Build To Order:
Endogenous Supply in Centralized Mechanisms*

Andrew Ferdowsian†, Kwok Hao Lee‡, Luther Yap§

This version: March 31, 2023

Abstract

How should the supply of public housing be optimally designed? Although commonly used queuing mechanisms treat the supply of goods as exogenous, designers can 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 matching quality. Batching applications artificially generates competition and is optimal when the planner places a high weight on match quality.

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

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

*We thank Sylvain Chassang, Moshe Hoffman, Adam Kapor, Alessandro Lizzeri, Daniel McGee, Nick Arnosti, Joseph Ruggiero, Can Urgun, Tim Wang, Alan Wei, and audiences at Princeton University, Stony Brook, 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.

†Department of Economics, Princeton University. Email: a.ferdowsian@princeton.edu
‡Department of Economics, Princeton University. Email: khl@princeton.edu
§Department of Economics, Princeton University. Email: lyap@princeton.edu

1 Introduction

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 strategy when supply is endogenous and evaluate the welfare loss when supply is exogenous. 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.

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 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 *insincere* applications are prevalent—applications where households apply to a queue that does not match their type. We show that under exogenous supply, insincere 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 would prefer 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 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

[1] See e.g., Waldinger (2021), Van Dijk (2019) for empirical work pertaining to Cambridge, MA and Amsterdam; 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. This implies that at least US$3 billion were transacted in apartment value of government-built housing, suggesting large potential gains to improvements in efficiency.

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.[3]

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.[4] 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.[5]

Using data on Singaporean housing applications, we undertake descriptive analysis in Section 6 to show that our focus on strategic behavior is empirically justified. We have suggestive evidence that households strategically apply to queues, and the government takes past demand into account when determining where to build new developments. Our descriptive findings aid us in characterizing the nature of the government's optimal strategy in this mechanism, but leave us short on the specifics. In view of the government's responsiveness to past information and households' strategic applications, we propose a dynamic model to deepen our understanding of the contrast between settings with exogenous and endogenous supply. In such a model, governments respond optimally to applicant behavior, and vice versa.

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. 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.[6] Accordingly, our model differs from standard mechanism design. We

[3]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 is far less than it would be in other public housing markets. For instance, in the US, public housing is primarily used as an alternative to homelessness. We show that in such a setting, where the cost of waiting is high, a pooling equilibrium where all applicants receive the same housing is optimal. This analysis corroborates results in Arnosti and Shi (2019) with the fact that many public housing programs ignore household preferences.

[4]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.

[5]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.

[6]In practice, a household can be given priority in a few specific instances. A mature neighborhood is one that has been a residential area for over 20 years. Non-mature neighborhoods are typically farther away from the city center and have fewer completed apartments. For example, if a household has previously been rejected twice, then applies for a flat in a non-mature neighborhood, they are more likely to receive an early queue number. Since non-mature queues are likely to be undersubscribed in the first place, we abstract from non-constant priorities in our model.

restrict the space of policies, only allowing the government to utilize lottery mechanisms, focusing our attention instead on the optimal supply.

As alluded to beforehand, to assess the importance of endogenous supply, we also consider an analogous setting in which supply is exogenous. With exogenous supply, vacancies and inefficient allocation occur even under arbitrarily high levels of patience. We analyze this setting and show that, compared to the setting with endogenous supply, there is strictly higher allocative inefficiency in all parameter regions.

Our second contribution is to add to a growing literature on thickness in dynamic markets through characterizing when batching multiple application cycles is optimal. We examine the effect of various forms of competition, both when supply is endogenous or when it is exogenous.[7] We show that when the government can control the supply of housing, oversubscription improves the ability of the government to generate responsive mechanisms.[8] 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, increasing the disparities between the exogenous setting and the endogenous setting. We utilize this insight to show that, in the optimal mechanism, thickness is artificially generated in the market by batching applications.

For our final contribution, we leverage novel data on historical BTO applications to develop a lower bound for the proportion of insincere applications. To do so, we compare fluctuations in applications to apartment types to the outflow of successful applicants. We find consistent evidence of households switching their applications between quarters. Such rapid application shifts are difficult to explain through preference changes, supporting our claim that they are the result of strategic applications. Notably, these switches are associated with the relative oversubscription of apartments. This provides evidence for our focus on the household trade-off between quality of match and wait times.

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. Similarly, Budish and Cantillon (2012) explore class selection at the Harvard Business School. They show that students strategically report their preferences in response to the course allocation mechanism.

Another surprising example can be found in the area of food donations. Prendergast (2016) and

[7] We refer to environments where households anticipate low odds of success in the queuing lottery as competitive. Oversubscription refers to environments where the average ratio of households to available apartments is high.

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

Altmann (2023) 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 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. For instance, in public housing, the government can control the type of apartments built and the allocation of apartments.

This paper is closely related to the theoretical literature in matching, much of which stems from studying the problem of optimal student assignment to schools (Abdulkadiroğlu and Sönmez 2003).[9] 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 to our setting, 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

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

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 in 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 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, Ferdowsian, and Yap (2023), we utilize the same dataset to construct a dynamic choice model over housing lotteries and estimate it. Using this model, we were 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, due to the demand response eclipsing the supply increase. However, improving the allocation procedure can complement increasing supply. Specifically, in combination with switching to 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, many government apartments are first introduced into the housing stock via the BTO scheme.

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 of the apartment 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 a first-in-first-out queue is used where households need only apply once, while under BTO households must reapply every quarter. BTO ensured that current applications were an accurate representation of current demand. This motivates our modelling restriction that precludes the government from using allocation mechanisms that reward seniority.[10]

Introduced in April 2001 and taking place each quarter, the BTO exercise allows potential homeowners to ballot for their preferred neighborhood and apartment size.[11] 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.[12] 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. Furthermore, households that have withdrawn more than once incur severe penalties to their priority in future applications. Given that apartments of a similar size are similarly furnished, this motivates our assumption that households who apply to an apartment size and are successful 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 are typically ready for homebuyers to move in within 3 years of the corresponding BTO exercise.[13] To prevent immediate arbitrage, apartments may not be sold on the secondary market before 5 years have elapsed after the initial move-in.

These apartments are often oversubscribed because they are sold at highly subsidized prices,

[10]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 Board 2014). By Financial Year 2018/2019, all residential units under construction were BTO apartments (Housing and Board 2019). See https://www.hdb.gov.sg/cs/infoweb/residential/buying-a-flat/new/modes-of-sale for more details.

[11]For more on the historical context for the BTO, see the government archives: http://eresources.nlb.gov.sg/history/events/d33acabb-a341-460c-8fde-99cf0a9270f4.

[12]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. These quotas are enforced at both the BTO stage and on subsequent resale. This aspect of the housing system has been extensively studied in previous work (see Wong 2013 and Wong 2014), so we will abstract from these concerns.

[13]Wait times recently have lengthened to 5 years because of labor shortages caused by the COVID-19 pandemic.

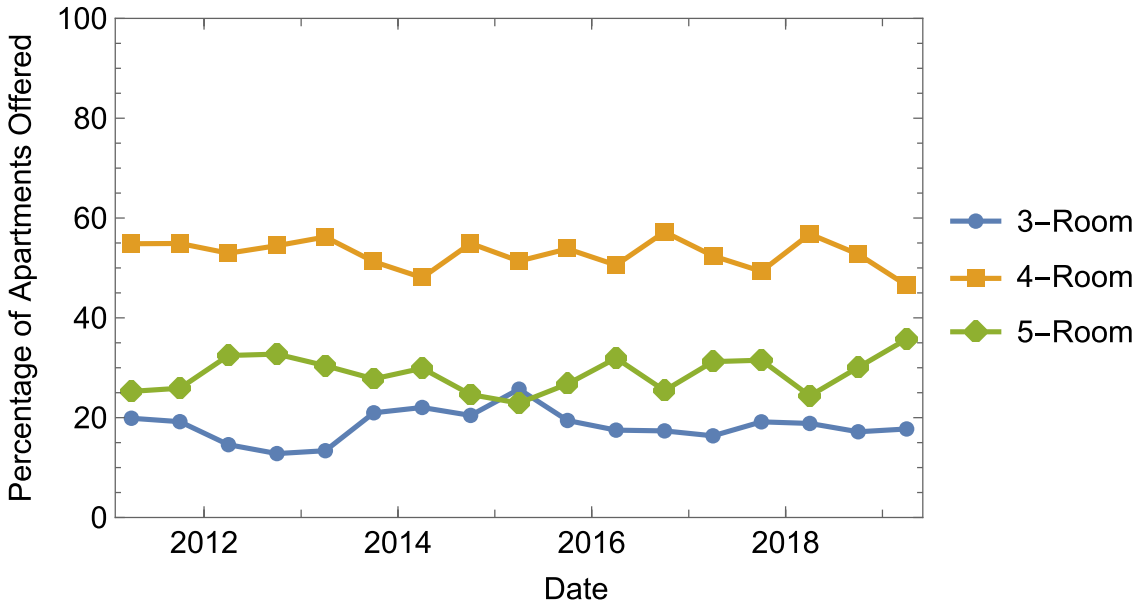


Figure 1: *Percentage of apartments per half year by type, built in non-mature neighborhoods.*

resulting in selection into BTO from the private market. These subsidies cause purchases on the private market to feature a selection effect. Buyers with greater wealth levels default towards the private market as a result of their inability to receive subsidies, biasing the housing purchases observed in the market. Thus, due to the selection of buyers, it is difficult to utilize private market prices as indicative of actual preferences over housing.[14]

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. For instance, while 3- and 5-room apartments initially are built in similar proportions, their relative ratios change dramatically. Comparing Figure 2 and Figure 1, we find a categorical difference between how the government allocates apartments in mature and non-mature neighborhoods. We

[14]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 sufficiently wealthy households that are not capital constrained.

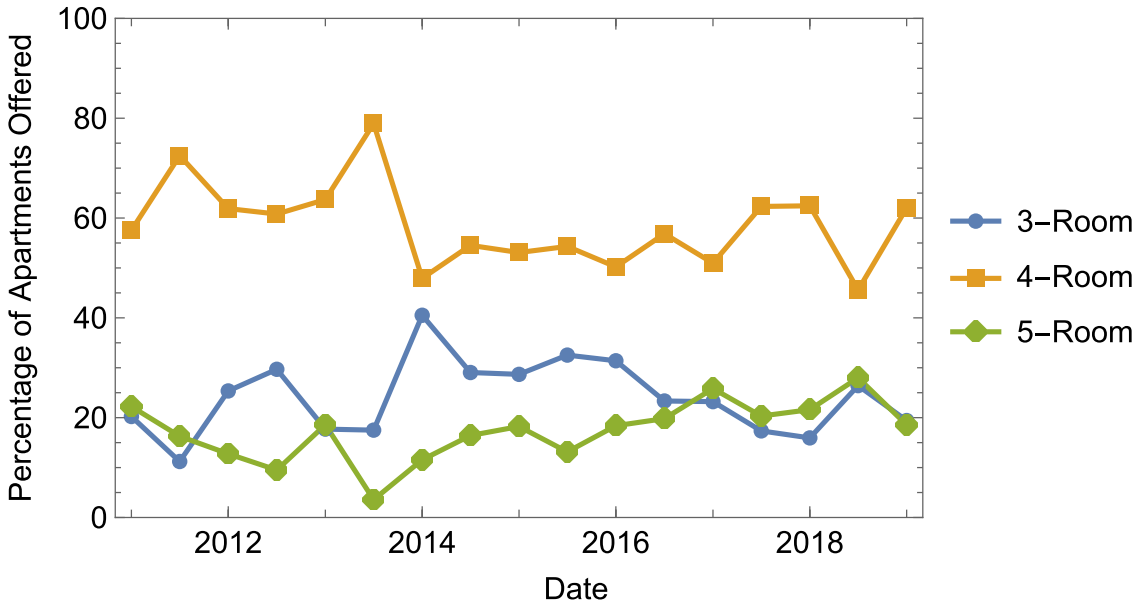


Figure 2: *Percentage of apartments per half year by type, built in mature neighborhoods.*

believe that this difference is due to the government accounting for neighborhood level demand, and using it to determine future offerings within that neighborhood. In Section 6, we conduct a more detailed analysis using BTO data to show that the government takes household demand into consideration.

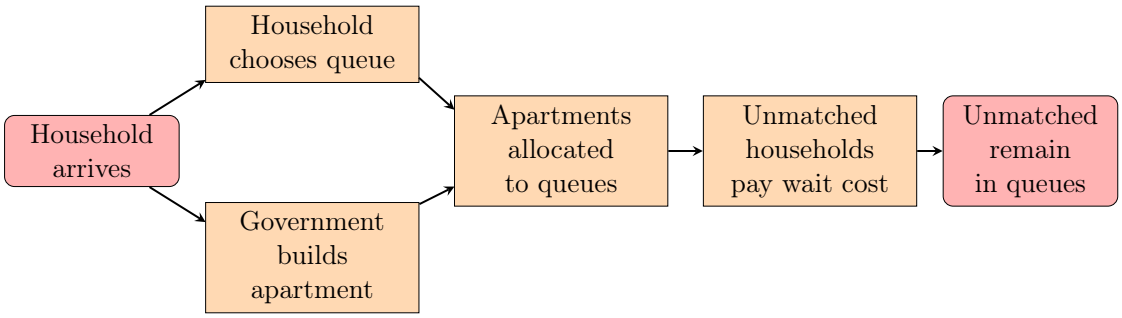
3 Model

The descriptive facts suggest that the Singaporean government is responsive, but leave us short on prescriptions of how to optimally design a responsive mechanism when supply is endogenous. We develop a 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. 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, $\Theta = \{A, B\}$.[15] 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 that is matched to an apartment of the same type receives utility h. A household that is matched to an apartment with a different type receives utility $l < h$. A household that is not matched incurs a flow cost of waiting, c, and remains in its queue.[16]

[15] In the appendix, we consider the model with three types. We show that the optimal mechanism is qualitatively similar to the case with two types.

[16] One worry is that agents can change the queues they have entered in BTO. We consider an alternative mechanism in the appendix. Households are able to choose the queue they wish to enter every period. We show that the natural adaption of the mechanisms presented in the body of the paper remain optimal. See Section B of the Appendix for further details.

Figure 3: *Market Timing*



There is one queue for each type of apartment. The two queues will be referred to as 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 $\phi_t = A$.[17] The timing of the market is summarized in Figure 3.

The government's strategy must satisfy *queue-anonymity*: it must treat households within a 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.[18]

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 $H_t = (x_0, x_1, \phi_1, \ldots, x_{t-1}, \phi_{t-1}) \in \{A, B\}^{2t-1}$.

Definition 1. *A **mechanism** μ is a sequence $\{\Phi_t^A\}_{t=1}^\infty$, 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. 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 $\Phi^A : \mathbb{Z}^2 \to [0, 1]$.[19]

[17] In practice, the HDB chooses the location of upcoming developments at least two cycles in advance, justifying our assumption that the government must choose Φ_t^A before observing household applications for the current cycle.

[18] We discuss our rationale for modelling the allocation of apartments as a uniform distribution in Section 3.1.

[19] A brief comment on our approach is warranted. 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 4.4 we show that given the government's choice of mechanism, the government's outcome parameters of interest are well-defined and unique.

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

In state s, $W^{\mu}_{\phi}(s)$ will be used to denote the expected wait time for an incoming household that enters queue ϕ. Similarly, $w^{\mu}_{\phi}(s)$ 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 for a household that enters queue ϕ is the match benefit from a type-ϕ apartment minus expected waiting costs from queue ϕ, such a household's expected utility can be written as:

$$U(d_t, \theta_t, s_t) = \mathbb{E}_{\mu}\left[l + \mathbb{1}_{\{d_t=\theta_t\}}(h-l) - c\left[d_t \mathbb{1}_{\{s_A \geq 0\}} \frac{s_A}{s_A+1}(1+w_A(s_{t+1})) + (1-d_t)\mathbb{1}_{\{s_B \geq 0\}} \frac{s_B}{s_B+1}(1+w_B(s_{t+1}))\right]\right].$$

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:

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

The government's objective depends on two elements: the quality of matches and the frequency of unassignment.[21] 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. 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 in Section 5.3 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 solution to the government's problem for all possible weightings of allocation and unassignment inefficiency. 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 $\mathbb{1}_{\theta_t \neq \phi_\tau}$ allocation inefficiency. 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*

[20] The focus on Markovian mechanisms is with loss of generality. In Appendix Section C, we show that in settings with higher degrees of oversubscription, a non-Markovian mechanism outperforms the optimal Markovian mechanism.

[21] Unmatched households need not be without a home. In Singapore, most of those applying under the BTO mechanism have outside options, such as rentals or living in their parents' homes.

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 4.4, 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 U. 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:

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

3.1 Model Discussion

Here we detail the rationale behind several important modelling 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. Additionally, abstracted from our model is an additional government concern: family formation. At 1.14 children per woman in 2018, Singapore has one of the lowest reproductive rates in the world. The Singaporean government has publicly stated that increasing their reproductive rate is one of their major objectives. In order to achieve this objective, the government gives married couples with children an extra draw from the housing lottery. If a first-in-first-out mechanism were utilized, the benefit to family formation would disappear as soon as a period had passed and the household was placed in the queue. By contrast, through running the weighted lottery every period, the incentive to form a family persists. This motivates our decision to focus on the specific mechanism utilized in Singapore. Even when the allocation method is constrained, welfare is substantially higher under endogenous supply relative to exogenous supply.

Private Information: The model implicitly assumes that the government cannot elicit household preferences through means outside the allocation mechanism. In particular, the government cannot use a Becker–DeGroot–Marschak (BDM) style mechanism to encourage households to report

[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.

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. Relatedly, Cason and Plott (2014) show that BDM mechanisms can confuse subjects due to its complexity, which in turn results in noisy or biased outcomes. A similar line of reasoning motivated the Singaporean government to primarily use household applications to estimate demand (Mah 2010).[23]

Linear Waiting Costs: We assume waiting costs are linear (as in related work on dynamic matching, e.g., Leshno 2022, Baccara, Lee, and Yariv 2020, Ashlagi et al. 2018). 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

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 d_t, the government learns 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. Therefore $\Phi^A(1, 0) = 1$, and 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 defining $\Phi^A(1, 0), \Phi^A(0, 1)$, and $d_t(\theta_t)$ defines a mechanism that we will refer to as the "first-best" mechanism, μ_{fb}. For illustrative purposes, the mechanism can be depicted by the finite state automaton in Figure 4.

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

[23] The following quotation is from the HDB website in response to a question regarding the possible introduction of a registry to track household preferences. "MND and HDB have considered the Member's suggestion to introduce a register for Build-To-Order (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."

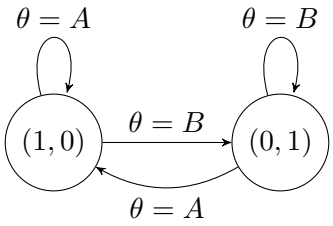


Figure 4: *Finite-state automaton depicting the first-best mechanism, μ_{fb}. Circles indicate states, while arrows depict transitions given an incoming household of type θ.*

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 $\frac{h-l}{c}$. 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 holds:

$$\begin{aligned} &u(d_t = A, \theta_t = A, s) \geq u(d_t = B, \theta_t = A, s) \\ &\qquad \implies \frac{h-l}{c} \geq \left(\Phi^A(s)\frac{s_A}{s_A+1} + 1 - \Phi^A(s)\right)(1 + w_A(s_{t+1})) - \left(\Phi^A(s)\right)(1 + w_B(s_{t+1})) \end{aligned}$$

Observe that the right-hand side is simply the difference in the expected wait times. 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. To simplify notation, let $\gamma \equiv \frac{h-l}{c}$ represent the ratio of the benefit from sincerely applying to the loss from waiting an additional period.

Now we return to the government's problem under private information, and ask if the first-best can be implemented. μ_{fb} is still the only mechanism that could achieve 0 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 $2/3$ and $4/3$ respectively in state $(1, 0)$. Then, households are willing to apply sincerely only if $4/3 - 2/3 \leq \gamma$:

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

For any value of α and $\gamma \geq 2/3$, 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 government response is slightly lagged; it occurs after the demand shock. Since households are more concerned about correct matching than wait times, the government can maximize efficiency through building apartments that are in high current demand.

4.3 Exogenous Supply Benchmark

In this section, we expand on the importance of endogenous supply. To do so, we illustrate what happens when the supply is exogenous. That is, we assume the government cannot choose the type of the incoming apartment. We restrict the government's choice to $\Phi^A(s) = 1/2$. Note that this setting is equivalent to what Leshno (2022) called a "balanced" setting. The proportion of incoming type-A households and apartments are equal, implying that in the long run, perfect allocation efficiency is still feasible, if not necessarily incentive compatible.

The first household will always wish to apply sincerely in equilibrium, since both queues have an equal wait time. 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 κ, it dictates that a type-θ household, in state $(s_\theta, s_{\theta'})$, enter queue θ if and only if $s_\theta < \kappa$.

Lemma 3. *Any equilibrium generates an outcome equivalent to that generated by some threshold strategy profile.*

When all households use the same κ as their threshold, 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, the threshold, κ, must be large enough to ensure that incoming households no longer wish to apply sincerely when the state reaches $(\kappa, -(\kappa - 1))$. A type-A household that enters queue A in state $(\kappa, -(\kappa - 1))$ gives

up the chance to immediately receive l, but gains h minus the expected wait cost. 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 constraints:

$$\text{IC}_\kappa(\kappa) : \gamma \leq \frac{\kappa}{2(\kappa+1)}\Big[1 + w_A^\kappa(\kappa, -(\kappa-1))\Big] + \frac{1 + w_A^\kappa(\kappa+1, -\kappa)}{2}$$
$$\text{IC}_\kappa(\kappa-1) : \gamma \geq \frac{\kappa-1}{2\kappa}\Big[1 + w_A^\kappa(\kappa-1, -(\kappa-2))\Big] + \frac{1 + w_A^\kappa(\kappa, -(\kappa-1))}{2}$$

The wait times $w_A^\kappa(s)$ can be computed as the solution to a linear system of $\kappa+1$ equations.[24] 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 $w_A^2(2,-1) = 5/2$, $w_A^2(1,0) = 2$, and $w_A^2(3,-2) = 10/3$. Substituting these values into the above equations implies the following:

Lemma 4. *If $\gamma \in [5/2, 10/3]$ and supply is exogenous, the strategy profile where all households use a threshold of 2 is an equilibrium.*

As κ, the threshold, increases, the lower bound for γ increases. 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 2. *As κ increases, the minimum γ under which the threshold κ strategy profile is an equilibrium also increases.*

We focus on $\kappa = 2$, because it is the minimal informative threshold. Suppose $\kappa = 1$ was the threshold in some equilibrium. Then, in state $(1,0)$, households enter queue B, no matter their type. In effect, households never apply sincerely. Then, the steady state under the $\kappa = 1$ strategy profile would have $1/2$ allocation inefficiency and 0 vacancy inefficiency. 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 $\gamma > 2/3$. 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[25] has a frequency in states $((1,0),(2,-1))$ of $(2/3, 1/3)$ generating an inefficiency level of $\alpha/6 + (1-\alpha)/3$. 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 $\gamma > 2/3$, there is 0 inefficiency when supply is endogenous.

[24] The general system of equations is listed in the appendix.

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

This suggests that the ability to manipulate the supply of apartments is incredibly important for the government. Even when only a single household is in a queue, households are tempted to apply insincerely. In the remainder of this paper, we explore the optimal mechanism in various situations with endogenous supply.

4.4 When the First-Best Cannot be Implemented

Returning to the setting with endogenous supply, we proceed by assuming that households are unwilling to apply sincerely under μ_{fb}.

Assumption 1. $\gamma < 2/3$.

Now, households weigh the cost of waiting more highly relative to the utility gain from matching to the most desirable apartment type. 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 arises when the government always builds type-A apartments and all incoming households apply to queue B. Then, the minimum wait is $2/3 + 2/3 \cdot 1/2 = 1$. Combined with Lemma 2, a type-A household prefers to apply sincerely only if, $\gamma \geq 1$, which violates Assumption 1. Therefore, in state $(2, -1)$ under Assumption 1 all households will always enter queue B. 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 5. *Suppose Assumption 1 holds. If* $\max\{s_A, s_B\} > 1$*, in state* (s_A, s_B)*, either type-A or type-B households do not apply sincerely.*

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 $(-1, 2)$,[26] 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 $(0, 1)$ has equal inefficiency. Then, the finiteness of the steady-state combined with the recurrence of the state space implies there is a unique steady state. Before stating the result, we define *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 6.

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

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 6 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 vacancy inefficiency 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.[27] Under the *pooling mechanism*, $\Phi^A(s) = 1$ and $d_t = A$.

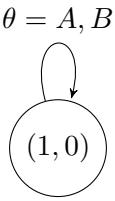


Figure 5: *Finite state automaton depicting the pooling mechanism.*

Note that half of the households were of type A. It then immediately follows that the level of allocation inefficiency is $1/2$ and the level of vacancy inefficiency is 0.

Remark 1. *Under μ_p, the level of allocation inefficiency is $1/2$ and the level of vacancy inefficiency is 0.*

The above logic further implies that under Assumption 1, a mechanism that involves always building an apartment of the type matching the non-empty queue causes all households to prefer allocation to the 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, the government must build the less desirable apartment with positive probability. This provides a cautionary tale for endogenous supply policy that responds myopically to demand. Doing so provides counterproductive incentives for households, and leads to households opting to not apply sincerely. Furthermore, a naïve estimation of public demand through observing queuing decisions overestimates how satisfied households are with the current policy regime.[28]

We proceed by searching for the optimal mechanism with allocation efficiency below $1/2$. In order to improve allocation, we must have $\Phi^A(1,0) \neq 1$. In particular, $\Phi^A(1,0)$ must be low

[27] 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.

[28] In Lee, Ferdowsian, and Yap (2023), households trade off higher chances of success against being closer to amenities.

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 regardless.

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 5. 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 $\Phi^A(2,-1)=1$.[29]

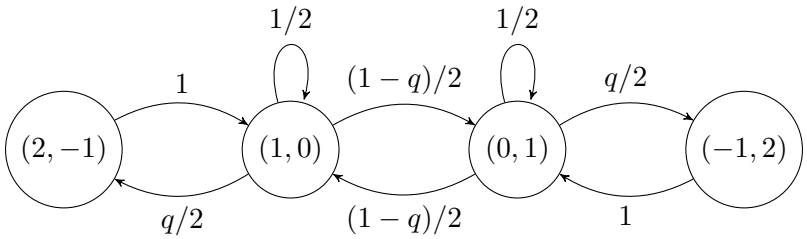


Figure 6: *Finite state automaton depicting the two-state mechanism, μ_q. Arrows denote transition probabilities.*

Proposition 3. *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 possible on-path wait times. Let q denote the probability with which the government builds a type-B apartment in state $(1,0)$. 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 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:

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

[29] As an aside, we note that a mechanism could potentially sometimes build a type-B apartment when the state is $(2,-1)$. In the appendix, we formally show that doing so never increases the extent to which households apply sincerely. As such, the optimal mechanism never does so.

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

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 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, we again refer to states (s_A, s_B) and (s_B, s_A) as (s_A, s_B) when displaying inefficiencies. This can be thought of as identifying two states that have equal numbers of households in the long queue. 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	$(1,0)$	$(2,-1)$
$(1,0)$	$1 - q^*/2$	$q^*/2$
$(2,-1)$	1	0

We use P_q to denote the steady state measure of μ_q. That is, $P_q(s)$ denotes the "amount of time" the steady state spends 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 $1/2$ 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, 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 $(2,-1)$—in this case $\frac{q^*}{2+q^*}$.

In principle, the choice of q does not need exactly render type-B households indifferent in state $(1,0)$, indeed 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 by:

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

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

The above logic implies that there exist two possibly optimal mechanisms, the pooling mechanism and the two-state mechanism. We determine which of the two is optimal, conditional on α, the social planner's preference parameter over allocation inefficiency versus 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 two optimal mechanisms.

Theorem 1 shows that either μ_p or μ_q must be optimal. Theorem 1 then allows us to consider the impact of an increase in selectivity of households. Formally, we show that μ_q improves in efficiency in response to a decrease in the ratio of wait cost to relative gain from applying sincerely.

As expected, the threshold at which μ_q 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.

For a direct welfare comparison, suppose the government were utilitarian. The government would then place a weight of $h - l$ on allocation inefficiency and a weight of c on unassignment inefficiency. Without loss of generality, and to provide consistency with the previous results, we normalize the government's objective.

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

Then, the pooling mechanism is optimal when $\gamma < \frac{9-\sqrt{65}}{4}$. Otherwise, μ_q is optimal.

The implications of Corollary 1 are similar to those of Theorem 1. When γ is small and households care more about wait times than applying sincerely, 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 1 enables a simple comparison of "take-it-or-leave-it" mechanisms and the BTO mechanism. Referencing Arnosti and Shi (2019), this implies that "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 as well as the preferences of the recipients need to be considered before determining the appropriate mechanism. For instance, in the US, public housing is primarily used as an alternative to homelessness. In our model, homelessness corresponds to a large value for c, the cost of being homeless 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 previous work. Changing the design of US public housing policy to incorporate household preferences may also leave more apartments vacant, increasing waste.

5 Competition and Market Thickness

Having characterized the constrained optimal government strategy, we consider the impact of competition on gains from endogenous supply. To begin this section and fix ideas, we informally define 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 competition may not imply thickness.

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 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 a greater deal of flexibility allowing the government to equalize expected wait times across queues. Additional policy flexibility dominates its multiplicative effect, causing competition to actually improve efficiency in certain parameter regions, though overall welfare still decreases. Thickness, in particular, is highly desirable for the government. We show that the government can artificially generate thickness through batching several applications together. Furthermore, when the government prioritizes sincere applications, i.e., α is large, it is optimal for the government to batch.

5.1 Persistence

To begin, we consider the impact of persistent household types. Recall that in the original model, household types were independently distributed with uniform probability. We modify that assumption here, $\theta_{t+1} = \theta_t$ with probability $p \geq 1/2$. The period-0 household still has its type drawn with probability $1/2$ from $\{A, B\}$.

We then find conditions under which the government can implement the natural analog of the first best mechanism, μ_{fb}. In order for it to do so, it must utilize 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 will 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 in p.

We proceed in a manner similar to 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 4. *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 to implement the first-best.

Lemma 7. *When demand is persistent, the threshold under which there exists an optimal mechanism with 0 inefficiency is increasing in p.*

Lemma 7 follows due to households' expectation 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}.

5.2 Oversubscription

Going forward, we return to the no persistence case, $p = 1/2$. We proceed by considering a natural form of competition, oversubscription. We call a housing market oversubscribed when the number of households balloting 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 $2-3\times$. In Section 4.4, we focused on oversubscription at a rate of $2\times$. Larger rates of oversubscription feature an increase in the market thickness, while also directly increasing wait times. In this section, we will show that the increase in thickness dominates, expanding the region within which the government can implement the first-best, thereby reducing inefficiency.

As a simple method of varying 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 appears as before. Let N denote the number of households that arrive in period 0, i.e., the surplus of households. Households have the same information as in previous sections. Household types are private, but households observe the queues other households have entered. In particular, at $t = 0$, all households make their application simultaneously, and do so without information about the other households that are present.

First, we change the number of households appearing at $t = 0$ from 1 to 2. We then derive the optimal mechanism that implements the first-best. In state $(2, 0)$ the government must build a type-A apartment with probability 1, to avoid vacancies. In state $(1, 1)$ the government builds a type-A apartment with probability $1/2$. Without loss of generality, suppose this probability was less than $1/2$. Then, in state $(0, 2)$ type-A households wish to apply sincerely only if type-B households wish to apply sincerely in state $(2, 0)$. The probability of building a type-A apartment could be increased, strictly improving type-A household's willingness to apply sincerely.

Since when attempting to achieve the first-best there are no vacancies, there are only three possible resulting states $(2,0),(1,1),(0,2)$. We treat the first and last state symmetrically.

Solving for wait times under this mechanism, we find that $w_A(2,0) = \frac{25}{17}, w_A(1,1) = \frac{27}{17}$. Using these values, we can then compute the incentive for new arrivals to apply sincerely in all states. First, in state $(1,1)$, the expected waiting time is independent of the queue entered, and so households always prefer to apply sincerely. Next, we ensure both types wish to apply sincerely when $s = (2,0)$, in order to do so, we compute the difference in expected waiting times conditional on entering queue A as opposed to queue B. The expected wait time from entering queue A is $2/3(1+w_A(2,0))$, while entering queue B gives a wait time of $1 + w_A(1,1)$. Taking the difference and simplifying generates a value of $\frac{16}{17}$.

Then, a type-A household never has an incentive to not apply sincerely in state $(2,0)$ since doing so entails increasing their expected waiting time. By comparison, type-B households have a strong incentive to not apply sincerely. We do not consider the incentive to apply sincerely in state $(1,1)$, because the mechanism implies waiting times are equal between queues. We show in the appendix that when $N = 2$, the first-best can be implemented iff $\gamma \geq \frac{16}{17}$.

In particular, comparing this value to the constraint when $N = 1$ implies that the immediate effect of oversubscription was to decrease the ability of the government to implement the first-best outcome. It is instructive to consider why this occurs. While the level of competition is higher, the government's hands are tied when it comes to designing the mechanism. The allocation of households to queues is fixed to avoid allocation inefficiency. In order to avoid unassignment, it must choose $\Phi^A(2,0) = 1$ and $\Phi^A(0,2) = 0$ to avoid unused apartments. Last, in state $(1,1)$ it must choose $\Phi^A(1,1) = 1/2$ as any other value either unbalances the wait times from entering queue A or queue B, and thereby fails to optimize. Then, $N = 2$ maintains the pernicious effects of competition that were present in the persistence extension, while not providing new tools to deal with the increase in wait times.

Corollary 2. *If γ is high enough to incentivize households to apply sincerely under the first-best mechanism when $N = 2$, then γ is also high enough to incentivize households to apply sincerely under the first-best mechanism when $N = 1$.*

The Markovian assumption has some bite. The previous result arises because we restrict consideration to mechanisms that only condition upon the current values of payoff relevant variables. When the state is $(1,1)$, both types of households are more than willing to apply sincerely. Then, the apartment construction probabilities can be altered in order to incentivize households to apply sincerely in extreme states. Since the binding constraint comes from a type-B household in state $(2,0)$, we increase the probability that the government builds a type-B apartment when the current state is $(1,1)$ and the previous state was $(2,0)$. There also exists an upper bound: when type-B apartments are built too often, new type-A households will not apply sincerely. In Appendix Section C, we show that when the Markovian assumption is weakened, this increase in competition is helpful and the first-best can be implemented for $\gamma \geq 0.48$.

Next, we consider the impact of further increasing N. We show that the negative impact of oversubscription on efficiency when there are two excess households is an anomaly. A mechanism that implements the first-best in this setting cannot allow vacancies, but also must incentivize households to apply sincerely. As such, when the state is $(N, 0)$, the government must build a type-A apartment. However, for any intermediate state $(k, N-k)$, where households of both types are present, the government can freely choose any probability for $\Phi^A(s)$. Of course, as discussed above, when $N = 2$, the government has no additional flexibility in choosing $\Phi^A(s)$. 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. Thus, 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 places a hard constraint on the value 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 under demanded 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 7 displays the optimal mechanism for varying numbers 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.[30]

Based on our previous analysis, 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. This outcome relates to the thickness of the market, and how it better enables the government to match households correctly.

To better distinguish between the two forces, we return to the previous baseline where the government randomly builds either apartment type with probability 1/2, $\Phi^A(s) = 1/2$. Households freely choose the queue they wish to enter as before. Such a mechanism can never implement the first-best for any value of γ. There always exists a sufficiently extreme state $(x, N-x)$ such that type-A households prefer to not apply sincerely for x large. Instead, we 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

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

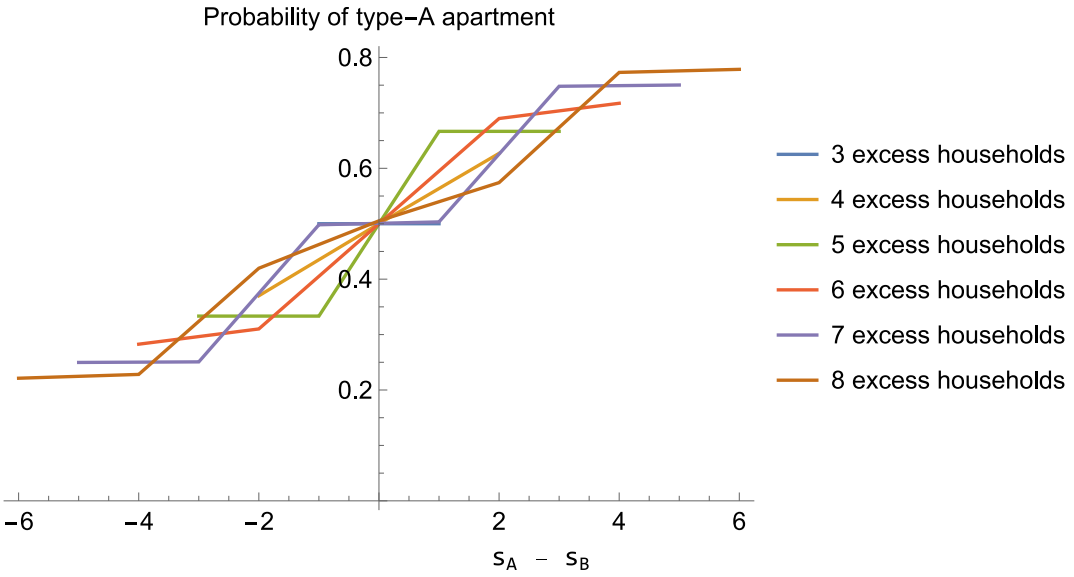


Figure 7: *The probability that a type-A apartment is built under the optimal mechanism with over subscription. The x-axis depicts the number of households in queue A minus the number in queue B. The y-axis is the probability that the government builds a type-A apartment in the corresponding state. States* $(0, N)$ *and* $(N, 0)$ *are omitted, the corresponding y-values are* 0 *and* 1 *respectively.*

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 8 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.

5.3 Batching

One natural concern is that by directly increasing oversubscription, all households are made worse off due to increased levels of unassignment. Furthermore, in practice, the government does not control household demand. Nonetheless, the government can manipulate the timing of apartment applications. For instance, the government could opt to have households apply quarterly or annually.[31] 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.

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

[31] In Singapore, the government batches applications on the quarterly level.

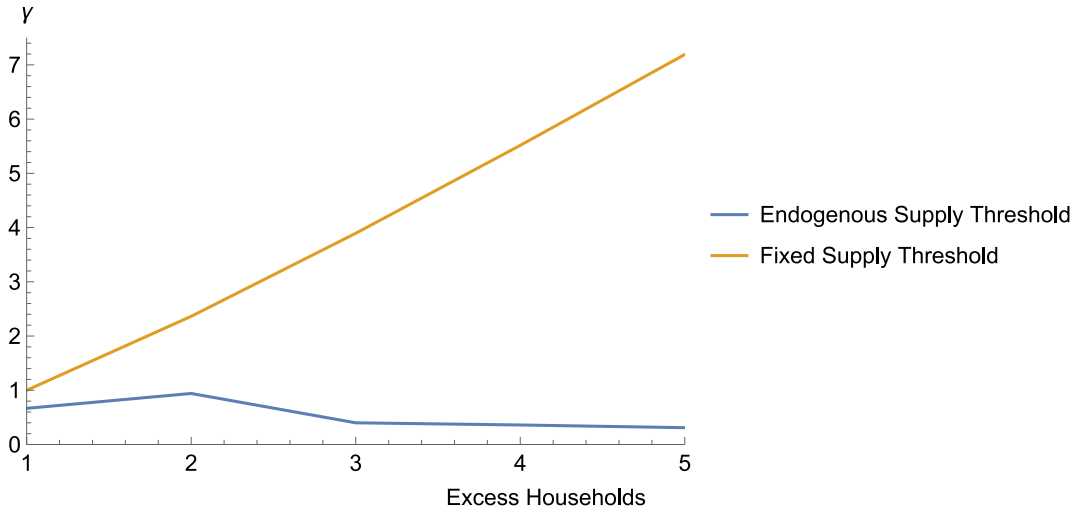


Figure 8: *Lower bound for γ under endogenous supply and exogenous supply. The x-axis displays the number of excess households (in the original model $x = 1$), while the y-axis displays the lower bound on γ. Endogenous Supply Threshold refers to the lower bound on γ under which the first-best can be implemented. Fixed Supply Threshold refers to the lower bound on γ under which individuals are willing to apply sincerely while no apartments are vacant.*

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 while it remains unmatched. 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, $\Phi^A : S \to \Delta\{A, B\}^T$. 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. This is due to the fact that 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 values of T, as the minimal unassignment inefficiency for a given value of T is $(T-1)/2$. In a sense,

batching procedures are overly penalized, implying that the government places a greater weight on the welfare of its citizens relative to the cost of vacant apartments. Nonetheless, we show that batching is 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 $T = 2$ 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 $\gamma \approx 0.294$. Hence, the batching mechanism with $T = 2$ induces sincere applications for $\gamma \geq 0.294$. Notably, batching attains a substantive improvement for the γ requirement under $T = 1$, namely $\gamma > 2/3$.

Since the mechanism achieves 0 allocation inefficiency, the only inefficiency is unassignment of which there are two sources. The two sources are the default $1/2$ 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 $1/2$ 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 A and B. The transition probabilities are given by:

	$(1, 0)$	$(2, -1)$
$(1, 0)$	$3/4$	$1/4$
$(2, -1)$	$3/4$	$1/4$

It is then immediate to observe that the expected time spent in state $(2, -1)$ is $P(2, -1) = 1/4$. Therefore, the total level of unassignment inefficiency generated by this mechanism is the sum of unassignment from $T = 2$ and state $(2, -1)$ or $1/2 + 1/4 = 3/4$. 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 $\gamma > 0.294$. It remains to determine if there exists a superior

mechanism for $T = 2$. We note that a superior mechanism also needs to overcome the pooling mechanism, which generates 0 unassignment inefficiency but 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 the appendix, 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 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 $\gamma \in (0.294, 2/3)$, the optimal mechanism depends on α. When α is small, $\alpha \approx 0.17$ the pooling mechanism is optimal. As α increases, μ_q becomes optimal. Finally, when α is large ≈ 0.94, the batching mechanism is optimal.

Theorem 2. *When $\gamma \in (0.294, 2/3)$, 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 this is the case, then through increasing the thickness of the market at the cost of increasing the temporary level of unassignment, batching can drastically improve the quality of matches.

6 Descriptive Analysis

In this section, we present several descriptive findings to show that, 1) the Singaporean government does indeed take household demand into consideration when determining supply and 2) the BTO mechanism matches households to apartments below their top preference a significant fraction of the time.

6.1 Data

Our data are taken from the HDB. The data set comprises applications to all BTO developments between 2012 and 2020. We refer to each period with applications and matching as a "cycle."[32] These data were constructed through liberal use of the Wayback Machine to scrape historical BTO results from the HDB website.[33] We split the apartment data into types based on the number of rooms, i.e., 3-, 4-, or 5- room apartments.[34] The types of apartments in the previous theoretical

[32]The number of cycles per year varies during this period: there are 6 cycles each year in 2012-2014; 3 cycles in 2015 and 2020; and 4 cycles in the remaining 4 years, resulting in a total of 40 cycles.

[33]http://www.archive.org/

[34]Applicants apply to a (location, size) pair in every cycle. The majority of the population chooses among 3-, 4-, and 5- room apartments. The Singaporean Public Housing Authority chooses the proportion of each type of apartments (characterized by size) in each period, but often does not have a choice of location.

Regressor	Estimate	Std. Error	t-stat	p-value
(Intercept)	0.495	0.192	2.581	0.015
$dd4_{t-3}$	0.435	0.167	2.595	0.015
$dd4_{t-4}$	0.159	0.176	0.904	0.374
$ss4_{t-1}$	0.183	0.183	1.004	0.324
$ss4_{t-2}$	0.295	0.161	1.829	0.078
$ss4_{t-3}$	-1.032	0.304	-3.395	0.002
$ss4_{t-4}$	0.010	0.336	0.029	0.977

Table 1: *Regression of proportional supply of 4-room apartments ($ss4_t$) in period t on proportional supply of 4-room apartments in period $t-1$ to $t-4$, and proportional demand of 4-room apartments in periods $t-3$ and $t-4$ (i.e., $dd4_{t-3}$ and $dd4_{t-4}$). There are $T = 40$ cycles.*

sections corresponds to the number of rooms provided. We use "supply" to refer to the number of apartments of a certain type available in a given period. Similarly, "demand" will refer to the number of applications in a period for those apartments. We aggregate the data across mature and non-mature neighborhoods.

6.2 Lagged Effects of Supply and Demand

To mimic the model, we use proportional demand and supply. Rather than using the total number of k-room apartments built, we transform the data to calculate the per-period proportion of k-room apartments built.

$$dd4_t = \beta_0 + dd4_{t-3}\beta_1 + dd4_{t-4}\beta_2 + \sum_{l=1}^{4} ss4_{t-l}\gamma_l + u_t$$

We regress proportional supply on proportional demand from 3-4 cycles ago and lags of supply from 1-4 cycles ago. This specification is chosen because the Public Housing Authority announces its developments approximately two cycles before the cycle takes place, so it can only condition on the observed demand from 3-4 cycles before. From the results of the regression, displayed in Table 1, the third lag of demand is significant at the 5% level in predicting supply. The coefficient is also positive, suggesting that when there is higher demand for 4-room apartments in previous periods, the corresponding supply increases in subsequent periods. Lags of supply are not significant when regressing demand on lags of demand and supply. Hence, demand likely does not depend on past supply. We have suggestive evidence that the government takes previous household requests into consideration when determining what types of new apartments to build, which justifies the modelling focus on the government's problem.

6.3 Estimating the Upper Bound on Sincere Applications

We only observe the total number of applicants in each queue. Hence, we cannot observe true household preferences, or even track households at the individual level to observe if their applications are changing over time. This makes it difficult to measure match inefficiency, as we do not know whether a match corresponds to a household receiving its top choice.

Nonetheless, we can estimate an upper bound on truth-telling with aggregate data. We assume individual-level preferences do not change over time and no exiting. Consider the number of households that apply for an apartment type in a given period minus the number of households that actually receive that apartment. Call this number the *oversubscription* (os) for that apartment type. In the next period, the *number of households applying to that apartment type* (dd) should be at least as high as the current period's oversubscription. If it is not, this means that households must have applied for different apartment types (i.e., they have switched). Then, in period t, the statistic $os_{t-1} - dd_t$ is a *lower bound on the number of period-t switchers* (lsw_t). If $os_{t-1} < dd_t$, we take the conservative view that there are no switchers. Hence,

$$lsw_t := \max\{0, os_{t-1} - dd_t\}.$$

Switching can occur at the individual level, but be indiscernible in the aggregate. As such, lsw is a lower bound for the number of switchers. For instance, if one household switches its application from 3-room apartments to 4-room apartments and another switches its application from 4-room apartments to 3-room apartments, aggregate applications would look identical for both periods. We would be unable to tell that two switches had occurred. The number of switchers is a lower bound for the extent of misreporting. With persistence in household preferences, switchers must have misreported their type at least once. Further, some households may consistently misreport their type, and would not be captured by the switching statistic. Hence, our lsw statistic is a conservative lower bound for the total extent of misreporting in the system.

Figure 9 shows there are several instances where the demand in a given period is below the oversubscription rate from the previous period, indicating the existence of switchers.

With a lower bound on the extent of misreporting, we can approximate a lower bound for mismatch present in the economy. Due to the uniform allocation lottery, the probability that any household receives an apartment in a given period for a given type is simply $P(match_t) = ss_t/dd_t$, where ss_t is the period t supply. Then, the expected number of switchers that are matched to an apartment in period t is:

$$lmis_t := P(match_t) lsw_t$$

Since lsw_t is a lower bound for the total number of people misreporting in period t, the expected number of mismatches in period t must be at least as high as $lmis_t$.[35] We then normalize $lmis_t$

[35] Since people are matched to their reports, the number of people who misreport and are matched is equal to the number of mismatches in a given period.

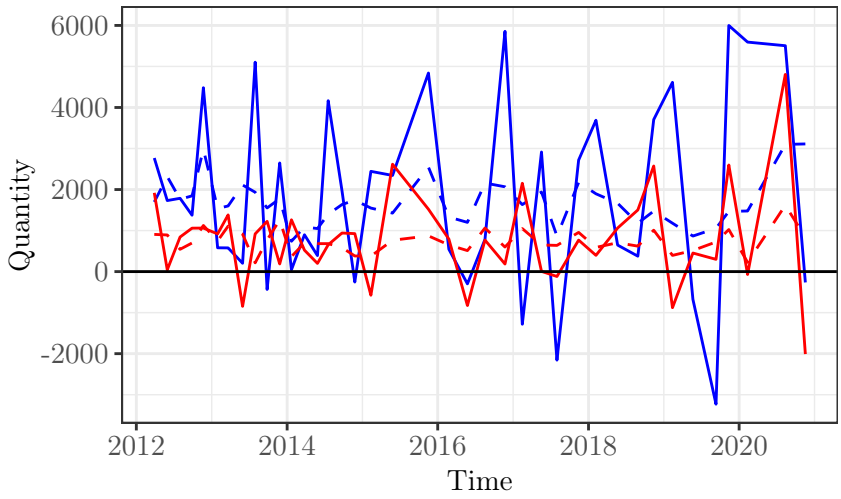


Figure 9: *Blue lines indicate 4-room apartments and red lines indicate 5-room apartments. Solid lines indicate the difference between previous oversubscription and current demand, $dd_t - os_{t-1}$. Dashed lines indicate the supply of apartments.*

by dividing by the total supply of apartments of that type, and use $lmis_t/ss_t$ as our measure of mismatch. By taking the mean of $lmis_t$ for 4-room and 5-room developments in our time period, we find that at least 6.4% of 4-room apartments and 10.2% of 5-room apartments are allocated to households that would have preferred a different apartment under this measure. Given that this measure underestimates actual levels of mismatch, the potential to improve allocative efficiency is large.

7 Discussion

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.[36] However, this notion of efficiency inherently fails to address certain societal objectives, such as avoiding inequality or racial segregation. That is, an environment where one "dictator" receives everything is considered efficient by this notion. If a government wishes to address those objectives, it runs the risk of distorting the market. This can potentially lead to a worse outcome than if the government had done nothing. In particular, when market forces cannot ensure demand is met by supply, the government struggles to accurately measure household preferences. Indeed, many attempts to create centralized markets in the past have failed.

Through the BTO program, the Singaporean government has treated concerns regarding racial and socioeconomic inequality, while also incentivizing truthful reporting. These concerns have often

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

failed to be addressed by private markets in other cities. For instance, the Singaporean government wishes to ensure that adequate housing is affordable for all families. In a city-state like Singapore, where land is a scarce commodity, leaving housing to market forces has historically failed to ensure affordability. Indeed, in many large modern cities such as New York City or Hong Kong, housing costs have skyrocketed in recent decades. In practice, this exacerbates the effects of inequality, dividing the rich and the poor. Those who seek jobs or amenities provided by the city are often forced to commute long hours or settle for cramped living quarters. While we will abstract from these concerns in our model, they provide a motivation for centralization of the market.

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 cost. This model suggests that the gains from doing so can be quite large. Indeed, the gains are likely bigger than a naïve estimate of household preferences would suggest. This follows from a household incentive to not apply sincerely when supply is exogenous. Current mechanism design setups generally focus on allocating objects that arrive exogenously. Where supply can be adjusted—such as housing, transportation, and foodstuffs—the conclusions for optimal design differ starkly from settings where goods arrive randomly.

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 (2023). "Choice, Welfare, and Market Design". In: *mimeo*.

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.

Cason, Timothy N and Charles R Plott (2014). "Misconceptions and game form recognition: Challenges to theories of revealed preference and framing". In: *Journal of Political Economy* 122.6, pp. 1235–1270.

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, and Luther Yap (2023). "The dynamic allocation of public housing: Policy and spillovers". 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 Appendix—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:

$$\begin{aligned} W_B(1,0) &= \sum_{i=0}^{\infty} (1/2 \cdot 1/2)^i \\ &= \frac{1}{1 - 1/4} \\ &= 4/3 \\ W_A(1,0) &= \sum_{i=0}^{\infty} 1/2(1/4)^i \\ &= 1/2\,[w_b(1,0)] \\ &= 2/3 \end{aligned}$$

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 $\gamma \geq W_B(1,0) - W_A(1,0) = 2/3$. 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 $\gamma \geq 2/3$. The mechanism never leaves the set of states $\{(1,0),(0,1)\}$, which implies that no other constraints are relevant. Then, when $\gamma \geq 2/3$, the first-best mechanism is implementable, and if $\gamma < 2/3$, no mechanism achieves 0 inefficiency. □

Proof of Lemma 3. 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. 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 $w_A^m(k, -(k-1))$, when $0 < k < m$, as the solution to the following system of equations using the transition matrix.

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

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.

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

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

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

The solution to the system of equations is given by $w_A^2(2,-1) = 5/2$ and $w_A^2(1,0) = 2$. Given the previous two values, $w_A^2(3,-2)$ can be computed and is equal to $10/3$. 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 $\gamma \in [5/2, 10/3]$. □

Proof of Proposition 2. 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. □

Proof of Lemma 6.

1. To begin, Lemma 5 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, at least one steady state exists by standard results in dynamics.

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 cannot be optimal due to the symmetry of the problem. Suppose a mechanism generated two steady states that were not queue symmetric, which yielded different efficiencies. 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 not optimal.

□

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

The following simple result will prove useful for the proof of Proposition 3.

Lemma 8. *The maximal equilibrium allocation inefficiency is $1/2$.*

Proof of Lemma 8. 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 $1/2$ 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 $1/2$. Last, since overall allocation inefficiency is a weighted average of the allocation inefficiency in each state, the overall allocation inefficiency must be at most $1/2$. □

Proof of Proposition 3. Lemma 5 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 inefficiency

is 1/2 and vacancy inefficiency is 1. Note that both inefficiency values are weakly larger than inefficiency in state $(1,0)$ by Lemmas 5 and 8. 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 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 must set $\gamma = W_B(1,0) - W_A(1,0)$.

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:

$$\frac{2(1+x^B(1,0))q - \alpha(q + x^B(1,0)(-2+q+2\Phi^B(2,-1)))}{2(2+q+x^B(1,0)q - 2\Phi^B(2,-1))}. \tag{5}$$

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

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

where the exact value of A is irrelevant since A^2 must be positive. The numerator is always negative, note that it can be rewritten as $B + \sqrt{4(2+\gamma(x^B(1,0)-3))x^B(1,0)(1-\Phi^B(2,-1)) + B^2}$, where $B = 4x^B(1,0)(1+x^B(1,0))(-2+\alpha+\alpha x^B(1,0))(-3+x^B(1,0)+\gamma(-1+x^B(1,0))(-1+\Phi^B(2,-1)) + 2\Phi^B(2,-1)$. However, $\gamma \leq 2/3$ implies that $4(2+\gamma(x^B(1,0)-3))x^B(1,0)(1-\Phi^B(2,-1)) \geq 0$ and therefore $\sqrt{4(2+\gamma(x^B(1,0)-3))x^B(1,0)(1-\Phi^B(2,-1)) + B^2} \geq -|B|$. 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:

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

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

Lemma 9. *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 9 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. These will prove useful for determining the region in which μ_q is optimal.

Wait time computations for two-state mechanism:

$$w_A(1,0)=1/2\cdot q[1+w_A(1,0)]+1/2\cdot q[1+w_A(2,-1)]+(1-q)/4\cdot[1+w_A(1,0)]$$
$$w_A(2,-1)=1/2[1+w_A(1,0)]$$

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:

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

□

Lemma 10. *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 10. 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 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 3 shows that either μ_q or μ_p is optimal. We proceed by comparing the inefficiencies generated by the two mechanisms. Lemma 10 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:

$$\frac{\alpha}{2}\frac{2-3\gamma}{7(2-\gamma)}(1-\alpha)\frac{2-3\gamma}{7(2-\gamma)} = (1-\alpha/2)\frac{2-3\gamma}{7(2-\gamma)} \tag{6}$$

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

Corollary 3. *The threshold which dictates whether μ_p or μ_q is optimal, $\frac{2-3\gamma}{8-5\gamma}$, is decreasing in γ.*

Proof of Corollary 3. We take the derivative of the threshold in Theorem 1 with respect to γ.

$$\frac{\partial \frac{2-3\gamma}{8-5\gamma}}{\partial \gamma} = \frac{-14}{(8-5\gamma)^2} < 0$$

□

Proof of Corollary 1. A computation similar to the above implies that the government prefers μ_q when $2(\gamma)^2 - 9\gamma + 2 < 0$ or when $\gamma > \frac{9-\sqrt{65}}{4} \approx .234$. □

Proof of Proposition 4. We begin by computing the difference in the expected wait times under the implementation of the first-best mechanism.

$$\begin{aligned}
W_B(1,0) &= p \cdot 1/2 + (p \cdot 1/2)^2 + \dots \\
&= \frac{1}{1-p/2} \\
W_A(1,0) &= 1/2 + 1/2\frac{1-p}{2} + 1/2\frac{1-p}{2}\frac{p}{2} + \dots \\
&= 1/2\frac{3-2p}{2-p} \\
W_B(1,0) - W_A(1,0) &= \frac{1+2p}{2(2-p)}
\end{aligned}$$

Lemma 2 then implies that this mechanism is an equilibrium only if γ bounds the differences in expected wait times. □

Proof of Lemma 7. The derivative of equation 4 with respect to p is given by:

$$\begin{aligned}\frac{\partial[w_B(1,0) - w_A(1,0)]}{\partial p} &= \frac{2(2(2-p)) - (1+2p)2(-1)}{4(2-p)^2} \\ &= \frac{5}{2(2-p)^2} > 0\end{aligned}$$

Then, since the derivative is positive, as p increases, so must γ in order to incentivize households to apply sincerely. □

Lemma 11. *When $N = 2$, the first-best can be implemented iff $\gamma \geq \frac{16}{17}$.*

Proof of Lemma 11. Wait times under the first-best mechanism, when $N = 2$ can be described according to the following system of equations

$$\begin{aligned}w_A(2,0) &= 1/2 \cdot 2/3(1 + w_A(2,0)) + 1/2 \cdot 1/2(1 + w_A(1,1)) \\ w_A(1,1) &= 1/2[1/2 \cdot 1/2(1 + w_A(1,1)) + 1/2(1 + w_A(2,0))] + 1/2 \cdot 1/2(1 + w_A(1,1))\end{aligned}$$

The solution to the above system of equations is $w_A(2,0) = 25/17, w_A(1,1) = 27/17$, the remainder of the proof follows directly from surrounding arguments. □

Proof of Corollary 2. Referring back to Proposition 1, under $N = 1$ the first-best could only be implemented when $\gamma \geq 2/3$. The threshold, 16/17 from Lemma 11 is higher, and therefore imposes a stricter condition on γ. □

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:

$$V(\mu) = \min_{\Phi^A(1,0) \in \Delta\{0,1\}^2} \alpha m + (1-\alpha)v(\mu) \tag{7}$$

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:

$$w_A(1,0) = 1/4[2q/3(2 + w_A(2,-1)) + (1-q)1/3(2 + w_A(1,0))] + 1/2 \cdot q/2(2 + w_A(1,0))$$
$$w_A(2,-1) = 1/4 \cdot 1/2(2 + w_A(2,-1)) + 1/2 \cdot 1/3(2 + w_A(1,0))$$

Algebra yields the following solutions with respect to q

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

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

$$w_A(1,0) - w_B(1,0) = 1/6\frac{-1725 + 2357q + 62q^2}{-231 + 50q}$$
$$w_A(2,-1) - w_B(2,-1) = \frac{453 - 142q}{1386 - 300q}$$

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 $\frac{-453+213/62(-833+\sqrt{753905})}{6(-231+75/62(-833+\sqrt{753905}))} \approx 0.294$. □

B Queue Hopping

Under the real BTO mechanism, households can freely switch queues between application cycles. In this section, we show that the ability to do so does not change the optimality of the mechanisms presented in the main body of the paper. Under the mechanisms presented in the main body of our paper, no household wishes to change its queue at the beginning of a period. We will formally prove this 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. As a result, correct allocation is impossible when households switch, causing inefficiency in mechanisms that take advantage of swapping to be high. 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.[37]

[37] 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.

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 more care. Previously, once a household had entered a given queue, its actual type became irrelevant for both the household and the government. Since it could not switch, incoming households only cared about its selected queue, and the government could no longer influence that household's match. However, now household's can switch queues between periods, implying that tracking their type is important. Furthermore, households can carry beliefs regarding other household's types. This is important because their type informs their switching probabilities. In practice, this behavior seems unrealistic. While households might observe application rates, they will not track applications on a household level. We then make the following assumption of naïvete, households only observe the length of each queue, they do not observe 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 it remains optimal in the new setting, namely when $\gamma \geq 2/3$. 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 matches its type never wishes to change its queue. If it 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 will 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 5. *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. q, the probability with which an apartment of type B is built, 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, the incoming household in this case. Furthermore, in the event the household does not receive an apartment in the current period, it prefers state $(1,0)$ to state $(0,1)$.

It remains to prove that μ_q is optimal in the current environment. Proposition 3 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 $0 < k < 1/2$, 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 $1/2$ 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 $k < 1/2$ 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 $1/2 \leq k < 1$. 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. □

C Suboptimality of Markovian Mechanisms

While throughout this paper we focus on Markovian mechanisms, we note that doing so is with potential loss of generality. To provide intuition for why the optimal mechanism is not necessarily

Markovian, we construct a mechanism that improves upon the oversubscription mechanism developed in Section 5.2 when $N = 2$. However, while the optimal probabilities change, the structure of the optimal mechanism is similar to the mechanisms utilized throughout the main body of the paper.

We weaken the Markovian assumption and allow the government to condition X_t^A, C_t not only on s_t, r_t but also on s_{t-1}, r_{t-1}. That is, the government can "reward" households from the previous period that have waited. We focus on allocation in state $(1,1)$, where incentive constraints are lax for incoming households of either type. The government can then modify the allocation probabilities in order to better incentivize truthful reporting in states $(2,0)$ and $(0,2)$. Altering notation slightly, we use $C_\theta^n(s)$ to indicate the probability with which the government builds a type-A apartment in state (s) when the previous state and report were $(n, 2-n)$ and θ.

Again, previous reasoning implies that in order to implement the first-best, if the state is $(2,0)$ or $(0,2)$ the government must always build a type-A or type-B apartment respectively. However, in state $(1,1)$, the government may build either apartment type freely. We will suppress the state when indicating $C_T^n(1,1)$, since $C_T^n(2,0) = 1$ and $C_T^n(0,2) = 0$. Utilizing our above methodology, we proceed by computing the expected waiting times for a household in each queue. We will slightly alter our notation to accommodate the new conditioning of w. Let $w_{(r,T)}^n(s)$ indicate the expected waiting time for a household in queue T when the previous report and state were $r, (n, 2-n)$ and the current state is s. Similarly, $W_{(r,T)}^n(s)$ indicates waiting times for the incoming household.

Note that we must have $C_A^2(1,1) = C_B^2(1,1)$. In state $(2,-1)$ all incoming households are allocated to queue B. If there are different wait times depending upon their reported type, they are incentivized to misreport their type in state $(2,-1)$. A similar argument implies $C_A^0(1,1) = C_B^0(1,1)$. In the following equations, we drop unnecessary notation. Then, the following system of equations defines expected wait time for an individual already in queue A

$$
\begin{aligned}
w_{(r,A)}^n(2,0) &= 1/2 \cdot 2/3(1 + w(2,0)) + 1/2 \cdot 1/2(1 + w^2(1,1)) \\
w_{(r,A)}^n(1,1) &= 1/2[C_r^n(1,1)1/2 \cdot (1 + w_{(a,A)}^1(1,1)) + (1 - C_r^n)(1 + w(2,0))] \\
&\quad + 1/2 \cdot (1 - C_r^n)(1 + w_{(b,A)}^1(1,1))
\end{aligned}
$$

We compute the the following expected wait times for each state, and use the symmetry of the problem to deduce that $C_a^1 = 1 - C_b^1$ and $C_b^2 = C_a^0$.

$$
\begin{aligned}
w(2,0) &= \frac{76 - 37C_a^1 - 15C_b^2}{44 - 29C_a^1 + 9C_b^2} \\
w^2(1,1) &= \frac{100 - 31C_a^1 - 61C_b^2}{44 - 29C_a^1 + 9C_b^2} \\
w_b^1(1,1) &= \frac{3(12 + 11C_a^1 + C_b^2)}{44 - 29C_a^1 + 9C_b^2} \\
w_a^1(1,1) &= \frac{100 - 95C_a^1 + 3C_b^2}{44 - 29C_a^1 + 9C_b^2} \\
w^0(1,1) &= \frac{36 - 31C_a^1 + 67C_b^2}{44 - 29C_a^1 + 9C_b^2}
\end{aligned}
$$

The next step is to then compare expected wait times for individuals in different settings, we begin with a type-B household in state $(2,0)$. The household knows that the supplied apartment will always be type-A, and must decide if waiting is worthwhile.

$$
\begin{aligned}
W_B(2,0) &= 1 + w_b^2(1,1) \\
W_A(2,0) &= 2/3(1 + w(2,0)) \\
W_B(2,0) - W_A(2,0) &= 1/3 + w_b^2(1,1) - w(2,0) \\
\implies \gamma &\geq \frac{16(C_a^1 - 5C_b^2)}{-44 + 29C_a^1 - 9C_b^2}
\end{aligned}
$$

An identical restriction can be computed in state $(0,2)$ by the symmetry of the problem.

Last, we compute the differences in wait times in state $(1,1)$ dependent upon the previous period's state and report.

When the previous state and report are s and r

$$
\begin{aligned}
W_{(r,B)}^n(1,1) &= C_r^n(1 + w(2,0)) + (1 - C_r^n)1/2(1 + w_A^1(1,1)) \\
W_{(r,A)}^n(1,1) &= C_r^n/2(1 + w_A^1(1,1)) + (1 - C_r^n)(1 + w(2,0))
\end{aligned}
$$

In effect, the wait times are the same except C_r^n is replaced with $(1 - C_r^n)$, hence why setting both equal to $1/2$ was previously optimal. In this instance, since this constraint in state $(1,1)$ is lax, changing these values may weaken constraints in states $(2,0)$ and $(0,2)$.

This generates the following differences in wait times

$$W^2_{(r,B)}(1,1) - W^2_{(r,A)}(1,1) = \frac{4(C^1_A + 3(-4 + C^2_B))(-1 + 2C^2_B)}{44 - 29C^1_A + 9C^2_B}$$
$$W^1_{(r,B)}(1,1) - W^1_{(r,A)}(1,1) = \frac{(4(-1 + 2C^1_A)(C^1_A + 3(-4 + C^2_B))}{44 - 29C^1_A + 9C^2_B)}$$

Then, we proceed by minimizing these differences in wait times through selection of C. Rudimentary optimization yields $C^1_A = 0.28 = C^2_B$ and $\gamma \geq 0.48$. This is a substantial improvement upon the minimal γ for $N = 2$.

D Labelling

In this section, we consider the impact of relabelling one of the apartment types on the set of achievable outcomes. In particular, suppose that apartment type-B was relabelled as two separate types $B1$ and $B2$. All preferences are maintained, that is if a household is of type B, then it gains h in utility from receiving either $B1$ or $B2$. Similarly, a type-A household gain l in utility from receiving either $B1$ or $B2$. The government can choose any of the three housing types when building an apartment.

Consider what happens if the government implements the previous first-best mechanism, replacing any instance of a type-B apartment with apartment type $B1$. Similarly, any household that would have previously applied for a type-B apartment instead applies for apartment type $B1$. Households still always wish to apply sincerely, the existence of an extra housing type that will never be constructed provides no incentive to deviate. Then, the previous strategy profile remains an equilibrium. Furthermore, suppose the mechanism involved some form of randomization between the two different housing types. Then, with positive probability, an apartment will lie vacant, implying that the first-best cannot be achieved.

Lemma 12. *The first-best in the standard two type case can be achieved if and only if it can also be achieved in the relabelling case.*

E M 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. It is simple to see that households wish to apply sincerely here. A household that did not receive an apartment in the previous period never has incentive to misreport because it knows an apartment of its 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. Last, 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, in doing so they must risk positive probability of vacancies. Last, this argument holds for all $|\Theta| > 2$, the above arguments did not utilize the fact that $|\Theta| = 3$.

Lemma 13. *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.*