency, and a weight of c on unassignment inefficiency to represent the household cost of unassignment, then normalize the sum of weights to 1 by dividing both weights by $h-l+c$.

3.1 Model Discussion

Here we detail the rationale behind several important modeling decisions.

Random Uniform Lottery: We restrict the government to mechanisms that cannot offer priority. In particular, the government cannot utilize a first-in-first-out mechanism under this assumption. Indeed, in the setting presented in this model, were the government permitted to choose an arbitrary mechanism, a first-in-first-out mechanism would always achieve the first-best outcome under any parameter region.[22] As previously mentioned, the RFS system previously utilized a first-in-first-out style mechanism which led to a surplus stock during the Asian financial crisis, draining the HDB's wealth through maintenance and holding expenses.

Private Information: The model implicitly assumes that the government cannot elicit household preferences through means outside the allocation mechanism. Indeed, the Singaporean government was motivated to primarily use household applications to estimate demand. The following quotation is from the HDB website in response to a question regarding the possibility of directly surveying household preferences. "[The Singapore Ministry of National Development] and HDB have considered the Member's suggestion to introduce a register for [BTO] flat applicants. However, there is no assurance that doing so will improve the planning of BTO launches to meet demand since an indication of interest may not accurately reflect actual demand, as there is no commitment to buy" (Mah 2010).[23]

Linear Waiting Costs: We assume waiting costs are linear (as in related work on dynamic matching, e.g., Ashlagi et al. 2018; Baccara, Lee, and Yariv 2020; Leshno 2022). The first reason is normative: exponential waiting costs would imply that an incoming household would be a higher priority for the mechanism designer than a household that has failed to match multiple times. Linear waiting costs take an agnostic stance on this front. The second reason is practical: if waiting costs were exponential, the state space would grow rapidly. Not only would the number of households of each type need to be tracked, but so would their time of arrival.

4 Results

When is it optimal to disregard household preferences when choosing the supply of housing? In this section, we first consider the complete information benchmark to provide context. We then characterize the optimal government strategy under endogenous supply. Using these results we show that the Singaporean and US public housing systems can be simultaneously optimal for their

[22] Consider the following mechanism. Let every incoming household be allocated to a single waiting queue, independent of their type. In every period $t \neq 0$, the government builds an apartment to match the type of the household at the beginning of the queue. Then, every household is incentivized to report truthfully, since their report does not change their expected waiting time.

[23] In particular, the government cannot use a Becker–DeGroot–Marschak (BDM) style mechanism to encourage households to report truthfully (Becker, DeGroot, and Marschak 1964). We draw on recent literature that has shown that BDM mechanisms may not accurately capture willingness to pay. For instance, Lehmann (2015) and Müller and Voigt (2010) show that BDM mechanisms may produce biased estimates of willingness to pay (WTP).

respective environments.

4.1 Perfect Information Benchmark

To begin, we analyze the benchmark case where household types are common knowledge and the government can choose the queue a household enters. Formally, rather than households choosing a queue, d_t, the government learns household types θ_t and also selects $d_t(\theta_t)$.

The government can easily ensure that allocative inefficiency is zero. In order to do so, the government allocates households to queues of their types. That is, to avoid allocating an apartment to a household with a different type, the government sets $d_t(\theta_t) = \theta_t$.

Consider state $(1,0)$. If the incoming household is of type-A, it will be allocated to queue A. Had the government built an apartment of type-B, a vacancy would result. As such, to avoid the possibility of a vacancy, the government must build a type-A apartment to fill the non-empty queue. Hence, $\Phi^A(1,0) = 1$; for similar reasons, $\Phi^A(0,1) = 0$. Under this strategy, the only possible states on the equilibrium path are $(1,0)$ and $(0,1)$. Therefore, these values of $\Phi^A(1,0), \Phi^A(0,1)$, and $d_t(\theta_t)$ define a mechanism that we will refer to as the "first-best" mechanism, μ_{fb}. This mechanism can be depicted by the finite state automaton in Figure 4.

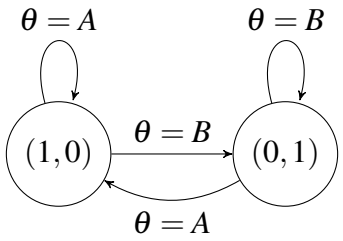

Figure 4: *Finite-state automaton depicting the first-best mechanism, μ_{fb}.*
Notes*: In the first-best mechanism, the government always builds an apartment matching the type of the non-empty queue. Circles indicate states, while arrows depict transitions given an incoming type-θ household.*

Lemma 1. *When the government knows θ_t and can select $d_t(\theta_t)$, μ_{fb} generates 0 inefficiency.*

It is worth noting that the government perfectly responds to demand when information is public. If the previous household is of type θ, then a type-θ apartment is built. Additionally, complete information implies the absence of vacancies in an optimal mechanism. We will later show that vacancies occur with positive probability when the government prioritizes minimizing allocation inefficiency in an optimal mechanism. Vacancies are present in non-mature neighborhoods in the real world, implying that the government lacks the ability to perfectly predict household preferences and that the government prioritizes reducing allocation inefficiency over minimizing vacancies.

4.2 Implementing the First-Best

For the remainder of this paper, we assume household types are not public information. To begin, we prove a simple lemma that restricts the state space. A type-A household prefers queue A if and only if the expected wait from entering queue A is no greater than the expected wait of queue B plus the *benefit-cost ratio*, $\gamma \equiv \frac{h-l}{c}$. To see why, suppose the state is $s = (s_A, s_B)$ and $s_A > 0$. Then, a simple calculation shows that the type-A household prefers to enter queue A if the following inequality holds:

$$u(d_t = A, \theta_t = A, s) \geq u(d_t = B, \theta_t = A, s),$$
$$\implies \gamma \geq \left(\Phi^A(s)\frac{s_A}{s_A+1} + 1 - \Phi^A(s)\right)(1 + w_A(s_{t+1})) - \left(\Phi^A(s)\right)(1 + w_B(s_{t+1})).$$

Observe that the right-hand side is the difference in the expected wait times for both queues. Similar computations can be done when $s_B > 0$ or for a type-B household considering queues B and A respectively.

Lemma 2. *A type-θ household prefers to apply sincerely if and only if $\frac{h-l}{c} \geq W_\theta(s) - W_{\theta'}(s)$.*

The proof of Lemma 2 and all future proofs are relegated to the appendix. The left-hand side of the constraint compares the benefit from sincere matching to the cost of waiting. As households care more about matching to an apartment of their type, they are more willing to accept increases in wait times. Similarly, as the cost of waiting increases, their focus shifts, placing a lower weight on sincerely applying and a higher weight on receiving an apartment immediately.

Now we return to the government's problem under private information, and ask if the first-best can be implemented. The first-best mechanism, μ_{fb}, is still the only one that can achieve zero inefficiency. Lemma 2 lets us determine if households are willing to select $d_t = \theta_t$ in all states by computing the difference in expected wait times for each state and comparing to γ.

We only need to check the incentives for a type-B household in state $(1,0)$. The symmetry of the mechanism implies that the incentives are the same for a type-A household in state $(0,1)$. A simple calculation shows that the wait times from queue A and queue B under the first-best mechanism are $\frac{2}{3}$ and $\frac{4}{3}$ respectively in state $(1,0)$. Then, households are willing to apply sincerely only if $\frac{4}{3} - \frac{2}{3} \leq \gamma$.

Proposition 1. *A mechanism can achieve $m = v = 0$ if and only if $\gamma \geq \frac{2}{3}$.*

For any value of α and $\gamma \geq \frac{2}{3}$, there is a single optimal mechanism. In state $(1,0)$, the government always builds an apartment of type A. Similarly, in state $(0,1)$, the government always builds an apartment of type B. Incoming households always apply sincerely. An outside observer

would see the government responding in a manner commensurate to demand. This response is slightly lagged; it occurs after the demand shock. Since households are much more concerned about correct matching than wait times, the government can maximize efficiency through building apartments that are currently highly demanded.

4.3 When the First-Best Cannot be Implemented

We proceed by assuming that households are impatient enough that a mechanism designed to eliminate all inefficiency would not motivate sincere applications. We show the existence and uniqueness of steady states under any mechanism, then characterize optimal mechanisms conditional on the benefit-cost ratio γ.

Assumption 1 (Non-triviality)**.** $\gamma < \frac{2}{3}$.

Importantly, Assumption 1 sharply restricts household behavior in state $(2,-1)$. In state $(2,-1)$, all households prefer to enter queue B. To see why, consider the expected wait times for each queue under any mechanism. The maximum wait for a household that enters queue B is 0, because a type-B apartment is available. In contrast, the minimum wait for a household that enters queue A is $\frac{2}{3}+\frac{2}{3}\cdot\frac{1}{2}=1$, which arises when the government always builds type-A apartments and all incoming households apply to queue B. Combined with Lemma 2, a type-A household prefers to apply sincerely only if, $\gamma \geq 1$, which violates Assumption 1. Therefore, in state $(2,-1)$, by Assumption 1, any incoming household will enter queue B regardless of her type. This statement also holds for any state with more than two households in queue A. The expected wait from queue A in such a state is strictly larger than in state $(2,-1)$, while the expected wait from queue B remains 0.

Lemma 3. *Under Assumption 1, if* $\max\{s_A, s_B\} > 1$*, in state* (s_A, s_B)*, either type-A or type-B households do not apply sincerely.*

The intuition from the above result is: in extreme states, i.e., when one queue is long and the other has a vacant apartment, all entering households prefer to take the available house. Therefore, the system experiences negative feedback and tends towards states in which demand is less imbalanced. It follows that the state space of any mechanism's steady state is finite. In particular, any optimal mechanism generates a steady state with a finite state space. Furthermore, since it cannot be optimal to remain permanently in state $(2,-1)$ or state $(-1,2)$,[24] the steady-state must be recurrent unless it never transitions between $(1,0)$ or $(0,1)$. Any mechanisms that fail to transition between $(1,0)$ and $(0,1)$ have equal inefficiency levels. Then, the finiteness of the steady-state combined with the recurrence of the state space implies there is a *unique* steady state.

[24] Such a mechanism would be dominated by a mechanism that always remains in state $(1,0)$.

Before stating the result, we define the term *queue symmetric*. Informally, two steady states are queue symmetric if they are equivalent up to relabelling queue A as queue B, and vice versa. Formally, a steady state, $\mathbb{S}$, is queue symmetric if there exists a permutation $\pi : \Theta \to \Theta$ such that the probabilities of any two states, $s, s' \in \mathbb{S}$, are equal under the permutation $\pi(s) = s'$. We next show that any optimal mechanism generates at least one steady state, and furthermore, that outcomes are effectively unique under optimal mechanisms.

Lemma 4.

1. *If μ is an optimal mechanism, then there exists at least one steady-state associated with μ.*
2. *If an optimal mechanism μ generates multiple steady states, those steady states are queue symmetric.*

Lemma 4 implies that the average level of inefficiency is well defined. Either the steady state is unique, or the two possible steady states feature equivalent levels of inefficiency. As such, we will proceed by evaluating mechanisms using the average level of inefficiency in any steady state.

Importantly, when Assumption 1 holds, behavior in state $(1,0)$ is tightly regulated. Suppose the government attempted to avoid vacancies through always building an apartment of type A, i.e., $\Phi^A(1,0) = 1$. Then, incoming households optimally respond by entering queue A irrespective of their type. Since in every period one type-A apartment is built and one household enters queue A, the state remains in $(1,0)$ indefinitely. We will refer to the described mechanism as the *pooling mechanism*, μ_p.[25] Under the *pooling mechanism*, without loss of generality, $\Phi^A(s) = 1$ and $d_t = A$.

Since half of the entering households are of type A, the level of allocation inefficiency is $\frac{1}{2}$ and the level of vacancy inefficiency is 0.

Remark 1. *Under μ_p, allocation inefficiency is $\frac{1}{2}$ and vacancy inefficiency is 0.*

By Assumption 1, any mechanism that always builds an apartment of the type matching the non-empty queue causes all households to prefer said non-empty queue. Then, the state will never change, because every period a new household enters the queue matching the one that the apartment is built for. To prevent this behavior, the government must build the less desirable apartment with positive probability. This insight cautions against endogenous supply policy that responds myopically to demand: households are incentivized to not apply sincerely. Furthermore, a naïve imputation of demand through observing queuing decisions, without a deeper understanding of the

[25] There are technically several mechanisms that result in similar allocations and equivalent levels of efficiency. For the sake of exposition, we will focus on the pooling mechanism described in the main text.

decision problem faced by the applicant, overstates how satisfied households are with the current policy regime.[26]

We proceed by determining the shape of an optimal mechanism with allocation inefficiency below $\frac{1}{2}$. In order to improve overall efficiency, we must have $\Phi^A(1,0) \neq 1$. In particular, $\Phi^A(1,0)$ must be low enough to incentivize type-B households to apply sincerely. In state $(2,-1)$, the government always builds type-A apartments to minimize vacancy inefficiency, since it cannot increase the incentive for households to apply sincerely.

Consider the following mechanism. In state $(1,0)$, the government builds a type-B apartment with probability q. In state $(2,-1)$, the government always builds a type-A apartment. There are never more households in queue A than in state $(2,-1)$, since households always enter queue B when $s = (2,-1)$ according to Lemma 3. Similarly, $\Phi^A(0,1) = q$ and $\Phi^A(-1,2) = 0$. We refer to this mechanism as the **two-state mechanism with parameter** q. Formally, the *two-state mechanism with parameter* q, μ_q, sets $\Phi^A(1,0) = 1-q$ and $\Phi^A(2,-1) = 1$.[27]

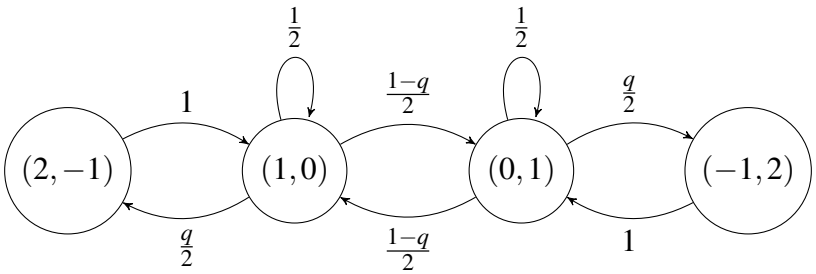

Figure 5: *Finite state automaton depicting the two-state mechanism,* μ_q.
Notes*: Under the two-state mechanism, in extreme states, the government builds an apartment of the type of the long queue. In states* $(0,1)$ *and* $(1,0)$*, the government builds an apartment "of the wrong type" with probability* q*. Arrows denote transition probabilities.*

Proposition 2. *Under Assumption 1, the optimal mechanism takes on one of two forms. Either the pooling mechanism is optimal, or there exists* q^* *such that* μ_{q^*} *is optimal.*

In order to determine the value of q^*, we compute all the equilibrium path wait times. Raising q increases $w_A(1,0)$, the expected wait time for a household in queue A in state $(1,0)$. In return, raising q increases the incentive for an incoming type-B household to apply sincerely.

Given q, we can compute the queue-specific wait times for incoming households in state $(1,0)$. Solving yields $w_A(1,0) = \frac{4q+1}{3-2q}$ and $w_A(2,-1) = \frac{2+q}{3-2q}$. Lemma 2 places a bound on the difference

[26] For instance, when estimating demand for BTO developments in our companion paper (Lee et al. 2024), we modeled households as trading off higher chances of success against being closer to amenities. This insight was corroborated by our conversations with applicants and government officials.

[27] As an aside, a mechanism could potentially sometimes build a type-B apartment when the state is $(2,-1)$. It turns out that the optimal mechanism never does so. In the appendix, we formally show that building "the wrong apartment" in extreme states never increases the extent to which households apply sincerely.

in wait times, if the bound is exceeded, households will not apply sincerely. In order to minimize the probability of entering state $(2,-1)$, $q^*(\gamma)$ is implicitly defined as the solution to:

$$\gamma = \frac{1-q}{2}(1+w_A(1,0)) - q(1+w_A(2,-1)). \tag{2}$$

The right-hand side of Equation 2 is the difference in expected wait times. Inputting the values of $w_A(1,0)$ and $w_A(2,-1)$ implies:

$$q^*(\gamma) = \frac{3\gamma-2}{2(\gamma-3)}.$$

In the remainder of this paper, when we refer to μ_q without specifying q, it is understood that q is optimally chosen, i.e., $q = q^*(\gamma)$.

When a type-B household is indifferent between entering queue A and queue B, a type-A household in state $(1,0)$ will strictly prefer to enter queue A. This result immediately follows from Lemma 2, since the difference in wait times between queues A and B is simply -1 times the difference in wait times between queues B to A, and therefore is less than 0. By the symmetry of the mechanism, households also apply sincerely in state $(0,1)$.

Given the symmetry of the optimal mechanism, when discussing inefficiencies, the relevant statistic is the number of households in the long queue. With some abuse of notation, we use "(s_A, s_B)" with $s_A > 0$ to refer to both states (s_A, s_B) and (s_B, s_A). We proceed by computing the steady state probabilities up to queue symmetric steady states. In the steady state, the transition probabilities at the beginning of a period are given by:

Origin \ Destination	$(1,0)$	$(2,-1)$
$(1,0)$	$1-\frac{q^*}{2}$	$\frac{q^*}{2}$
$(2,-1)$	1	0

We use P_q to denote the steady state measure of μ_q. That is, $P_q(s)$ denotes the probability that the steady state is in state s. Let $M(q)$ denote the transition matrix generated by μ_q, then $P_q = P_q M(q)$. Inverting and solving for P_q yields $P_q(1,0) = \frac{2}{2+q^*}$ and $P_q(2,-1) = \frac{q^*}{2+q^*}$.

The level of inefficiency in the steady state is directly proportional to $P_q(2,-1)$. In state $(2,-1)$ there is a $\frac{1}{2}$ probability that the new household is of type A, while all households enter queue B. State $(2,-1)$ is the only source of allocation inefficiency in equilibrium, since in state $(1,0)$ households always apply sincerely. Then, the average level of allocative inefficiency is proportional to the fraction of time that the steady state is in state $(2,-1)$, and equals $\frac{q^*}{2(2+q^*)}$. Similarly, the level of unassignment is equal to the proportion of time that the steady state is in state $(2,-1)$, contributing another $\frac{q^*}{2+q^*}$ in inefficiency.

In principle, the choice of q need not exactly render type-B households indifferent in state $(1,0)$; larger values of q can also convince households to apply sincerely. The upper limit for q is the point at which type-A households in state $(1,0)$ prefer to enter queue B. This constraint is given the following condition, which compares the benefit-cost ratio to a function of q:

$$\gamma \geq q(1+w_A(2,-1)) - \frac{1-q}{2}(1+w_A(1,0)). \tag{3}$$

Solving Equation 3 for q implies that $q = \frac{3\gamma+2}{2(\gamma+3)}$. Then, households apply sincerely in state $(1,0)$ when $q \in \left[\frac{3\gamma-2}{2(\gamma-3)}, \frac{3\gamma+2}{2(\gamma+3)}\right]$. We utilize the previously calculated values of inefficiency by taking the derivative of inefficiency with respect to q. Unsurprisingly, both derivatives are positive: increasing the probability that the undesired housing is built, increases inefficiency. Since inefficiency is increasing in q, and thus minimized by q^*, we focus on the two-state mechanism with parameter q^*.

The above implies that there exist two potentially optimal mechanisms, the pooling mechanism and the two-state mechanism. In the proof of Theorem 1 we show that no other mechanisms are optimal. Which of the two is optimal is conditional on α, the social planner's preference parameter over allocation inefficiency relative to unassignment inefficiency.

Theorem 1. *Let Assumption 1 hold.*

For $\alpha < \frac{2-3\gamma}{8-5\gamma}$, μ_p is the optimal mechanism. For $\alpha > \frac{2-3\gamma}{8-5\gamma}$, the optimal mechanism is μ_q. Finally, if $\alpha = \frac{2-3\gamma}{8-5\gamma}$, μ_p and μ_q are the only optimal mechanisms.

Theorem 1 shows that either μ_p or μ_q must be optimal. We next consider how the choice of optimal mechanism depends on the selectivity of households, γ. μ_q improves in efficiency in response to a decrease in the ratio of wait cost to relative gain from applying sincerely, while the efficiency of the pooling mechanism remains fixed.

Corollary 1. *Let Assumption 1 hold.*

If the two-state mechanism is optimal for some $\underline{\gamma}$, then it is optimal for all $\gamma \geq \underline{\gamma}$.

As expected, the threshold at which the two-state mechanism is optimal decreases when the relative gain from applying sincerely increases. The intuition for this comparative static is simple. As the relative gain from the correct match increases, households become more willing to apply sincerely, improving the ability of the government to match households properly. Since μ_q takes advantage of sincere applications while the pooling mechanism does not, the inefficiency of μ_q is decreasing in γ.

To understand the direct welfare impact of the chosen mechanism, suppose the government were utilitarian, and aimed to maximize household welfare:

Corollary 2. *Let Assumption 1 hold, and* $\alpha = \frac{h-l}{h-l+c}$. *That is, the government aims to minimize* $\frac{h-l}{h-l+c}m + \frac{c}{h-l+c}v$.

The two-state mechanism is optimal when $\gamma \geq \frac{9-\sqrt{65}}{4}$. *Otherwise,* μ_p *is optimal.*

The implications of Corollary 2 are similar to those of Corollary 1. When γ is small and households care more about wait times than matching to apartments of their type, mechanisms that ignore preferences are optimal. When γ is large and households care about allocation, mechanisms involving sincere applications do better.

In the context of public housing, Corollary 2 enables a simple comparison of the BTO mechanism and "take-it-or-leave-it" mechanisms which offer households no choice. Referencing Arnosti and Shi (2019), there exist welfare parameters for which "take-it-or-leave-it" mechanisms may be optimal. Indeed, more intricate mechanisms feature losses that may not be immediately apparent. Both "take-it-or-leave-it" mechanisms and endogenous supply can be optimal in different parameter regions. Crucially, the social planner's objective and the preferences of the recipients need to be considered before determining the appropriate mechanism. For instance, in the US, public housing is primarily used to prevent homelessness. In our model, homelessness would correspond to a large value for c, the cost of being homeless for an additional period. For contrast, in Singapore, the alternative to receiving an apartment is generally renting or living with family for an additional period. Then, our model suggests that the differences in housing policy between the US and Singapore could be optimal, in contrast to the findings of Arnosti and Shi (ibid.).[28] Changing the design of US public housing policy to incorporate household preferences may leave more apartments vacant, increasing waste.

5 Competition and Market Thickness

How can the government improve on the above mechanisms? Having characterized the constrained optimal mechanism, we proceed by considering the impact of competition on gains from endogenous supply. To begin this section and to fix ideas, we informally characterize our notions of competition and thickness. A household considers a queue to be *competitive* if the household believes there is a high probability another household will later enter that queue. Competition implies that a household expects there are or will be several other households in the same queue it is in. *Thickness*, while related, refers to the number of households applying for queues simultaneously. Thickness implies competition, but a competitive setting may not have a thick market.

We will proceed by showing that competition is undesirable when supply is exogenous. To see why, note that competition has a multiplicative effect on expected wait times. Importantly, when

[28] "The system in New York City... is comparable to disregarding preferences and making take-it-or-leave-it offers" (Arnosti and Shi 2019).

supply is exogenous, competition increases the expected difference in wait times. Then, high levels of competition and exogenous supply dissuade households from applying sincerely. On the other hand, when supply is endogenous, competition gives the government more flexibility to equalize expected wait times across queues. Additional policy flexibility dominates the multiplicative effect of competition on wait times, thus improving efficiency in certain parameter regions (though overall welfare still decreases). In particular, thickness is highly desirable for the government. We show that the government can artificially generate thickness by batching several applications together. Furthermore, when the government prioritizes sincere applications, i.e., when α is large, it is optimal for the government to batch.

5.1 Oversubscription

We proceed by considering a natural form of competition, oversubscription. A housing market is oversubscribed if the number of households applying is larger than the number of apartments available. For reference, under the BTO scheme, apartments of all sizes are oversubscribed, generally at a *minimum* of 3 times. In the previous section, we focused on settings with oversubscription at a rate of 2 times. Larger rates of oversubscription feature a high level of competition, while also directly increasing wait times. We show that the increase in competition dominates, expanding the region within which the government can implement the first best, thereby reducing inefficiency.

To vary the level of oversubscription, we change the number of households that arrive in period 0, without changing the supply of apartments. In every subsequent period, one household arrives, as before. Let N denote the number of households that arrive in period 0, i.e., the surplus of households. At $t = 0$, all households privately choose a queue to enter simultaneously. Households have the same information as in previous sections. Household types are private, but households observe the queues other households have previously entered, and know the state of the market.

Similar to μ_{fb}, a mechanism that implements the first-best cannot allow vacancies, but also must incentivize households to apply sincerely. When the state is $(N,0)$, the government must build a type-A apartment. However, unlike μ_{fb}, which only had a single possible configuration that could possible eliminate all inefficiency, now in any intermediate state $(k,N-k)$, where households of both types are present, the government can freely choose $\Phi^A((k,N-k))$. For $N > 2$, increasing the level of oversubscription always improves the ability of the government to implement the first-best. This is a direct result of the government's increased ability to equalize wait times between the two queues.

We proceed by solving for the optimal supply probabilities, as well as the associated restrictions on γ. It is worth focusing on the shape of the optimal mechanism under oversubscription. The optimal mechanism randomizes; it sometimes builds an apartment in lower demand. In turn, the

state is pushed towards a more extreme level: sometimes the less demanded apartment is built, and the incoming household wants the more demanded apartment. In order to achieve the first best, the mechanism cannot build an apartment of the less-demanded type when there are no households in the corresponding queue. This constraint limits the values of γ under which the first-best can be implemented. A naïve solution would be to default to building the apartment type that is in higher demand. However, doing so strongly disincentivizes incoming households of the underdemanded type from applying sincerely.

Except in this most extreme state $s = (N,0)$, both apartment types are always built with positive probability by the government. Figure 6 displays the optimal mechanism conditional on the number of excess households. In general, the probability that a less desired apartment is built is larger than the fraction of households in the corresponding queue.[29]

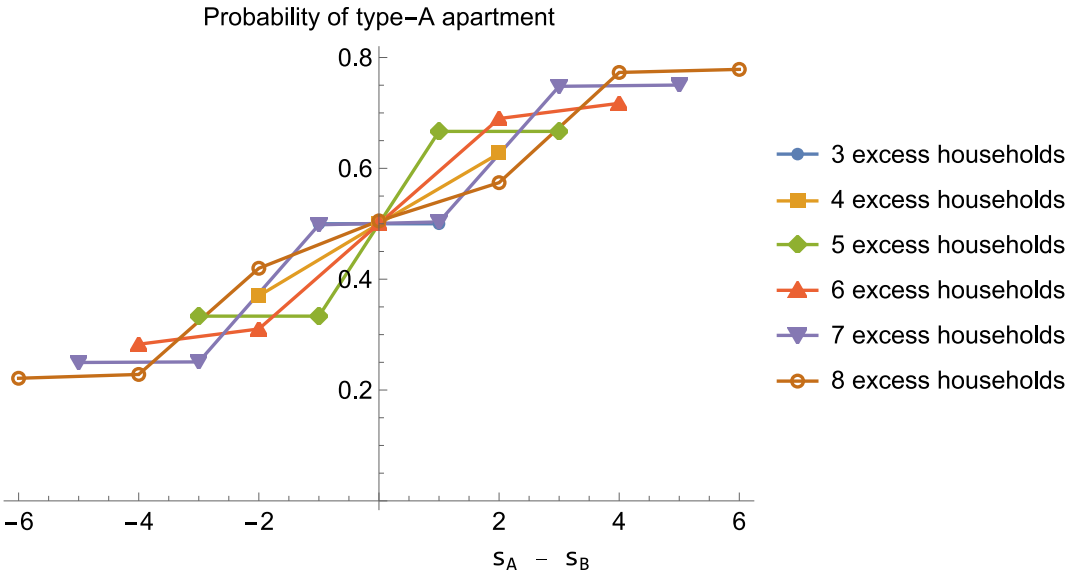

Figure 6: *The probability that a type-A apartment is built under the optimal mechanism.*
Notes*: The x-axis depicts the number of households in queue A minus the number in queue B. The y-axis is the probability that the government builds a type-A apartment in the corresponding state. States* $(0,N)$ *and* $(N,0)$ *are omitted, their corresponding y-values are* 0 *and* 1 *respectively.*

Two forces are at play. On the one hand, increasing the level of competition exacerbates the loss from missing a match in the current period, because waiting times increase across the board. On the other hand, increasing oversubscription increases the number of free variables the government can use in order to normalize wait times between reports. Alternatively stated, the thickness of the market better enables the government to incentivize households to apply sincerely.

To better distinguish between the higher loss from remaining unmatched and the larger latitude to "balance out" wait times, we consider the case of exogenous supply. We frame exogenous

[29] We solve explicitly for the optimal mechanism when $N < 5$, and numerically compute it when $N \geq 5$.

supply as the setting where for any state s, either apartment type is built with equal probability: $\Phi^A(s) = \frac{1}{2}$. Households remain able to freely choose the queue they wish to enter as before. Such a mechanism can never achieve the first-best for any value of γ. There always exists a sufficiently extreme state $(k, N-k)$ such that type-A households prefer to not apply sincerely for a large k. As such, we instead find a weaker condition under which households apply sincerely until there exists a vacant apartment. That is, households apply sincerely except in state $(N+1,-1)$, where all households enter the queue for type-B apartments. We then determine conditions on γ under which households are incentivized to follow this strategy profile. Such a measure underestimates the direct effect of competition while supply is exogenous, yet this further serves our point to show that the government's flexibility in choosing the apartment allocation is crucial.

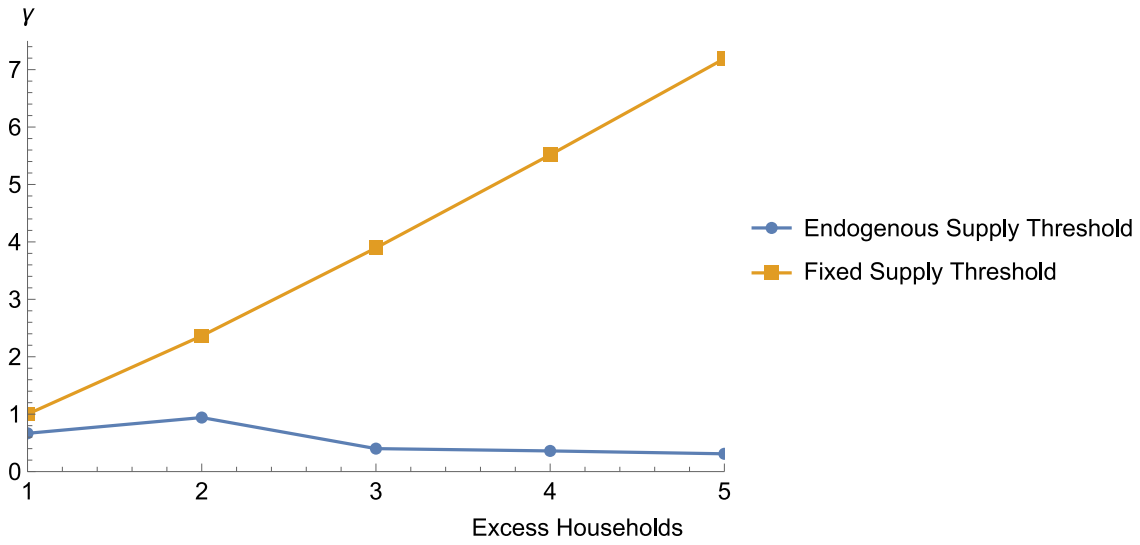

Figure 7: *Lower bound for the benefit-cost ratio, γ, under endogenous and exogenous supply.*
Notes*: The x-axis displays the number of excess households, while the y-axis displays the lower bound on γ. Endogenous Supply Threshold refers to the lower bound on γ such that the expected inefficiency is* 0*. Fixed Supply Threshold refers to the lower bound on γ under which households are willing to apply sincerely so long as no apartments are vacant.*

Figure 7 displays the minimal values of γ as the level of competition increases under this mechanism. Despite the increase in thickness generating an increase in expected wait times, the added government control reduces the difference in wait times. Therefore, when competition increases, the necessary value of γ also rises under exogenous supply, while the opposite holds under endogenous supply.

5.2 Batching

One natural concern is that by directly increasing oversubscription, all households are made worse off due to increased levels of unassignment. Indeed our optimality measure penalizes in-

efficiencies, rather than measuring household surplus. This makes it difficult to compare across different levels of oversubscription. Furthermore, in practice, the government does not control household demand. Nonetheless, the government can (and does) manipulate the timing of apartment applications. For instance, the government could opt to have households apply quarterly or annually.[30] Through delaying the timing of applications, the government increases the level of unassignment in the short run, as incoming households must wait until the next application cycle to enter the market. However, we show that the corresponding increase in market thickness improves the government's ability to incentivize sincere applications. This actually can reduce the overall level of inefficiency, and increase the expected welfare of agents in the market.

We adjust the timing of the model to allow the government to choose the amount of time that elapses between application cycles. At the beginning of the game, before $t = 0$, the government declares the length of application cycles, T. Households that arrive during periods that are not a multiple of T, must wait for the period to reach a multiple of T before entering a queue. While a household waits, the household's type remains hidden, and the household continually pays the flow cost of waiting each period. When the period is a multiple of T, all households not in a queue simultaneously choose a queue to enter. At the same time, except when $t = 0$, the government chooses the types of T different apartments to be built. In effect, the government stockpiles its supply of apartments, then builds all of them simultaneously when a cycle begins.

The government must still choose the apartment supply, $\Phi^A(s)$, before observing agent reports. Since the government can now build T apartments simultaneously, we update our notation. Formally, the government's strategy is $\Phi^A : S \to \Delta\{A,B\}^T$. That is, the government declares a probability distribution over the types of T apartments. Then, the government's objective is:

$$V(\mu) = \min_{(\Phi^A(s),T)} \alpha m(\mu) + (1-\alpha)v(\mu),$$

where m and v are the values of allocation and unassignment inefficiency in any steady state of μ as before. The government faces the same objective as before: to minimize the weighted sum of inefficiencies. In this section, unassignment and vacancies are no longer equivalent: new households continue to arrive while the government delays the building of apartments. Here, we focus on the original definition of unassignment. Notably, unassignment penalizes high levels of batching, as the minimal unassignment inefficiency for a given value of T is $\frac{T-1}{2}$. This formulation stacks the deck against batching, as it implies that the government places a greater weight on the welfare of its citizens than the cost of vacant apartments. Nonetheless, we show that batching is

[30] At the time of the initial release of this paper, in Singapore, the government batched applications at the quarterly level. The Singaporean government has since changed its policy to batch at a lower frequency, only accepting applications three times a year, in line with our recommendations.

still optimal when the government cares deeply about the quality of matches.

We proceed by describing a mechanism with $T = 2$ and finding conditions under which it is optimal. In the appendix, we consider the general setting and show formally the following mechanism is optimal under these conditions. We will abuse notation and refer to it as the $T = 2$ batching mechanism where appropriate.

Suppose the government aimed to ensure all households applied sincerely, $m = 0$. In state $(1,0)$, in order to properly incentivize households, the government randomizes between building one apartment of each type or building two type-A apartments. Let $\Phi^A(1,0) = [(q,(1,1)),(1-q,(2,0))]$. In state $(2,-1)$, the government always builds two type-A apartments, $\Phi^A(2,-1) = [(1,(2,0))]$. We can then compute expected wait times conditional upon the state. These in turn allow the expected wait times conditional on reports to be computed. By Lemma 2 the difference must be bounded by γ in order for households to apply sincerely.

In the appendix, we compute the difference in wait times with respect to q. Then, we can solve for the optimal level of q that minimizes the difference in wait times. We find the minimal difference is given by $\gamma \approx .294$. Hence, the batching mechanism with $T = 2$ induces sincere applications for $\gamma \geq .294$. Notably, batching attains a substantive improvement in motivating sincere applications for the γ requirement under $T = 1$, which was $\gamma \geq \frac{2}{3}$.

Since the mechanism achieves zero allocation inefficiency, the only inefficiency is unassignment, of which there are two sources. The two sources are the default $\frac{1}{2}$ unassignment inefficiency from batching with $T = 2$ and the unassignment inefficiency in state $(2,-1)$. State $(2,-1)$ is only entered from state $(1,0)$ when two type-A households arrive and with $1-q$ probability the government builds two type-A apartments, or when two type-B households arrive and with q probability the government builds one apartment of each type. Notably, since a household has $\frac{1}{2}$ probability of being either type, this implies that in the steady state the probability of state $(2,-1)$ is independent of q. Then, so long as γ is high enough such that households apply sincerely and q is accordingly chosen, the level of unassignment inefficiency is independent of q.

As in the previous welfare analysis, we combine states that are symmetric in queue A and queue B. The transition probabilities are given by:

	$(1,0)$	$(2,-1)$
$(1,0)$	$\frac{3}{4}$	$\frac{1}{4}$
$(2,-1)$	$\frac{3}{4}$	$\frac{1}{4}$

It is then immediate to observe that the expected time spent in state $(2,-1)$ is $P(2,-1) = \frac{1}{4}$. Therefore, the total level of unassignment inefficiency generated by this mechanism is the sum of unassignment from $T = 2$ and state $(2,-1)$, or $\frac{1}{2} + \frac{1}{4} = \frac{3}{4}$. We observe that any mechanism with $T > 2$ occurs a minimum unassignment inefficiency of 1 and therefore is dominated by the optimal

$T = 2$ mechanism when $\gamma > .294$. It remains to determine if a dominating mechanism exists for $T = 2$. We note that a superior mechanism also needs to overcome the pooling mechanism, which generates 0 unassignment inefficiency but $\frac{1}{2}$ allocation inefficiency.

A superior mechanism with $T = 2$ must generate a lower level of unassignment inefficiency. At the same time, it must improve upon the allocation inefficiency generated by the pooling mechanism. However, in order to improve upon the inefficiency of the pooling mechanism, households must not be matched uniformly. In our proof, we show that no mechanism can do both. The key tension is that mechanisms that improve the level of unassignment inefficiency do so at the cost of increasing allocation inefficiency. However, due to the inherent $\frac{1}{2}$ unassignment inefficiency due to setting $T = 2$, such mechanisms are dominated either by the previously defined $T = 2$ batching mechanism or by the pooling mechanism for all α.

As an aside, we note that for $\gamma \in (.294, \frac{2}{3})$, the optimal mechanism depends on α. When α is small, the pooling mechanism is optimal. As α increases, μ_q becomes optimal. Finally, when α is large, the batching mechanism is optimal.

Theorem 2. *When $\gamma \in (.294, \frac{2}{3})$, the optimal mechanism for $\alpha < \frac{2-3\gamma}{8-5\gamma}$ is the pooling mechanism. For $\alpha \in \left[\frac{2-3\gamma}{8-5\gamma}, \frac{34-9\gamma}{38-15\gamma}\right]$ it is μ_q. Last, for $\alpha > \frac{34-9\gamma}{38-15\gamma}$ the optimal mechanism is the T=2 batching mechanism.*

The key implication of Theorem 2 is that batching is only a useful tool when allocation is a greater concern than unassignment. If so, then despite a higher temporary level of unassignment, batching can drastically improve the quality of matches by increasing the thickness of the market.

6 Discussion

We developed a model of allocation with endogenous supply. The model predicts that extreme shifts in preferences are underestimated by simple counts of applications. For instance, if a given apartment type experiences a commonly known surge in popularity, a portion of households will strategically apply for less desirable housing to avoid extended wait times. The model shows that market thickness improves the government's ability to match households to apartments correctly. One policy implication is that the government should delay the timing of housing developments to increase market thickness artificially.

We also provide a normative statement regarding the added benefit from endogenous supply as opposed to exogenous supply. In many situations, the mechanism designer has the ability to change the flow of incoming goods, potentially at some cost. This model suggests that the gains from doing so can be quite large. Indeed, these gains are likely more than a naïve estimate of

household preferences would suggest. This result follows from households having stronger incentives to manipulate when supply is exogenous. Current mechanism design setups generally focus on allocating objects that arrive exogenously. When supply can be adjusted—such as for housing, transportation, and food—the conclusions for optimal design differ starkly from settings where goods arrive randomly.

Beyond the public-housing setting, our study holds broader implications for any market where the mechanism designer controls the supply of goods available. In economics, it is generally taken for granted that markets can achieve an "efficient" solution.[31] However, this notion of efficiency does not speak to other societal objectives, such as avoiding inequality or racial segregation. Through the BTO program, the Singaporean government has treated concerns regarding racial and socioeconomic inequality, while also incentivizing truthful reporting. While we abstract from these concerns in our model, they motivate centralizing the market for new public housing.

[31] For instance, the First Welfare Theorem states that in a competitive market with minor regularity conditions, any equilibrium is Pareto efficient.

References

Abdulkadiroğlu, Atila and Tayfun Sönmez (2003). "School choice: A mechanism design approach". In: *American Economic Review* 93.3, pp. 729–747.

— (2013). "Matching markets: Theory and practice". In: *Advances in Economics and Econometrics* 1, pp. 3–47.

Agarwal, Nikhil, Itai Ashlagi, Michael A Rees, Paulo J Somaini, and Daniel C Waldinger (2019). *An empirical framework for sequential assignment: The allocation of deceased donor kidneys*. National Bureau of Economic Research.

Akbarpour, Mohammad, Shengwu Li, and Shayan Oveis Gharan (2020). "Thickness and information in dynamic matching markets". In: *Journal of Political Economy* 128.3, pp. 783–815.

Altmann, Sam M (2024). "Choice, Welfare, and Market Design". In: *mimeo*.

Armentano, Vincent, Craig McIntosh, Felipe Monestier, Rafael Piñeiro-Rodriguez, Fernando Rosenblatt, and Guadalupe Tuñón (2024). "Movin' on up? The impacts of a large-scale housing lottery in Uruguay". In: *Journal of Public Economics* 235, p. 105138.

Arnosti, Nick and Peng Shi (2019). "How (Not) to Allocate Affordable Housing". In: *AEA Papers and Proceedings*. Vol. 109, pp. 204–08.

— (2020). "Design of Lotteries and Wait-Lists for Affordable Housing Allocation". In: *Management Science*.

Ashlagi, Itai, Maximilien Burq, Patrick Jaillet, and Amin Saberi (2018). "Maximizing efficiency in dynamic matching markets". In: *arXiv preprint arXiv:1803.01285*.

Baccara, Mariagiovanna, SangMok Lee, and Leeat Yariv (2020). "Optimal dynamic matching". In: *Theoretical Economics* 15.3, pp. 1221–1278.

Becker, Gordon M, Morris H DeGroot, and Jacob Marschak (1964). "Measuring utility by a single-response sequential method". In: *Behavioral science* 9.3, pp. 226–232.

Budish, Eric and Estelle Cantillon (2012). "The multi-unit assignment problem: Theory and evidence from course allocation at Harvard". In: *American Economic Review* 102.5, pp. 2237–71.

Galichon, Alfred and Yu-Wei Hsieh (2018). "Aggregate stable matching with money burning". In: *Available at SSRN 2887732*.

Guo, Yingni and Johannes Hörner (2020). "Dynamic allocation without money". In: *TSE Working Paper*.

Housing and Development Board (2014). *HDB Annual Report*. URL: https://www20.hdb.gov.sg/fi10/fi10320p.nsf/ar2014/pdf/HDB_Key%20Statistics_13_14_d9_HiRes.pdf.

— (2019). *HDB Annual Report*. URL: https://services2.hdb.gov.sg/ebook/AR2019-keystats/html5/index.html?&locale=ENG&pn=9.

Lee, Kwok Hao, Andrew Ferdowsian, Yiying Tan, and Luther Yap (2024). "Public housing at scale". In: *Mimeo*.

Lehmann, Sebastian (2015). "Toward an Understanding of the BDM: Predictive Validity, Gambling Effects, and Risk Attitude". In: *Working Paper Series*.

Leshno, Jacob D (2022). "Dynamic matching in overloaded waiting lists". In: *American Economic Review* 112.12, pp. 3876–3910.

Mah, Bow Tan (2010). *Reflections on Housing a Nation*.

Müller, Holger and Steffen Voigt (2010). "Are there gambling effects in incentive-compatible elicitations of reservation prices? An empirical analysis of the BDM-mechanism". In: *Working Paper Series*.

Prendergast, Canice (2016). "The Allocation of Food to Food Banks." In: *EAI Endorsed Trans. Serious Games* 3.10, e4.

Shi, Peng (2022). "Optimal priority-based allocation mechanisms". In: *Management Science* 68.1, pp. 171–188.

Thakral, Neil (2019). "Matching with stochastic arrival". In: *AEA Papers and Proceedings*. Vol. 109, pp. 209–12.

Van Dijk, Winnie (2019). "The socio-economic consequences of housing assistance". In: *Mimeo*.

Verdier, Valentin and Carson Reeling (2022). "Welfare effects of dynamic matching: An empirical analysis". In: *The Review of Economic Studies* 89.2, pp. 1008–1037.

Waldinger, Daniel (2021). "Targeting in-kind transfers through market design: A revealed preference analysis of public housing allocation". In: *American Economic Review* 111.8, pp. 2660–96.

Wong, Maisy (2013). "Estimating ethnic preferences using ethnic housing quotas in Singapore". In: *Review of Economic Studies* 80.3, pp. 1178–1214.

— (2014). "Estimating the distortionary effects of ethnic quotas in Singapore using housing transactions". In: *Journal of Public Economics* 115, pp. 131–145.

A Proofs

Proof of Lemma 2. As noted in the text, in state $s = (s_A, s_B)$, a type-A household prefers to apply sincerely if and only if $u(d_t = A, \theta_t = A, s) \geq u(d_t = B, \theta_t = A, s)$. Rearranging the utility function yields:

$$\begin{aligned} u(d_t = A, \theta_t = A, s) &\geq u(d_t = B, \theta_t = A, s) \\ h - cW_A(s) &\geq l - cW_B(s) \\ h - l \geq c[W_A(s) - W_B(s)] \implies \frac{h-l}{c} &\geq W_A(s) - W_B(s). \end{aligned}$$

An identical computation holds for type-B households. The claim follows. □

Proof of Proposition 1. As shown in the text, the first-best mechanism is the only mechanism that achieves 0 inefficiency. Under the first-best mechanism, the expected wait times for incoming households in state $(1,0)$ are:

$$\begin{aligned} W_B(1,0) &= \sum_{i=0}^{\infty} \left(\frac{1}{2} \cdot \frac{1}{2}\right)^i \\ &= \frac{4}{3}, \\ W_A(1,0) &= \sum_{i=0}^{\infty} \frac{1}{2} \left(\frac{1}{4}\right)^i \\ &= \frac{2}{3}. \end{aligned}$$

Because $W_B(1,0) > W_A(1,0)$, type-A households will always be willing to apply sincerely in state $(1,0)$. Lemma 2 implies that type-B households will be willing to apply sincerely if $\gamma \geq W_B(1,0) - W_A(1,0) = \frac{2}{3}$. By design, the mechanism is symmetric in states $(1,0)$ and $(0,1)$. Therefore, in state $(0,1)$, type-B households are always willing to apply sincerely, and type-A households are willing to apply sincerely if $\gamma \geq \frac{2}{3}$. The mechanism never leaves the set of states $\{(1,0),(0,1)\}$, which implies that no other constraints are relevant. Then, when $\gamma \geq \frac{2}{3}$, the first-best mechanism is implementable, and if $\gamma < \frac{2}{3}$, no mechanism achieves 0 inefficiency. □

In the text, we focused on threshold strategies in equilibrium. Lemma 5 shows that this approach was justified.

Lemma 5. *Any equilibrium generates an outcome equivalent to that generated by some threshold strategy profile.*

Proof of Lemma 5. Suppose not. Then, there exist two states s, s', both of which occur with positive probability, such that $s_\theta > s'_\theta$, but type-θ households enter queue-θ in state s and not state s'. However, $W_\theta(s) > W_\theta(s')$ implies that type-θ households must either strictly prefer to enter queue-θ in state s' or the other queue in state s. Therefore, in at least one of these states, households must have a profitable deviation. Then, the original strategy profile cannot be an equilibrium. □

Proof of Lemma 4.

1. To begin, Lemma 3 implies that in state $(2,-1)$ all incoming households strictly prefer to enter queue B. Then, the state space is bounded by $(2,-1)$ and $(-1,2)$, and a finite number of states are recurrent. Therefore, by standard results in dynamics, at least one steady state exists.

2. For the steady state to fail to be unique, there must be an absorbing state separating either $(2,-1)$ from $(1,0)$ or $(1,0)$ from $(0,1)$. A mechanism that remains in $(2,-1)$ indefinitely cannot be optimal, since such a mechanism would have both allocation and unassignment inefficiency higher than the pooling mechanism. On the other hand, a mechanism that fails to transition between $(1,0)$ and $(0,1)$ implies uniqueness up to queue-symmetric steady states, unless the mechanism has different steady states when beginning with a type-A household as opposed to a type-B household. However, this state non-uniqueness cannot be optimal due to the symmetry of the problem. Suppose a mechanism generated two steady states that were not queue symmetric, with differing levels of efficiency. Then, because both must involve equilibrium behavior of the part of the household, the government could simply use a strategy equivalent to that of the steady state with lower inefficiency everywhere. Through doing so, household behavior must still be equilibrium behavior, and inefficiency would have been lowered, proving that the original mechanism was suboptimal.

□

Remark 2. *When the government utilizes μ_q with parameter q^* an incoming type-A household in state $(1,0)$ does not have an incentive to deviate.*

The following simple result will prove useful for the proof of Proposition 2.

Lemma 6. *The maximal equilibrium allocation inefficiency is $\frac{1}{2}$.*

Proof of Lemma 6. Let $W_A(s) - W_B(s) < \gamma$ for some state s, then each type-A household strictly prefers to enter queue A. As such, allocation inefficiency in s is at most $\frac{1}{2}$ since all type-A households apply sincerely. When $W_B(s) - W_A(s) < \gamma$ a similar argument holds for type-B households. Then, in any state, allocation inefficiency is at most $\frac{1}{2}$. Last, since overall allocation inefficiency

is a weighted average of the allocation inefficiency in each state, the overall allocation inefficiency must be at most $\frac{1}{2}$. □

Proof of Proposition 2. Lemma 3 allows us to focus on mechanisms with state spaces restricted to states $(2,-1)$ through $(-1,2)$. Since incoming households always enter the short queue in $(2,-1)$ and $(-1,2)$, the state can never exceed $(2,-1)$ or $(-1,2)$. In state $(2,-1)$, allocation and unassignment inefficiencies are $\frac{1}{2}$ and 1. Note that both values are weakly larger than inefficiency in state $(1,0)$ by Lemmas 3 and 6. It follows that the government aims to minimize the proportion of time spend in states $(2,-1)$ and $(-1,2)$.

Importantly, if allocation inefficiency is lower than $\frac{1}{2}$, the difference in expected wait times in states $(1,0)$ and $(0,1)$ can be at most γ. That is, $|W_A(1,0) - W_B(1,0)| \leq \gamma$, otherwise type-$A$ households or type-B households strictly prefer to not apply sincerely.

We then optimize over all such possible mechanisms, and show that μ_q minimizes inefficiency. Firstly, note that if a type-B household in state $(1,0)$ strictly prefers to apply sincerely, $\Phi^A(1,0)$ can be increased while ensuring type-B still has incentive to continue applying sincerely. Furthermore, increasing $\Phi^A(1,0)$ reduces the probability that the mechanism enters state $(2,-1)$. Then, the optimal mechanism has $\gamma = W_B(1,0) - W_A(1,0)$.

We proceed by considering a mechanism more general than μ_q. There are two primary differences. First, the government sometimes builds the wrong apartment in state $(2,-1)$, that is $\Phi^B(2,-1)$ is not necessarily 0. Second, type-B households are incentivized to apply sincerely with probability below 1 in state $(1,0)$. Let $x^B(1,0)$ be the probability with which a type-B household enters queue A in state $(1,0)$.

The level of inefficiency generated by this mechanism is:

$$\frac{2(1+x^B(1,0))q - \alpha(q + x^B(1,0)(-2+q+2\Phi^B(2,-1)))}{2(2+q+x^B(1,0)q - 2\Phi^B(2,-1))}.$$

Taking the derivative of the above equation with respect to $\Phi^B(2,-1)$ yields:

$$-\frac{4x^B(1,0)(1+x^B(1,0))(-2+\alpha+\alpha x^B(1,0))(-3+x^B(1,0)+\gamma(-1+x^B(1,0))(-1+\Phi^B(2,-1))+2\Phi^B(2,-1))}{U^2} +$$
$$\frac{\sqrt{-4(2+\gamma(-3+x^B(1,0)))x^B(1,0)(-1+\Phi^B(2,-1))+(-3+x^B(1,0)+\gamma(-1+x^B(1,0))(-1+\Phi^B(2,-1))+2\Phi^B(2,-1))^2}}{U^2}.$$

where $U \neq 0$ is constant with respect to $\Phi^B(2,-1)$. The exact value of U is irrelevant since U^2 must be positive. The numerator is always negative: Note that it can be rewritten as $V + \sqrt{4(2+\gamma(x^B(1,0)-3))x^B(1,0)(1-\Phi^B(2,-1))+V^2}$, where $V = 4x^B(1,0)(1+x^B(1,0))(-2+\alpha+\alpha x^B(1,0))(-3+x^B(1,0)+\gamma(-1+x^B(1,0))(-1+\Phi^B(2,-1))+2\Phi^B(2,-1)$. However, $\gamma \leq \frac{2}{3}$ implies that $4(2+\gamma(x^B(1,0)-3))x^B(1,0)(1-\Phi^B(2,-1)) \geq 0$ and therefore $\sqrt{4(2+\gamma(x^B(1,0)-3))x^B(1,0)(1-\Phi^B(2,-1))+V^2} \geq -|V|$. Then the derivative of inefficiency

with respect to $\Phi^B(2,-1)$ is always positive, and it is optimal to minimize $\Phi^B(2,-1)$ by setting it equal to 0.

We then proceed by taking the derivative of inefficiency with respect to $x^B(1,0)$. The derivative is:

$$\frac{8x^B(1,0)(-3+\gamma+x^B(1,0)+4\alpha x^B(1,0)-\gamma x^B(1,0)}{U^2}$$
$$+\frac{\sqrt{4(2+\gamma(-3+x^B(1,0)))x^B(1,0)+(-3+\gamma+x^B(1,0)-\gamma x^B(1,0))^2)}}{U^2},$$

where again U is a constant we omit because U^2 must be positive. Furthermore, by a similar line of reasoning, this derivative is positive; therefore $x^B(1,0)$ should be set to 0 in an optimal mechanism. We summarize these two findings in the following Lemma.

Lemma 7. *Under Assumption 1, if the pooling mechanism is not optimal, then the optimal mechanism sets* $x^B(1,0)=0$ *and* $\Phi^B(2,-1)=0$.

Then, Lemma 7 implies that the last variable to consider is the value of q. As shown above, q must be minimized subject to the constraint that type-B households in state $(1,0)$ apply sincerely. Therefore, μ_q is optimal anytime type-B households in state $(1,0)$ have incentive to apply sincerely. Then, given the previous computation of q^*, μ_q is optimal. □

We proceed by determining the wait times for μ_q.

Wait time computations for two-state mechanism:

$$w_A(1,0)=\frac{1}{2}\cdot q[1+w_A(1,0)]+\frac{1}{2}\cdot q[1+w_A(2,-1)]+\frac{1-q}{4}\cdot[1+w_A(1,0)],$$
$$w_A(2,-1)=\frac{1}{2}[1+w_A(1,0)].$$

Solving yields $w_A(1,0)=\frac{4q+1}{3-2q}$ and $w_A(2,-1)=\frac{2+q}{3-2q}$. Then, in order for households to apply sincerely, Lemma 2 implies the following constraints on γ:

$$\gamma\geq(1-q)[1+w_A(1,0)]-[q(1+w_A(2,-1))+\frac{1-q}{2}(1+w_A(1,0))],$$
$$\gamma\geq\frac{1-q}{2}(1+w_A(1,0))-q(1+w_A(2,-1)).$$

□

Lemma 8. *Under μ_q, the level of allocation inefficiency is $\frac{2-3\gamma}{14(2-\gamma)}$ and the level of unassignment inefficiency is $\frac{2-3\gamma}{7(2-\gamma)}$.*

Proof of Lemma 8. Since μ_q only generates inefficiency in states $(2,-1)$ and $(-1,2)$, the total inefficiency is directly proportional to the proportion of time spent in those two states. Under the optimal mechanism, the proportion of time in states $(2,-1)$ and $(-1,2)$ is $\frac{2-3\gamma}{7(2-\gamma)}$. With $\frac{1}{2}$ probability, a household fails to apply sincerely, and there is always a vacant apartment in $(2,-1)$ or $(-1,2)$. Therefore, the allocation and vacancy inefficiencies are given by $\frac{2-3\gamma}{14(2-\gamma)}$ and $\frac{2-3\gamma}{7(2-\gamma)}$. □

Proof of Theorem 1. Proposition 2 shows that either μ_q or μ_p is optimal. We proceed by comparing the inefficiencies generated. Lemma 8 directly provides the individual inefficiencies for μ_q. Summing them with weights from the government's objective implies that the total level of inefficiency is:

$$\frac{\alpha}{2}\frac{2-3\gamma}{7(2-\gamma)}+(1-\alpha)\frac{2-3\gamma}{7(2-\gamma)}=\left(1-\frac{\alpha}{2}\right)\frac{2-3\gamma}{7(2-\gamma)}.$$

The level of inefficiency under the pooling mechanism is $\alpha\frac{1}{2}$; there is no vacancy inefficiency and half of the households fail to apply sincerely. Comparing the two and solving for α yields the threshold $\frac{2-3\gamma}{8-5\gamma}$. □

Proof of Corollary 1. We take the derivative of the optimality threshold in Theorem 1 with respect to γ.

$$\frac{\partial\left[\frac{2-3\gamma}{8-5\gamma}\right]}{\partial\gamma}=\frac{-14}{(8-5\gamma)^2}<0.$$

Where the final inequality holds because $\gamma<2/3$ by Assumption 1. □

Proof of Corollary 2. Substituting $\alpha=\frac{h-l}{h-l+c}$ into the threshold in the proof of Theorem 1 implies that the government prefers the two-state mechanism when $2\gamma^2-9\gamma+2<0$ or when $\gamma>\frac{9-\sqrt{65}}{4}\approx .234$. □

Proof of Theorem 2. In order to show that the $T=2$ batching mechanism is optimal, we proceed in a similar manner to the proof of the optimality of the q^* mechanism. As shown in the main body of the paper, whenever the $T=2$ mechanism encourages households to apply sincerely, the $T=2$ mechanism dominates any mechanism with $T>2$. This follows from the fact that if $T>2$, then unassignment inefficiency is 1 at a minimum.

Since we have already characterized the optimal mechanisms with $T = 1$, we proceed by proving the $T = 2$ mechanism is optimal among all mechanisms with $T = 2$. Note that for reasons similar to that in the q^* mechanism, it is never optimal to build type-B apartments in state $(2,-1)$. Formally, the problem is the following:

$$V(\mu) = \min_{\Phi^A(1,0)\in\Delta\{0,1\}^2} \alpha m + (1-\alpha)v(\mu).$$

Standard optimization techniques imply that $\Phi^A(1,0)[0,2] = 0$. Last, we determine the optimal value for q under the $T = 2$ mechanism. We can compute expected wait times for a household already in queue A as the solution to the following pair of equations:

$$\begin{aligned} w_A(1,0) &= \frac{1}{4}\left[\frac{2q}{3}(2+w_A(2,-1)) + (1-q)\frac{1}{3}(2+w_A(1,0))\right] + \frac{1}{2}\cdot\frac{q}{2}(2+w_A(1,0)), \\ w_A(2,-1) &= \frac{1}{4}\cdot\frac{1}{2}(2+w_A(2,-1)) + \frac{1}{2}\cdot\frac{1}{3}(2+w_A(1,0)). \end{aligned}$$

Algebra yields the following solutions with respect to q:

$$\begin{aligned} w_A(1,0) &= \frac{42+196q}{231-50q}, \\ w_A(2,-1) &= \frac{162+4q}{231-50q}. \end{aligned}$$

Conditional on the state, the difference in wait times is:

$$\begin{aligned} w_A(1,0) - w_B(1,0) &= \left(\frac{1}{6}\right)\frac{1725-2357q-62q^2}{231-50q} \\ w_A(2,-1) - w_B(2,-1) &= \frac{453-142q}{1386-300q} \end{aligned}$$

This batching mechanism minimizes the maximum of the two differences when $q = \frac{3}{124}(-833+\sqrt{753905})$, implying that households apply sincerely for γ above $\frac{-453+\frac{213}{62}(-833+\sqrt{753905})}{6(-231+\frac{75}{62}(-833+\sqrt{753905}))} \approx .294$. □

Remaining appendix sections for online publication:

B Preference Intensity

In the main paper, we assumed that agents were one of two types, A or B. Both types valued matched and mismatched apartments comparably (h,l), and shared wait costs c; they only differed in the types of apartments that they preferred. We now consider what happens if there are $2K$ types of agents. These types are given by θ_i, where $\theta \in \{A,B\}$ and $i \in \{1,2,\ldots,K\}$. Each type's utility is characterized by three parameters: h_i, l_i, and c_i. Agents of type θ prefer that variety of apartment. Without loss of generality, we order the types by increasing value of $\gamma_i = \frac{h_i - l_i}{c_i}$.[32] We assume that in each period the incoming household's type is drawn uniformly across all possible types.

To begin, we observe an analogue of Lemma 2 for the generalized type space. Agent i is only willing to apply sincerely if the increase in expected wait time is below γ_i. The proof follows directly from the proof of Lemma 2.

Lemma 9. *A type-θ_i household prefers to apply sincerely if and only if $\gamma_i \geq W_\theta(s) - W_{\theta'}(s)$.*

For the remainder of this section, we identify types with their respective values of γ_i, justified by Lemma 9.

B.1 Strong and Weak Preferences

To begin, we let $K = 2$. First, suppose that $\gamma_1 \geq 2/3$, which implies that $\gamma_2 \geq 2/3$. Then, all types are willing to apply sincerely under the first-best mechanism, which must be optimal.

Next, suppose that $\gamma_1 < 2/3$. As shown in the main body of the paper, the first-best mechanism then fails to motivate households to apply sincerely. Consider the two-state mechanism and choose q such that type-A_1 households are willing to apply sincerely in state $(1,0)$. This two-state mechanism automatically generates an equilibrium. As shown in the main text, $W_A(1,0) > W_B(1,0)$ for any value of q consistent with the two-state mechanism. Therefore, there is no worry that type-A_1 households will manipulate in state $(1,0)$. Furthermore, in state $(2,-1)$, Lemma 3 implies that all households will enter queue B.

It remains to be determined when the two-state mechanism is optimal. There are two other mechanisms that require consideration. First, recall the pooling mechanism, which has been previously considered; second, we introduce a "blended" mechanism, in which q is chosen such that type-B_2 households apply sincerely in state $(1,0)$ while type-B_1 households manipulate. We will

[32] We also assume each type has a unique value of γ_i. If two types have the same θ and γ_i we treat them as a single type.

refer to this mechanism as the *mixed mechanism*. Under the mixed mechanism, households still always enter queue B in state $(2,-1)$. However, only type B_2 households enter queue B in state $(1,0)$, while households of types A_2, A_1, and B_1 all enter queue A instead.

The level of inefficiency under the mixed mechanism can be computed as follows:

$$u_{mm} = \frac{3(-5+\gamma+\sqrt{41-30\gamma+\gamma^2})(1-\alpha/2)}{1+3\gamma+3\sqrt{41-30\gamma+\gamma^2}} + \frac{(4\alpha)}{1+3\gamma+3\sqrt{41-30\gamma+\gamma^2}}$$

We can then compare inefficiencies under the mixed mechanism to those under the two-state mechanism. Notably, for $\alpha > 1/2$ and any values of γ_1 and γ_2, the mixed mechanism is strictly less efficient than the two-state mechanism. To see why, observe that in the mixed mechanism, in every state there is a minimum of $1/4$ matching inefficiency. To contrast, under the two-state mechanism, all households apply sincerely in state $(1,0)$, and the probability of state $(2,-1)$ can be kept sufficiently low. Indeed, the maximum probability of $(2,-1)$, comes when $\gamma = 0$, and is still only $1/7$.

However, as α decreases, the relative value of the mixed mechanism increases. Indeed, when γ_1 and γ_2 are sufficiently far apart, the mixed mechanism benefits from allowing type-B_1 households to manipulate in state $(1,0)$, decreasing the value of q necessary to ensure the correct households apply sincerely.

B.2 Continuum of Household Types

The results in the previous section generalize to when $K = \infty$, and there is a continuous distribution G over each household's realization of γ_i. Let $G(x)$ indicate the probability that a household's value of γ is below x. We require that $G(0) = 0$. Households can be divided into two groups: those who intend to apply sincerely in state $(1,0)$, and those who do not. These groups are characterized by a threshold t: households with γ_i below the threshold manipulate, while those above apply sincerely. Then, the mass of sincere applicants is $1 - G(t)$. We denote this mixed mechanism by μ_t

Theorem 3. *For any distribution G, there exists $t_G \in [0,1]$ such that μ_{t_G} is optimal.*

Proof. To begin, the steady state probabilities are continuous functions of t_G. In turn, since both m and v are continuous functions of the steady state probabilities, this implies that $\alpha m(\mu_{t_G}) + (1-\alpha)v(\mu_{t_G})$ is also a continuous function of t_G. Then, by the Weierstrass extreme value theorem, there exists t_G such that the government's objective achieves a maximum on the compact set $[0,1]$. Last, by previous arguments in Lemma 3, the state can never exceed $(2,-1)$ or $(-1,2)$ and furthermore, all households will manipulate in both of those states. Then, choosing t on $[0,1]$ exhausts

the government's strategy space, and so the previously found value of t_G must define an optimal mechanism. □

In the previous setting, we characterized when the two-state mechanism and the pooling mechanism were optimal. These previously optimal mechanisms admit simple structures: the mixed mechanism with thresholds of $G(0)$ and $G(1)$ respectively. When the threshold is $G(0)$, households always apply sincerely in state $(1,0)$; when the threshold is $G(1)$, households never apply sincerely.

C Queue Hopping

Under the real BTO mechanism, households can freely switch queues between application cycles. In this section, we show that allowing for queue switching does not change the optimality of the mechanisms presented in the main body of the paper. Under those mechanisms, no household wishes to change their queue at the beginning of a period. We formally prove this result below.

To provide intuition for the results that follow, note that the incoming household in a given period always has more information than households currently in a queue. In particular, if a household in a given queue is willing to change queues, then all incoming households will strictly prefer to enter the queue that that household swapped to. Then, correct allocation is impossible when households switch, resulting in high inefficiency in mechanisms that take advantage of swapping. This insight is specific to stationary markets, wherein households continually arrive to be matched. In a static market, where no new households arrive, swapping could very well be part of an optimal mechanism.[33]

The timing is as follows. In every period when the incoming household would enter a queue, all present households simultaneously also choose a queue to enter. That is, households choose the queue they wish to enter without knowledge of the queue other households are about to enter. We denote the period τ choice of the household that arrived in period t by d_t^τ.

In this setting, we need to define the state variables with care. Previously, once a household had entered a given queue, their actual type became irrelevant for both the household and the government. Since they could not switch, incoming households only cared about their selected queue, and the government could no longer influence that household's match. However, now households can switch queues between periods, implying that tracking their type is important. Furthermore, households can carry beliefs regarding the types of other households. This observation is important because a household's type informs their switching probabilities. In practice,

[33] Consider a simple static setting with two households and one apartment of each type. If both households initially apply to the same queue, the household that loses the resulting lottery would prefer to switch queues.

this behavior seems unrealistic. While applicants might observe application rates, they will not track applications household-by-household. Hence, we make the following assumption of naïvete: households only observe the length of each queue, not the types of other households or the history of household-level applications.

Assumption 2. *Households are Markovian —their strategies are a function of the state.*

It is trivial to show that whenever the first-best mechanism was implementable in the original model (i.e., when $\gamma \geq \frac{2}{3}$) it remains optimal in the new setting. To see why, recall that the first-best mechanism instructed households to report truthfully, and always built an apartment matching the type of the household currently present. Then, a household in the queue that matched their type would never wish to change their queue. If they were to do so, that household could not receive an apartment this period, and furthermore ensures the government will build the "wrong" apartment next period. We then focus our attention on the case where the first-best is not implementable; namely, when Assumption 1 holds.

The new incentive constraints implied by the ability to switch are never violated by the pooling mechanism. Switching queues merely ensures that the household cannot receive an apartment in the given period, and will not change the types of apartments the government builds in the future.

We show that the natural translation of Section 4.2's two-state mechanism continues to be an equilibrium in household strategies. As before, μ_q is optimal whenever the pooling mechanism or first-best mechanism are not optimal.

Proposition 3. *Households never switch their queues under μ_q.*

Furthermore, if neither the pooling nor the first-best mechanisms are optimal, then μ_q is optimal.

Proof. We first show formally that no household wishes to change its queue under μ_q on the equilibrium path. We translate μ_q to the current setting by fixing the government's strategy and incoming household's strategies. Old households reenter their queue in every period. Notably, in state $(2,-1)$ incoming households report type B independently of their type. In addition, we require that households do not swap the queue they have entered in later periods. This generates two new constraints, one for each possible state.

In state $(1,0)$, it is easy to see that the type-A household does not wish to switch queues. The probability with which an apartment of type B is built, q, was selected to render incoming type-B households indifferent between the two queues. Relative to incoming type-B households, current type-A households expect less competition in queue A and prefer to match correctly. Then, if an incoming type-B household is indifferent, current type-A households strictly prefer to remain in queue A.

Similarly, in state $(2,-1)$ current type-A households expect the same wait time independent of the queue they select. All incoming households select queue B and the government always builds a type-A apartment. By remaining in queue A, they compete with one other household for one apartment: the other household from the previous period. By switching to queue B, they compete with one other household for one apartment as well: in this case, the incoming household. Furthermore, in the event the household does not receive an apartment in the current period, they prefer state $(1,0)$ to state $(0,1)$.

It remains to prove that μ_q is optimal in the current environment. Proposition 2 implies that μ_q is optimal among mechanisms that do not utilize swapping on the equilibrium path. Next, suppose a mechanism involved households swapping queues with probability k, in state $(2,-1)$, where $0<k<1$. The willingness to randomize would imply that present households are indifferent between the two queues. Such a mechanism must fail to improve upon μ_q with respect to allocative inefficiency in state $(2,-1)$. To see why, note that under the two state mechanism, households always sincerely apply except when in state $(2,-1)$ in which case they are immediately matched and exit the market. It remains to show that such a mechanism cannot reduce the proportion of time spent in state $(2,-1)$ through swapping in either state.

Suppose a mechanism involved households swapping their queue in state $(1,0)$ with probability $0<k<\frac{1}{2}$, while incoming households sincerely apply. The willingness of the present household to randomize implies they are indifferent between the two queues. This is despite the fact that the present household knows there is a $\frac{1}{2}$ chance that the incoming household is of type-B and enters queue B. However, the incoming household of type A then must strictly prefer to enter queue B. To see why incoming households prefer to enter queue B, note that there is a $k<\frac{1}{2}$ chance that the present household enters queue B. If the present household was indifferent between the two queues, then the incoming household must have a strict preference for queue B. Then, allocative inefficiency under such a mechanism is equal to that of the pooling mechanism.

Last, suppose instead that $\frac{1}{2}\leq k<1$. For households to be willing to switch queues, either $\Phi^A(1,0)<\Phi^B(1,0)$ or incoming households must not be applying sincerely. In both cases, allocative inefficiency is comparable to that of the pooling mechanism, which by assumption is suboptimal. □

D More Apartment types

Here we consider the impact of actually increasing the number of apartment types. Suppose there are now $|\Theta|=3$ different types. Households are still born with a type in Θ. If they receive an apartment of their type they gain h in utility, if they receive a different apartment type they receive l.

Consider the first-best outcome. Begin with μ_{fb} from the $|\Theta| = 2$ case. Households still wish to apply sincerely here. A household that did not receive an apartment in the previous period never has an incentive to manipulate because they know an apartment of their type will be built in the current period. Incoming households of a different type face the exact same incentive constraint as in the original model, and so face no incentives to switch. Lastly, incoming households that match the current type never have incentive to switch under this mechanism, and so it remains an equilibrium.

Furthermore, no other mechanism can achieve the first-best, because they risk a positive probability of vacancies. Lastly, this argument holds for all $|\Theta| > 2$; the above arguments do not utilize the fact that $|\Theta| = 3$.

Lemma 10. *The first-best in the standard two type case can be implemented if and only if it can also be achieved in the m type case.*

E Optimal Exogenous Mechanism

Leshno (2022) explores a mechanism with exogenous supply that minimizes manipulation. His paper allows for any method of constructing queues, whereas we have only studied mechanisms similar to BTO. A key assumption in Leshno (ibid.) is that $(1/2)\gamma \geq 1$, or that $\gamma \geq 2$. As shown in Section 4.1, when $\gamma \geq 2$ with endogenous supply, inefficiency is zero. To contrast, when $\gamma = 2$ the best mechanism with exogenous supply can only maintain a queue length of 1. Theorem 2 shows that the policy that maximizes the length of the queue, which in the exogenous setting is equivalent to minimizing manipulation, is the Load Independent Expected Wait (LIEW) policy.

Under the LIEW policy, all agents expect to wait 2 periods upon entering a non-empty queue. This policy minimizes the average wait time, but nonetheless generates manipulation whenever the queue stretches past 1 agent in length. Theorem 2 in Leshno (ibid.) shows that the policy generates a misallocation rate of:

$$m = \frac{2p(1-p)}{(1-p)k + pk + 1},$$

where $k = 2p\gamma - 1 = 1$. In our setting, there is an equal arrival rate of both type-A and type-B households, so $p = 1/2$. This implies that the overall misallocation rate is $m = 1/4$.[34]

E.1 Exogenous Supply Benchmark

Separate from LIEW, we proceed by illustrating what happens when the supply is exogenous and the government cannot change the queueing mechanism. We restrict the government's choice

[34] Vacancies also occur under LIEW, but since LIEW is not designed to reduce vacancies, to provide a fairer comparison, we only consider efficiency losses arising from misallocation.

to $\Phi^A(s) = \frac{1}{2}$, equivalent to what Leshno (2022) called a "balanced" setting. The proportion of incoming households and apartments of each type are equal, implying that in the long run, perfect allocation efficiency is still feasible.

The first household will always wish to apply sincerely in equilibrium since both queues have the same wait times. The strategy of subsequent households will depend upon the current state. When queue B has no households, i.e., the state is (s_A, s_B), for $s_A > 0$, the wait time in queue A is strictly greater than that of queue B. Therefore, incoming type-B households always strictly prefer to apply sincerely. The choice of a type-A household will depend upon the difference in expected wait times. Importantly, the wait time for queue A is increasing in the number of households in queue A, because supply is exogenous. A strategy is a **threshold strategy** if, for some threshold κ, it dictates that a type-θ household, in state $(s_\theta, s_{\theta'})$, enter queue θ if and only if $s_\theta < \kappa$. Under exogenous supply, strategy profiles in equilibrium are equivalent to threshold strategies.

When all households use the same threshold κ, we denote the expected wait time at the start of the period in state s for a household in queue ϕ by $w^\kappa_\phi(s)$. In equilibrium, κ must be large enough to ensure that incoming households no longer wish to apply sincerely when the state reaches $(\kappa, -(\kappa-1))$. When $\kappa > 1$, in equilibrium, two constraints must hold. When there are κ households in either queue, incoming households must prefer the empty queue. Second, when there are $\kappa - 1$ households in either queue, incoming households must prefer to apply sincerely. Lemma 2 then implies the following inequalities that relate the threshold κ to the benefit-cost ratio γ:

$$\gamma \leq \frac{\kappa}{2(\kappa+1)}\Big[1 + w^\kappa_A(\kappa, -(\kappa-1))\Big] + \frac{1 + w^\kappa_A(\kappa+1, -\kappa)}{2},$$
$$\gamma \geq \frac{\kappa-1}{2\kappa}\Big[1 + w^\kappa_A(\kappa-1, -(\kappa-2))\Big] + \frac{1 + w^\kappa_A(\kappa, -(\kappa-1))}{2}.$$

The wait times $w^\kappa_A(s)$ can be computed as the solution to a linear system of $\kappa+1$ equations. We focus on the case $\kappa = 2$ to illustrate a point of comparison with the first-best mechanism. To begin, we compute the solution to the system of equations. This yields wait times of $w^2_A(2,-1) = \frac{5}{2}$, $w^2_A(1,0) = 2$, and $w^2_A(3,-2) = \frac{10}{3}$. Substituting these values into the above equations implies the following:

Lemma 11 (Minimal Exogenous Equilibrium)**.** *If $\gamma \in [\frac{5}{2}, \frac{10}{3}]$ and supply is exogenous, the strategy profile where all households use a threshold of 2 is an equilibrium.*

Proof of Lemma 11. The wait time in a given state and queue depends on the probability of receiving an apartment immediately, as well as the transition probabilities. As such, we can recursively define the wait times $w^m_A(k, -(k-1))$, when $0 < k < m$, as the solution to the following system of

equations using the transition matrix:

$$\begin{aligned}
w_A^m(k,-(k-1)) &= \frac{1}{2}\left[\frac{1}{2}\cdot\frac{k}{k+1}(1+w_A^m(k,-(k-1)))+\frac{1}{2}\cdot\frac{k-1}{k}(1+w_A^m(k,-(k-1)))\right]+ \\
&\quad \frac{1}{2}\left[\frac{1}{2}(1+w_A^m(k+1,-k))+\frac{1}{2}(1+w_A^m(k,-(k-1)))\right], \\
w_A^m(m,-(m-1)) &= \frac{1}{2}\frac{m-1}{m}(1+w_A^m(m-1,-(m-2)))+\frac{1}{2}(1+w_A^m(m,-(m-1))) \\
&= \frac{2m-1}{m}+\frac{m-1}{m}w_A^m(m-1,-(m-2)).
\end{aligned}$$

Next, note $w_A^m(m+1,-m)$ only occurs when a household has deviated and entered a queue that is already at capacity while the government simultaneously fails to build a type-A apartment. Its value comes directly from the previous equations:

$$\begin{aligned}
w_A^m(m+1,-m) &= \frac{1}{2}\frac{m}{m+1}(1+w_A^m(m,-(m-1)))+\frac{1}{2}(1+w_A^m(m+1,-m)), \\
&= \frac{m}{m+1}(1+w_A^m(m,-(m-1)))+1.
\end{aligned}$$

When $m=2$ this process generates the following system of equations:

$$\begin{aligned}
w_A^2(2,-1) &= \frac{1}{2}\cdot\frac{1}{2}(1+w_A^2(1,0))+\frac{1}{2}(1+w_A^2(2,-1)), \\
w_A^2(1,0) &= \frac{1}{2}\left[\frac{1}{2}\cdot\frac{1}{2}(1+w_A^2(1,0))\right]+\frac{1}{2}\left[\frac{1}{2}(1+w_A^2(2,-1))+\frac{1}{2}(1+w_A^2(1,0))\right].
\end{aligned}$$

The solution to the system of equations is given by $w_A^2(2,-1)=\frac{5}{2}$ and $w_A^2(1,0)=2$. Given the previous two values, $w_A^2(3,-2)$ can be computed and is equal to $\frac{10}{3}$. Since $\kappa=2$ is an equilibrium only if both $IC_\kappa(\kappa)$ and $IC_\kappa(\kappa-1)$ are satisfied, this implies the threshold equilibrium $\kappa=2$ requires $\gamma\in[\frac{5}{2},\frac{10}{3}]$. □

As the threshold κ rises, the lowest benefit-to-cost ratio γ that can be sustained in equilibrium also rises. To see why, observe that as the number of households in a queue increases, each household expects to wait for a longer period of time in that queue. Critically, when the state is $(\kappa-1,-(\kappa-2))$, the incentives of an incoming type-A household determine the binding lower bound on γ.

Proposition 4 (Increased threshold for sincere applications requires a higher benefit-cost ratio)**.** *As κ increases, the minimum γ under which the threshold κ strategy profile is an equilibrium also increases.*

Proof of Proposition 4. To begin, note that as the threshold κ increases, the average wait time does as well. To see why, consider the exact difference between a threshold κ strategy and a threshold $\kappa+1$ strategy. In particular, the different arises when there are $\kappa-1$ households in a given queue and the incoming household is of that type. Under the threshold $\kappa-1$ strategy, the incoming household enters the empty queue and is immediately allotted an apartment. Under the threshold strategy, the incoming household enters the full queue, further increasing all present households wait times. Therefore, the ex-ante expected wait time is greater under a threshold equilibrium in any state.

Then, consider the $IC(\kappa)$ constraint. Since each $w_A^{\kappa}(s) > w_A^{\kappa-1}(s)$ and $w_A^{\kappa}(s+1) > w_A^{\kappa}(s)$, it follows that both $w_A^{\kappa}(\kappa,-(\kappa-1)) > w_A^{\kappa-1}(\kappa-1,-(\kappa-2))$ and $w_A^{\kappa}(\kappa+1,-\kappa) > w_A^{\kappa-1}(\kappa,-(\kappa-1))$. Last, a direct comparison of $IC_{\kappa}(\kappa-1)$ and $IC_{\kappa-1}(\kappa-2)$ then implies that the solution to the first must be larger than the solution to the section. □

We focus on $\kappa=2$, the minimal threshold that generates a responsive equilibrium. Suppose $\kappa=1$ was the threshold in some equilibrium. Then, in state $(1,0)$, households enter queue B, no matter their type. Such an equilibrium is not responsive and is equivalent to allocating apartments independently of type.

For contrast, if the government had controlled the supply of apartments, the first-best could have been implemented when $\gamma > \frac{2}{3}$. Furthermore, the level of allocation inefficiency is higher for equilibria with $\kappa \geq 2$ relative to the first-best implementation. We compute the level of allocation inefficiency under exogenous supply when $\kappa=2$. The resulting steady state[35] has a frequency in states $((1,0),(2,-1))$ of $(\frac{2}{3},\frac{1}{3})$ generating an inefficiency level of $\frac{\alpha}{6}+\frac{1-\alpha}{3}$. This inefficiency is directly increasing in the proportion of time spent in state $(2,-1)$. State $(2,-1)$ inherently contains a vacancy and furthermore, households do not apply sincerely. By comparison, when $\gamma > \frac{2}{3}$, there is 0 inefficiency when supply is endogenous.

This analysis suggests that the ability to control the supply of apartments is incredibly important for the government. Even when only a single household is in a queue, households are tempted to manipulate their reports.

F Persistence

In this section, we allow for persistence in the type of arriving households. Recall that in the original model, household types were independently distributed with uniform probability. Here, we assume instead that $\theta_{t+1}=\theta_t$ with probability $p \geq \frac{1}{2}$. The period-0 household still has its type drawn with probability $\frac{1}{2}$ from $\{A,B\}$.

[35] We treat queue-symmetric states as one state, i.e., $(1,0)$ and $(0,1)$ are reduced to $(1,0)$.

We then find conditions under which the government can implement the natural analog of the first best mechanism, μ_{fb}. As in Section 4.1, households must prefer to apply sincerely, and the government must build apartments matching the queue of the old household. Namely, $d(\theta, s) = \theta$ and $\Phi^A(s) = \mathbb{1}_{s=(1,0)}$. We emphasize that the government's hands are equally tied in the setting with persistent types and the previous implementation of the first-best in Section 4.2. There is increased competition, but the level of market thickness remains the same. In a given period, the same number of households are present relative to before, but if a household applies sincerely, it expects an increased level of competition in the following period. We show that persistence hinders the government's ability to implement the first-best outcome. To be exact, the parameter region where households always apply sincerely in μ_{fb} shrinks when p increases.

We proceed in a manner similar to that of Section 4.2. Computing the wait times conditional upon entering a queue implies the difference in ex-ante expected wait is $\frac{1+2p}{2(2-p)}$. Lemma 2 then implies that the difference in expected wait times is the lower bound on γ.

Proposition 5. *Under persistence p, there exists an optimal mechanism with 0 inefficiency when:*

$$\frac{1+2p}{2(2-p)} \leq \gamma. \tag{4}$$

We consider the welfare impact of increasing persistence. To do so, we take the derivative of the difference in wait times with respect to p. The result is positive: as the level of persistence increases, it becomes more difficult for the planner to implement the first-best.

Lemma 12. *When demand is persistent, the threshold under which there exists an optimal mechanism with 0 inefficiency is increasing in p.*

Lemma 12 follows because households expect that future applicants are more likely to compete for the same queue. A household's incentive to not apply sincerely increases in p. If a household applies sincerely, and is not matched in the current period, the household expects a longer overall wait time relative to settings with lower values of p. By not applying sincerely, households significantly decrease their expected wait times, due to decreased future competition. Then, γ must be higher to motivate households to apply sincerely under μ_{fb}.