American Economic Review 2022, 112(12): 3876–3910
https://doi.org/10.1257/aer.20201111

Dynamic Matching in Overloaded Waiting Lists[†]

By Jacob D. Leshno*

This paper introduces a stylized model to capture distinctive features of waiting list allocation mechanisms. First, agents choose among items with associated expected wait times. Waiting times serve a similar role to that of monetary prices in directing agents' choices and rationing items. Second, the expected wait for an item is endogenously determined and randomly fluctuates over time. We evaluate welfare under these endogenously determined waiting times and find that waiting time fluctuations lead to misallocation and welfare loss. A simple randomized assignment policy can reduce misallocation and increase welfare. (*JEL* C78, D61, D82, D83)

From nursery schools to nursing homes, waiting lists are a common tool for allocating scarce goods that arrive stochastically over time to agents that accumulate over time.[1] In such settings it is impossible to allocate all items at once, and waiting lists are used to dynamically allocate items over time. We focus on overloaded waiting lists, where items are scarce and many agents are waiting for any arriving item. Examples include waiting lists for public housing, organ transplants, nursing homes, and daycare centers.[2] Given the high demand, all items can be readily assigned. But

* University of Chicago Booth School of Business (email: jacob.leshno@ChicagoBooth.edu). Sylvain Chassang was the coeditor for this article. This paper was initially circulated in 2012 under the title "Dynamic Matching in Overloaded Systems." I thank the editor and three anonymous referees, Nikhil Agarwal, Sam Asher, Itai Ashlagi, Eduardo Azevedo, Omer Ben-Neria, Dirk Bergemann, Shai Bernstein, Simon Board, Emily Breza, Eric Budish, Rene Caldentey, Gabriel Carroll, Lorenzo Casaburi, Yeon-Koo Che, Rebecca Diamond, David Gamarnik, Josh Gottlieb, Refael Hassin, John Hatfield, Moshe Haviv, Gur Huberman, Stephanie Hurder, Scott Kominers, Greg Lewis, Ohad Noy-Feldheim, Parag Pathak, David Parkes, Assaf Romm, Philipp Strack, Seth Stephens-Davidowitz, Omer Tamuz, and many others for helpful discussions, suggestions, and advice. I am grateful to Al Roth, Ariel Pakes, Susan Athey, and Drew Fudenberg for their comments and guidance. I thank Ethan Che and Anand Shah for providing invaluable research support. This work is supported by the Robert H. Topel Endowment at the University of Chicago Booth School of Business. This work was inspired by the efforts of Dr. Stéphane Lemire to improve nursing home assignments in Quebec, Canada.

[†] Go to https://doi.org/10.1257/aer.20201111 to visit the article page for additional materials and author disclosure statement.

[1] Waiting lists are common in practice. Some examples include publicly provided medical services (Lindsay and Feigenbaum 1984; Martin and Smith 1999), organs for transplant (Kessler and Roth 2014), and public housing (Kaplan 1984). The inspiration for this work came from problems arising from allocating senior citizens to publicly provided nursing homes in Quebec, Canada. Barzel (1974) describes how waiting time can be used to ration goods when monetary prices are fixed to be below the market price. Waiting lists are also used by for-profit firms; for example, many National Football League teams hold waiting lists for season tickets (Forbes 2007).

[2] The rate at which applicants join the waiting list often exceeds the rate at which items arrive, leading to long and growing waiting lists. For example, the Chicago Housing Authority runs a lottery to determine who can join its long waiting list, because the number of potential applicants is too large and the median applicant on the waiting list drops out without being assigned (Chicago Housing Authority 2016). See also a similar assumption in Kaplan

when items are heterogeneous and agents have heterogeneous preferences,[3] to maximize welfare the assignment needs to efficiently match agents to items.

This paper introduces a new stylized model to capture distinctive features of waiting list allocation mechanisms. First, agents choose among items with associated expected wait times. Waiting times serve a similar role to that of monetary prices in directing agents' choices and rationing items. Second, the expected wait for an item is endogenously determined and randomly fluctuates over time. In particular, different agents make their choice at different times and may face different menus of options.

In the model, two kinds of items, *A* and *B*, stochastically arrive over time and are assigned to agents as they arrive. Agents join the overloaded waiting list according to an exogenous arrival process.[4] Agents incur waiting costs until they are assigned, regardless of whether or not they participate in the waiting list mechanism.[5] All agents incur the same waiting costs but differ in their preferences over the two items. We say that agents are mismatched if they are assigned to their less-preferred item. Because waiting is costly, agents will choose to wait for their preferred item only if the required expected wait is sufficiently low.

As an illustration, consider two agents, α and β, where α prefers an *A* item and β prefers a *B* item. Suppose that in period 1 a *B* item arrives, and in period *t* an *A* item arrives. Consider the two possible assignments: either both agents get their preferred item and α waits, or both agents get their mismatched item and β waits. Total waiting costs are equal across the two assignments, and therefore the former assignment maximizes welfare. But if *t* is large then α prefers the inefficient latter assignment because the costly wait is transferred to β. Thus, a mechanism that allows α to choose between the two assignments can generate a socially inefficient assignment.

As illustrated by the example, total waiting time cost are constant across assignments in the overloaded waiting list model. Agents join the waiting list exogenously, and an agent's costs of waiting can only be reduced by assignment of an item. Each arriving item reduces the waiting cost of one assigned agent, while other agents remain waiting. Thus, total waiting costs are constant across assignments that assign all items, and welfare is entirely determined by the fraction of mismatched agents.

Expected waiting times serve a similar role to monetary prices in guiding the allocation. Given a sufficiently long expected wait for *A* items, agents prefer an earlier assignment to a *B* item (even if it is their less-preferred item) because of the reduced waiting costs.[6] Offering equal waiting times for both items induces all agents to choose their preferred item. Alternatively, items can be rationed by

(1986). As of December 2016, more than 90,000 patients in the United States are waiting for a kidney transplant, with 22 people a day dying while waiting for transplant (UNOS 2016). Private conversations with daycare centers and the administration of the public nursing home system in Quebec indicate that long waiting lists are common there as well.

[3] Patients differ in their preferences over organs for transplant based on immunocompatibility and proper organ size. Applicants to public housing, daycare centers, and nursing homes differ in their geographical preferences over locations.

[4] For example, applicants sign up to join the public housing waiting list when they become eligible.

[5] For example, public housing applicants incur a waiting cost from having to pay higher private market rent. Applicants pay this waiting cost until receiving subsidized public housing regardless of whether or not they registered to the waiting list.

[6] The reduced waiting costs are not eliminated, but transferred to other agents who remain waiting.

offering a sufficiently longer expected wait for overdemanded items. Like monetary prices, waiting times can communicate to agents whether items are overdemanded.

But in contrast to monetary prices, expected waiting times are endogenously determined and fluctuate over time. If the mechanism already promised future arriving items to some agents, the mechanism is forced to offer a longer wait to the next agent. Moreover, expected waiting times vary with the current state of the system, which randomly fluctuates because of the random arrival of items and the unknown preferences of approached agents. Even if A and B items are equally likely to arrive and agents are equally likely to prefer each item, it is possible for many B items to arrive in succession while the mechanism approaches many agents who prefer A. In such case, some of these agents must be offered long waiting times for an A item, inducing some agents to mismatch.

To investigate welfare under endogenously determined and fluctuating waiting times, we analyze a common waiting list mechanism: the waiting list with declines.[7] When an item arrives, this mechanism approaches an agent and offers a choice between taking the current item or declining the item and keeping their position. If the item is declined, the mechanism immediately approaches the next agent in line. The mechanism informs agents of their position, giving agents all available information about their expected wait.[8]

The waiting-list-with-declines mechanism embodies the two features listed above. Agents can choose between immediate assignment to the current item or an expected wait and assignment to the other item, and the expected wait for the other item is endogenously determined by the number of agents ahead of them. If agents face an appropriate expected wait, their choices will be socially efficient. But expected waiting times randomly fluctuate because of the randomness in the item arrival process and the random composition of agents in the waiting list. Moreover, expected waiting times fluctuate even if the waiting list holds many agents. Analysis of the waiting-list-with-declines mechanism allows us to quantify welfare under the endogenously determined waiting times, explain how waiting time fluctuations cause welfare loss, and suggest alternative designs.

To analyze the waiting list with declines, it is useful to represent it as a buffer-queue mechanism. Under the buffer-queue representation, an agent who declines an item to wait for his preferred item is said to join a buffer queue for the preferred item. Two buffer queues, one for each kind of item, hold agents who declined a mismatched item and are waiting to be assigned their preferred item. The waiting list with declines is equivalent to a buffer queue mechanism with the first-come-first-served (FCFS) queuing policy.

The buffer-queue representation allows us to determine how often agents choose a mismatched item and calculate welfare. The system's dynamics are captured by a tractable Markov chain whose states are the number of agents in each buffer queue. The number of agents in a buffer queue can be thought of as the current imbalance between the demand from approached agents and the supply of arriving items. In

[7] This mechanism is a simplification of common mechanisms. Variants of this mechanism or equivalent formulations of it are commonly used for organ allocation (UNOS 2014) and allocation of spots in daycare centers. Thakral (2016) provides arguments in favor of this mechanism.

[8] This is a stylized modeling assumption. In many applications, agents are only given partial information about the current state of the system and their expected wait. See Section III and the comments below.

states of higher imbalance (many agents accumulating in the buffer queue), agents need to wait longer for their preferred item; under FCFS, the expected wait for an agent who joins the buffer queue increases linearly with the number of agents already in the buffer queue ahead of him.

Even if demand from agents and supply of items are balanced on average, the random arrivals of items and random preferences of approached agents cause the state to randomly fluctuate. In states of high imbalance, the mechanism offers a long expected wait that induces agents to prefer an immediate mismatched item. We calculate the misallocation rate by calculating the stationary distribution of the fluctuating state and the probability that imbalance becomes sufficiently high for the offered expected wait to be unacceptable. A similar approach can be applied to analyze other waiting list mechanisms as well.[9]

Given the welfare loss, we use the model to consider alternative designs. The planner would like to provide agents with information about expected waiting times to convey useful long-run information (e.g., all agents prefer A items, and therefore A items need to be rationed), but shield agents from random fluctuations (e.g., temporary shortage of A items). We derive alternative queuing policies and information designs that reduce fluctuations in expected waiting times, and thus reduce misallocation. These alternative designs can increase welfare without imposing additional delays or changing the way the mechanism approaches agents.[10]

Better information design can reduce misallocation. If the mechanism can limit information about the current state, it can improve welfare by only giving agents a binary signal about their position: suggesting whether the agent should wait for the preferred item or take the immediate mismatch. Agents make their choice based on the average expected wait across states, allowing the mechanism to signal agents to wait and endure longer wait in some states of high imbalance if on average the expected wait is acceptable because of the low wait in states of low imbalance. In other words, by hiding information, the mechanism can offer the same acceptable wait even if the state fluctuates. In practice many waiting list mechanisms provide infrequently updated information to applicants, limiting their response to fluctuations.[11]

Without hiding information, the mechanism can reduce misallocation by using a randomized queuing policy.[12] The FCFS queuing policy is wasteful in that it offers

[9] Examples include the analysis of Caldentey, Kaplan, and Weiss (2009) and Adan and Weiss (2012), and the subsequent work of Baccara, Lee, and Yariv (2020). Online Appendix C analyzes the disjoint-queues mechanism which holds a separate queue for each item and asks agents to choose a single queue when they join the waiting list. The analysis uses a similar Markov chain that tracks imbalance between supply of items and demand from agents. Similarly, misallocation occurs when the imbalance is too large, inducing agents to choose to join the shorter queue regardless of their preferred item.

[10] A different approach is to change the assignment process to avoid the underlying fluctuations. Assigning items in large batches can reduce the randomness due to the item arrival process, but requires agents to incur additional waiting costs while items accumulate. A lottery that offers agents a chance to get the current item or leave unassigned can eliminate dynamic considerations, but removes multiple agents from the waiting list for every assigned item and therefore requires that multiple agents join for every assigned item. Such approaches also eliminate the useful information provided by endogenously determined waiting times. Evaluation of such approaches is left for future work.

[11] While such policies can reduce misallocation (which is the focus of the model), providing current information to applicants is important for reasons that are beyond the scope of our model. See the discussion in Section III.

[12] Randomization is used in practice both implicitly and explicitly. For example, priority for liver transplants is determined by a lab test score (Wiesner et al. 2003). Because the score contains some random variation, an organ will be randomly assigned to one of the sickest patients. Arnosti and Shi (2017) and van Dijk (2019) analyze

a very short expected wait to agents approached in states of low imbalance, but an unacceptably long wait in states of high imbalance. By increasing the expected wait in low-imbalance states, a randomized queuing policy can decrease the expected wait in high-imbalance states. By having an expected wait that varies less with the state, the mechanism can offer an expected wait that is indicative of long-run information in more states, reducing misallocation due to random fluctuations.

A practical recommendation is the simple service-in-random order (SIRO) queuing policy. A SIRO buffer-queue mechanism has a simple description: agents who decline an item are allowed to join a priority pool for their preferred item, and agents in each priority pool have an equal probability of receiving an arriving item. We characterize the SIRO buffer-queue mechanism as the robustly optimal mechanism. This simple randomization does not fully equalize the expected wait across states, but it lessens the expected wait fluctuations and therefore reduces the misallocation probability and achieves higher welfare in equilibrium than FCFS.

In summary, this paper offers two messages for the practical design of allocation through waiting lists. First, although many public-housing authorities have waiting list policies that discourage applicants from declining items, the analysis suggests agents should be encouraged to decline mismatched items. When the system is overloaded, an agent who declines a mismatched item allows the system to search further and assign the item to a matching agent. Furthermore, such an agent reduces the waiting costs of others by allowing them to be assigned before him. Second, equalizing the expected wait agents face when making their choice can improve welfare. This can be achieved by the SIRO buffer-queue mechanism or by partial information mechanisms. Both are practical mechanisms that offer agents more equal options at the time they make their choice, and thus reduce misallocation and improve welfare.

Related Literature

A growing literature studies dynamic assignment markets. Public housing is a prominent example of assignment through waiting lists, studied by Kaplan in a series of papers (1984, 1986, 1987, 1988). Su and Zenios (2004, 2005, 2006) study the assignment of transplant organs through waiting lists and suggest mechanisms that induce agents to accept marginal kidneys to reduce wastage of organs. By contrast, our findings suggest that if the waiting list is long, patients should be induced to decline organs that can be better assigned to other agents. Bloch and Cantala (2017) analyze dynamic assignment to agents with idiosyncratic preferences and find an FCFS policy maximizes welfare. Schummer (2016) follows up on the current paper and derives conditions under which welfare improves when agents are induced to decline items. Thakral (2016) argues theoretically and empirically that waiting list mechanisms should allow agents to decline items without penalty. Arnosti and Shi (2017) analyze trade-offs between efficiency and targeting in dynamic assignments.

Subsequent to the initial version of this paper, a growing empirical literature evaluates the allocative efficiency of waiting list mechanisms. Agarwal et al. (2019) empirically study the allocation of kidneys, and van Dijk (2019) and Waldinger

mechanisms that allocate housing via explicit lotteries. Verdier and Reeling (2022) study the allocation of hunting licenses through a mechanism that uses lotteries for tie-breaking.

(2018) study the allocation of public housing. Verdier and Reeling (2022) study the dynamic allocation of hunting licenses.

The subsequent work of Baccara, Lee, and Yariv (2020) analyzes a dynamic two-sided matching market using a similar Markov chain to capture fluctuating imbalances. They find that agents would wait longer than is socially efficient, and that welfare can be improved by side payments or batching. Doval and Szentes (2018) analyze a dynamic two-sided matching market and characterize when agents will be more or less impatient than socially optimal. Doval (2015) develops a notion of stability in dynamic environments.

Ünver (2010); Akbarpour, Li, and Gharan (2020); Anderson et al. (2017); Ashlagi et al. (2019); and Das et al. (2015) explore the related issue of thickness of dynamic markets. This literature finds that a myopic policy can be optimal under some assignment feasibility constraints.

Our model is connected to but differs from standard queuing models. Starting with Naor (1969), a large literature considers waiting costs in strategic queuing settings; see Hassin and Haviv (2003) for a survey. In contrast, this paper analyzes the matching between agents and items. From a technical perspective, our stochastic model is closer to the FCFS infinite bipartite matching problem studied by Caldentey, Kaplan, and Weiss (2009). Our analysis relies heavily on their Markovian representation. Adan and Weiss (2012) and Adan et al. (2018) provide expression for calculating performance metrics, but conjecture that calculating welfare from these expressions is computationally hard.

The current paper is related to the literature on dynamic mechanism design (see Bergemann and Said 2011 for a survey), but differs from it in that we do not allow transfers. While expected waiting times serve as prices, the mechanism can offer only expected waiting times that can be feasibly generated by the stochastic dynamics.[13] Section V uses ideas from the literature on robust mechanism design (Bergemann and Morris 2005).

Finally, our results demonstrate how fluctuations adversely affect the efficiency of resource allocation. De Vany (1976) and Carlton (1977, 1978) study how demand fluctuations affect firms and market behavior. Asker, Collard-Wexler, and De Loecker (2014) provide empirical evidence that fluctuations cause misallocation and lower productivity.

Organization of the Paper

Section I introduces the model and shows that in an overloaded waiting list, welfare is maximized by maximizing the value of assigned items. Section II analyzes the waiting-list-with-declines mechanism and calculates the welfare loss from fluctuations. Section III shows that information design can be used to reduce welfare loss. Section IV gives the technical intuition for the results by providing a buffer-queue representation for the waiting-list-with-declines mechanism. Section V leverages the technical results to the design of queuing policies that help control expected wait fluctuations and reduce welfare loss. It derives the practical SIRO policy. Section VI

[13] The dynamic mechanisms we derive have features similar to Levin (2003). In both problems, the mechanism has a finite stock of value (the value of future arrivals) that is used for generating good incentives for agents.

uses simulations to assess concerns of realized envy and presents heuristics that can mitigate such concerns. Section VII concludes.

Appendix A provides the details regarding the Markov chain used in our analysis. Online Appendix B discusses nonlinear waiting costs. Online Appendix C analyzes the disjoint-queues mechanism and finds similar welfare loss from random fluctuation. Online Appendix D contains omitted proofs.

I. Model of Dynamic Matching in Waiting Lists

Each period $t \geq 1$ begins with the arrival of an item x_t and ends when the item is assigned to an agent. The item is of kind A with probability p_A and of kind B with probability $p_B = 1 - p_A$, independently across periods.[14]

Agents arrive and enroll in the waiting list according to an exogenous agent arrival process, which is discussed below. Agents are infinitely lived, risk neutral, and incur a common linear waiting cost $c > 0$ per period until they are assigned. We assume that registering for the waiting list does not change the agent's waiting costs.[15]

Agents are of two types: agents of type α prefer A items and agents of type β prefer B items. We refer to the agent's nonpreferred item as a *mismatched item*. Given an item, we refer to agents of the type that prefers the item as *matching*, and other agents are *mismatched*. Agent types are private information.

Once assigned, agents stop paying the waiting cost and receive a value $v > 0$ if they are assigned their preferred item, or a value of 0 if they are assigned their mismatched item. That is, the utility of an α agent who is assigned after waiting w periods is $v - c \times w$ if he is assigned an A item, or $-c \times w$ if assigned a B item. We assume agents break indifferences in favor of their preferred item. Because of the reduction in waiting costs, agents prefer receiving a mismatched item to never being assigned.

An *assignment* is $\mu : \{t \geq 1\} \rightarrow \mathcal{I}$, where $\mathcal{I}$ is the set of agents and $\mu(t) \in \mathcal{I}$ is the agent assigned the item x_t. We say the item arriving in period t is *misallocated* if $\mu(t)$ is a mismatched agent for the item x_t. Assignments are final, and assigned agents leave the system.[16]

The mechanism dynamically assigns items as they arrive. At the beginning of each period, the mechanism learns which item arrived and may have information about the preferences of agents approached in previous periods. We say that an agent is unapproached if the agent enrolled in the waiting list, but had no other interaction with the mechanism. The mechanism may assign the item given its current information or sequentially approach new unapproached agents from the waiting list to learn their preferences. We assume all such unapproached agents appear interchangeable

[14] The model and analysis would remain essentially unchanged if item arrivals follow a Poisson processes with rates λ_A, λ_B, as we can normalize time so that $\lambda_A + \lambda_B = 1$ and set $p_A = \lambda_A$. Kaplan (1986) argues the arrival process of public housing apartments should be modeled as a Poisson process.

[15] In particular, the agent's outside option of opting out of the waiting list is equivalent to never getting assigned and entails a utility of $-\infty$.

[16] In some applications, misallocated agents may eventually be able to trade their items. Such considerations are left out of the analysis in this paper, but can be partially captured by setting the utility of a mismatched agent to be $v' - c \times w$, where $0 \leq v' < v$ is the value of getting a mismatched item and trading it later (even if trade is possible, receiving a mismatched item is strictly worse than receiving the preferred item, because the agent spends some time assigned to the mismatched item before trading).

to the mechanism.[17] The probability that an unapproached agent is of type α is p_α, independently across agents. We denote $p_\beta = 1 - p_\alpha$. We say the system is *balanced* if $p_A = p_\alpha$, and in that case, denote $p = p_A = p_\alpha$.

The mechanisms analyzed in this paper[18] approach agents only when seeking to assign an arriving item. If all agents in the waiting lists are interchangeable and there is always another agent in the waiting list when the mechanism seeks to approach one, then these mechanisms are invariant to agents' arrival times. This observation allows us to simplify the model and abstract away from the details of the arrival process.

We say that the agent arrival process generates an overloaded waiting list for mechanism $\mathcal{M}$ if there exists an $M > 0$ such that (i) at any point in time, mechanism $\mathcal{M}$ holds at most M approached agents that have not been assigned items, and (ii) the waiting list always contains at least M agents. That is, if the waiting list is overloaded for mechanism $\mathcal{M}$, mechanism $\mathcal{M}$ never runs out of new agents to approach. For the mechanisms[19] and applications considered in this paper, it is highly unlikely that the mechanism runs out of new agents to approach.[20] By neglecting that possibility and assuming that the waiting list is overloaded for the mechanisms considered, the analysis can abstract away from further details of the agent arrival process.

The goal of the social planner is to allocate the limited supply of items to maximize total utility. Each assigned item makes two contributions to welfare: (i) agent's value of the item, and (ii) reduction in waiting costs. The following lemma shows that any assignment reduces total waiting costs by the same amount.[21] Intuitively, a public housing apartment generates a reduction of waiting costs equal to one month's rent reduction for each month the apartment is assigned. Any public housing assignment that immediately assigns all apartments as they arrive generates the same reduction in total rent paid by applicants. Since waiting costs are potentially unbounded due to the infinite time horizon, the lemma compares assignments up to an arbitrary finite time horizon.

LEMMA 1: *The difference between the total utility under assignments μ and μ' up to period T depends only on the number of misallocations under μ and μ' up to period T.*

[17] In particular, the time at which an agent enrolled in the waiting list is not correlated with the agent's preferences.

[18] That is, any mechanism equivalent to a buffer-queue mechanism (defined in Section IV).

[19] In contrast, for the disjoint-queues mechanism analyzed in online Appendix C, there is no agent arrival process that generates an overloaded waiting list. Because the disjoint-queues mechanism approaches each agent as they arrive, there are never any unapproached agents in the waiting list. Moreover, the agent arrival process can generate fluctuations under the disjoint-queues mechanism. Therefore, the analysis of the disjoint-queues mechanism cannot similarly abstract away from the specifics of the agent arrival process.

[20] Section IV bounds the number of approached agents in the system under buffer-queue mechanisms, implying that the waiting list is overloaded for a buffer-queue mechanism if the waiting list always contains at least M agents, where M is moderately small for plausible parameter values. Because in the applications considered items are scarce and there are likely to be many agents waiting for any arriving item, the overloaded waiting list assumption is likely to hold.

[21] The lemma can be generalized to allow a common discount rate, as in online Appendix B. In the presence of discounting, we define social welfare as the sum of agents' utilities evaluated at time 0.

By Lemma 1, the social planner can ignore waiting costs when comparing different assignments. Welfare is determined by the number of misallocations, as each misallocated item generates a value of 0 instead of v. Therefore, the social planner's objective is to minimize welfare loss from misallocation.

DEFINITION 1: *Given an assignment* μ, *let* $\xi_t \in \{0,1\}$ *be an indicator equal to 1 if the item* x_t *is misallocated under* μ. *The long-run misallocation rate* $\xi = \xi(\mu)$ *is given by* $\xi = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} \xi_t$.*We define welfare loss from misallocation* (WFL) *to be*

$$WFL = v \times \left(\xi - |p_A - p_\alpha|\right).$$

If $p_\alpha \neq p_A$, one of the items is overdemanded and $|p_A - p_\alpha|$ of agents must be assigned to their mismatched item.[22] A misallocation rate approximately equal to $|p_A - p_\alpha|$ can be obtained if agents are patient (Corollary 1) or if the mechanism has full information (Lemma 2). Thus, WFL can be interpreted as the additional loss due to the dynamic allocation problem.

As an illustration, we consider two simple mechanisms.

Sequential Assignment without Choice.—To highlight the importance of facilitating agent choice, consider *the sequential assignment* mechanism. Each period, the sequential assignment mechanism assigns the arriving item to an (arbitrary) unapproached agent without offering that agent a choice.[23] This mechanism induces a misallocation rate equal to $\xi^{SA} = p_A p_\beta + p_B p_\alpha$. If the system is balanced, the misallocation rate simplifies to $\xi^{SA} = 2p(1-p)$ and WFL is equal to $v \times 2p(1-p)$.

Full Information Mechanism.—To clarify the mechanism design challenge, consider a simple mechanism that has full information of agent preferences. When an item arrives, the *full information mechanism* searches the entire waiting list for a matching agent and assigns the item to a matching agent if there is one. The full information mechanism will mismatch agents only if there is no matching agent in the waiting list.

LEMMA 2: *Suppose the waiting list includes M agents at any point in time. The full information mechanism achieves a misallocation rate* ξ^{FI} *such that* $\lim_{M\to\infty} = \xi^{FI} = |p_A - p_\alpha|$.

II. The Waiting-List-with-Declines Mechanism

The waiting-list-with-declines mechanism holds a single ordered waiting list. Arriving items are offered to the first agent on the waiting list. When an agent is offered the current item, the agent can decline the item and keep his position in the

[22] Formally, if there is a constant that upper-bounds the number of agents in the waiting list at any time, $|p_A - p_\alpha|$ is the minimal expected misallocation rate among all allocations.

[23] Agents who decline the item are removed from the waiting list. If the waiting list is sufficiently long, even if an agent can decline an item and reenter the waiting list, the agent will not benefit from doing so.

waiting list. A declined item is immediately offered to the following agent. Note that the mechanism approaches a new agent only when it is trying to assign an arriving item that has been declined by all approached agents that are still in the system.

Agents know their position in the waiting list, which we denote by k. This assumption implies that agents can fully observe and react to the fluctuating state of the system. Section III explores partially informed agents.

Consider an α type agent who is offered a B item when he is in position k (the treatment of β agents who are offered an A item is symmetric). The agent faces a choice between taking B immediately or declining it and waiting for an A item. Let w_k denote the expected wait for an A item for an agent in position k. The α agent receives zero utility from taking the B item immediately and $v - c \times w_k$ from waiting for an A item.[24] Thus, the α agent prefers to wait for the preferred A item if the expected wait w_k is below $\bar{w} = v/c$.

The expected wait w_k serves a similar role to prices in guiding the allocation of items. An expected wait $w_k \leq \bar{w}$ induces α agents to wait for an A item. An expected wait $w_k > \bar{w}$ rations A items by inducing α agent to take the immediate B item. Similarly to monetary transfers in standard competitive equilibrium models, waiting costs can only be transferred between agents (as the total waiting costs are constant across assignments). In particular, if $p_\alpha = p_A$, it is socially inefficient for an α agent to take a B item, as the waiting costs w_k are transferred to other agents.[25]

If the planner could choose the expected wait offered to agents, the planner could implement the optimal assignment by offering an expected wait above $\bar{w}$ only when items need to be rationed. But the planner cannot directly choose the expected wait offered to agents. If a B item is offered to the agent in position k, it has been declined by agents in positions $1, \ldots, k-1$, who are also waiting for an A item. Because these k agents are assigned in an FCFS priority order, the expected wait for the agent in position k is the expected number of periods until k copies of A arrive, which is $w_k = k/p_A$. Different α agents face different expected waits for A depending on their position k when offered an item. If k is sufficiently large, an α agent prefers to take the immediate B item. Agent behavior is summarized in the following lemma.

LEMMA 3: *The waiting list with declines has a unique equilibrium, under which an α agent in position k declines a B to wait for an A item if and only if*[26] $k \leq K^A = \lfloor p_A \bar{w} \rfloor$. *Likewise, β agents wait for B items if and only if* $k \leq K^B = \lfloor p_B \bar{w} \rfloor$.

The waiting list with declines incurs misallocation and WFL because agents are offered a randomly fluctuating expected wait. Depending on the randomly evolving state of the system, some agents will be offered a higher expected wait for A while others may be offered a higher expected wait for B. As the mechanism accumulates α agents who declined a B, it approaches agents with higher k who are offered a higher expected wait w_k. Even if $p_\alpha = p_A$, the state of the system randomly evolves

[24] Note that an agent who declined a B once will prefer to decline all subsequent offers of B items to wait for an A, because past costs are sunk and the expected wait for an A can only decrease. Therefore, it is immaterial whether an agent who declined a B item will be offered B items again.

[25] If $p_\alpha > p_A$, the assignment must ration A items and assign $p_\alpha - p_A$ of α agents to a B, but it is socially inefficient to have more than a fraction $p_\alpha - p_A$ of α agents take a B item.

[26] We use the notation $\lfloor x \rfloor = \max\{n \in \mathbb{Z} | n \leq x\}$.

over time, and the expected wait offered to agents randomly fluctuates. When the mechanism randomly accumulates more than K^A agents waiting for an A item, it offers an expected wait that exceeds $\bar{w}$, inducing α agents to choose mismatched items.

THEOREM 1: *The waiting list with declines has a unique equilibrium, under which welfare loss when* $p_\alpha \neq p_A$ *is given by*

$$\begin{aligned} WFL^{WLWD} &= v\left(\xi^{WLWD} - |p_\alpha - p_A|\right) \\ &= v \times 2|p_\alpha - p_A|\left\{\left[\left(p_\alpha/p_A\right)^{K^A+1}\left(p_\beta/p_B\right)^{-(K^B+1)}\right]^{\operatorname{sgn}(p_\alpha - p_A)} - 1\right\}^{-1}, \end{aligned}$$

with $K^A = \lfloor p_A \bar{w} \rfloor$ *and* $K^B = \lfloor p_B \bar{w} \rfloor$. *When* $p_\alpha = p_A = p$, *welfare loss simplifies to*

$$WFL^{WLWD} = v\frac{2p(1-p)}{(1-p)K^A + pK^B + 1}.$$

The proof of Theorem 1 calculates WFL by solving for the stationary distribution in closed form. Section IV provides the intuition and technical analysis.

The two following corollaries show how WFL varies with agent preferences. First, misallocation decreases as the cost of waiting c decreases. When agents are more patient $\bar{w} = v/c$ is larger, and K^A, K^B are larger. In other words, expected waiting times can fluctuate within a larger range without exceeding $\bar{w}$ and causing misallocation.

COROLLARY 1: *As the cost of waiting* c *decreases,* $\lim_{c \to 0} \xi^{WLWD} = |p_A - p_\alpha|$ *and* $\lim_{c \to 0} WFL^{WLWD} = 0$.

Second, when the system is balanced (i.e., $p_\alpha = p_A$), WFL can be substantial even if mismatched items are undesirable. As v increases, agents are willing to wait longer for their preferred item, reducing the misallocation rate. However, as v increases, each misallocation is a greater loss. Taken together, these two countervailing effects roughly cancel each other, $\xi^{\text{WLWD}} \approx 1/\bar{w}$ and WFL is approximately equal to $v \times 1/\bar{w} \approx c$.

COROLLARY 2: *If the system is balanced, we have that* $\lim_{v \to \infty} WFL^{WLWD} \to c$. *If* $p_\alpha \neq p_A$, *we have that* $\lim_{v \to \infty} \xi^{WLWD} = |p_A - p_\alpha|$ *and* $\lim_{v \to \infty} WFL^{WLWD} \to 0$.

Figure 1 depicts the welfare loss under the waiting list with declines (labeled FCFS). When v is close to zero, the preferred item and the mismatched item are almost identical; little loss results from misallocation, and agents do not wait for their preferred item. As v increases, each misallocation becomes more costly, but agents are willing to wait for their preferred item at higher positions, reducing the misallocation rate. Discontinuity points correspond to values for which agents in some position are indifferent between waiting for their preferred item and taking an immediate mismatched item.

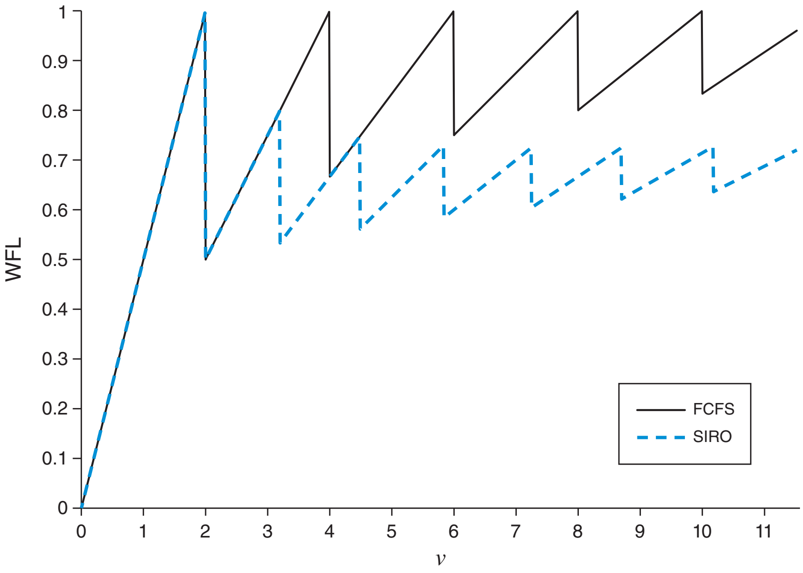

FIGURE 1

Note: Equilibrium WFL for $p_\alpha = p_A = 1/2$, $c = 1$, and varying values of v under the waiting list with declines (labeled FCFS) and the SIRO buffer-queue mechanism.

Figure 1 also provides a comparison with an alternative mechanism with lower welfare loss analyzed in Section V: the SIRO buffer queue. Under the SIRO buffer-queue mechanism, agents who decline an item join a priority pool for their preferred item, and agents in the pool have an equal probability of receiving an arriving item. By doing so, the expected wait offered to agents varies less with the agent's position, as shown in Figure 2. This allows the SIRO buffer-queue mechanism to offer more agents an expected wait that is below $\bar{w}$ and reduce misallocation and welfare loss.

III. Information Design

Providing information to agents about their expected wait is useful for rationing overdemanded items (e.g., when $p_\alpha \neq p_A$), but the previous analysis shows that revealing the fluctuating expected wait leads to misallocations. By hiding the agent's position, the mechanism can control the agent's perceived expected wait and reduce welfare loss.

Formally, consider a partial information mechanism that is identical to the waiting list with declines except (i) agents who decline an item are not offered that item again,[27] and (ii) agents may be given partial information about their position. Let S denote the state space of the mechanism (see Lemma 4) and let $\mathfrak{S}$ denote a set of signals. A partial information mechanism commits to information disclosure

[27] Under full information, it is immaterial whether agents who previously declined a B item are offered a B again because the expected wait for the preferred item can only decrease over time. Che and Tercieux (2020) consider a partial information setting in which agents may learn over time and may change their decision with time.

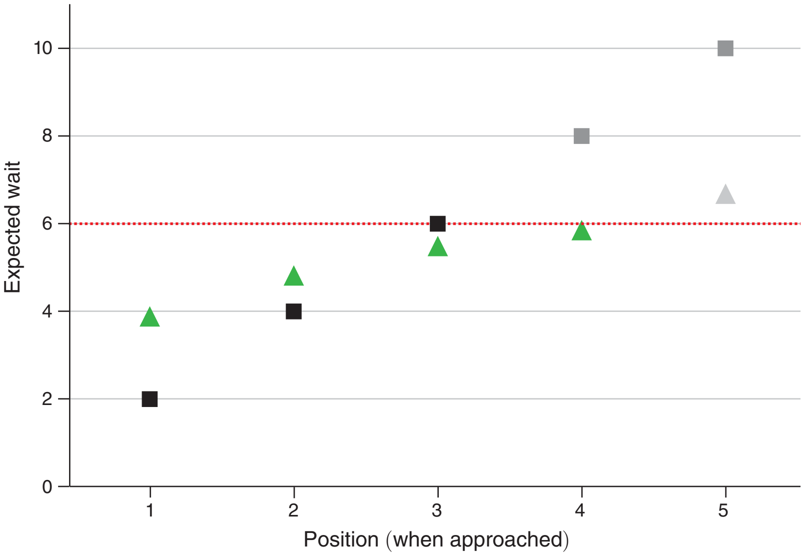

Figure 2

Notes: Equilibrium expected wait w_k for agent in position k under the waiting list with declines (black squares) and the SIRO buffer-queue mechanism (green triangles). Parameters used are $p_\alpha = p_A = 1/2$, $c = 1$, and $v = 6$. The dotted line indicates $\bar{w} = v/c = 6$. In equilibrium, agents wait for their preferred item in positions $k \leq 3$ under the waiting list with declines, or positions $k \leq 4$ under the SIRO buffer queue. Markers for positions in which agents are not willing to wait for their preferred item are shaded.

$\Upsilon : \{A,B\} \times S \to \Delta(\mathfrak{S})$ that discloses a signal given the state of the system and the current item kind. An α agent who is offered a B item and observes the disclosed signal $\mathfrak{s} \in \mathfrak{S}$ believes that the expected wait for A is[28] $w_{\mathfrak{s}} = E[w_{\tilde{k}} \mid \mathfrak{s}]$, and will prefer to wait for an A if $w_{\mathfrak{s}} \leq \bar{w}$.

As an illustration, consider an α agent in the setting $p_\alpha = p_A = 1/2$, $\bar{w} = 6$ depicted in Figure 2. Under full disclosure, the α agent is willing to decline a B and wait for an A in positions 1, 2, or 3. But an agent in position 4 prefers to take an immediate B as $w_4 = 4/p_A = 8 > \bar{w}$. Because the expected wait in position 2 is equal to $w_2 = 2/p_A = 4$ and is strictly below $\bar{w}$, an α agent is willing to wait for a B if he believes that his position is equally likely to be 2 or 4. That is, by hiding information the mechanism can induce an α agent in position 4 to wait for an A and avoid misallocation.

For general $p_\alpha = p_A$ and $\bar{w}$, a simple information disclosure allows the mechanism to minimize welfare loss.

THEOREM 2: *Suppose that $p_\alpha = p_A = p$, that $2p_A\bar{w}$, $2p_B\bar{w}$ are integers,*[29] *and assume that agents do not know their position k. Consider the information disclosure*

[28] We assume that agents know the steady state distribution of the system and infer the distribution of their possible position $\tilde{k}$ given the signal $\mathfrak{s}$ to calculate their expected wait $w_{\mathfrak{s}} = E[w_{\tilde{k}} \mid \mathfrak{s}]$.

[29] If $2p_A\bar{w}$ is not an integer, the information disclosure that minimizes welfare loss sends a randomly selected message to agent in position $\lfloor 2p_A\bar{w} \rfloor$. Let the message space be $S = \{$"wait", "mismatch"$\}$. Agents in positions $k < \lfloor 2p_A\bar{w} \rfloor$ are sent the message "wait." Agents in positions $k > \lfloor 2p_A\bar{w} \rfloor$ are sent the message "mismatch." Agents in position $k = \lfloor 2p_A\bar{w} \rfloor$ are sent the message "wait" with probability q and the message "mismatch" with

Υ^* under which agents offered an arriving B item are only informed whether $k \in \{1,\ldots,2p_A\bar{w} - 1\}$ or $k > K_A^* = 2p_A\bar{w} - 1$ (and symmetrically for an arriving A item). Under information disclosure Υ^*, there is a unique equilibrium that yields the minimal welfare loss of any equilibrium under any information disclosure policy. Welfare loss under information disclosure Υ^* is equal to

$$WFL^{WLWD:\Upsilon^*} = v\frac{2p(1-p)}{(1-p)K_A^* + pK_B^* + 1} = c/2.$$

Under the information disclosure Υ^*, an α agent who is informed that his position is $k \geq \lfloor 2p_A\bar{w} \rfloor$ prefers to take the current B item, as the expected wait for A is above $\bar{w}$. Under the stationary distribution, an α agent who is informed that $k < \lfloor 2p_A\bar{w} \rfloor$ is equally likely to be in either of the positions $1,\ldots,\lfloor 2p_A\bar{w} \rfloor - 1$, and thus believes his expected wait is $\lfloor 2p_A\bar{w} \rfloor / 2p_A \leq \bar{w}$ and prefers to wait for an A. For an illustration, see Figure 5. A similar approach can be used when $p_\alpha \neq p_A$.

This partial information mechanism eliminates approximately half of the welfare loss of the waiting list with declines. By equalizing the expected wait offered to agents, the mechanism can offer an acceptable expected wait to more agents.

The partial information disclosure Υ^* does not eliminate all welfare loss. For example, in the setting depicted in Figure 2, an α agent in position 6 will take an immediate B item. Because the mechanism reveals some information by disclosing the currently available item, no partial information mechanism can achieve lower welfare loss. For example, if agents in position 6 are given a signal to wait for an A, the average expected wait of agents who receive the signal will be above $\bar{w} = 6$, and some α agents would have preferred to take an immediate B item.

A mechanism that hides all information, including which item is currently offered,[30] can induce all agents to choose their preferred item if the system is balanced and agents hold correct steady-state beliefs. However, hiding all information may be problematic in practice for reasons that are beyond the scope of the model.[31] For example, the mechanism may want to provide information to agents to ensure they are not misinformed, or to reduce agents' incentive to collect information from other sources. In addition, the planner may want to provide agents with information to help manage expectations. Similar concerns arise for partial information mechanisms as well. First, in many situations the planner will have imperfect control over the information available to agents. Under partial information disclosure, agents will have an incentive to collect information about their position from external sources, such as online forums. Equity concerns may arise, as agents with access to better information will receive more favorable outcomes. Second, to implement Υ^* the planner needs to know the environment parameters, while the waiting list with

probability $1 - q$, where q is selected so that the expected wait conditional on receiving a "wait" message is equal to $\bar{w}$.

[30] That is, the mechanism asks agents to declare their preference without knowing the currently offered item. If agents are mismatched, then they join the queue for their preferred item type.

[31] In addition, such a mechanism can accumulate an unbounded number of agents in the buffer queue, and will not satisfy the overloaded waiting list assumption. Because the state follows an unbiased random walk, the expected wait for one item can grow unbounded. Moreover, the system can spend an arbitrarily long time in states in which agents face a long wait for A items while B items are offered for immediate assignment, making it difficult to hide the difference between waiting times.

declines is parameter free. Moreover, the planner will need to adjust the mechanism with any change in item arrival rates or agent preferences. Third, the behavior of agents depends on their beliefs about preferences of other agents and the stationary distribution of the system. For example, if the planner implements Υ^* for $p_\alpha = p_A = 1/2$, an α agent who believes $\hat{p}_\alpha = 3/4$ will refuse to wait for an A.

In practice, the mechanism should strike a balance between the need to provide agents with useful information and hiding the system's fluctuations. This can be achieved by disclosing historic expected waits that are updated sufficiently frequently to be relevant and sufficiently infrequently to avoid random fluctuations.

IV. Analysis via a Buffer-Queue Representation

This section introduces the class of buffer-queue mechanisms that allows for a tractable representation and analysis of the waiting-list-with-declines mechanism. This class of mechanisms is also useful for considering alternative mechanisms that use different queuing priorities to control expected wait fluctuations. Section V discusses incentives in this class of mechanisms and leverages the results in this section to derive improved incentive compatible mechanisms.

A buffer-queue mechanism represents an α agent declining a *B* to wait for an *A* as having the α agent join a *buffer queue* for *A* items. A buffer-queue mechanism maintains a buffer queue for *A* and a buffer queue for *B*. All approached agents who previously declined a *B* item are held in the *A* buffer queue until assigned. Arrivals of *A* items are assigned to an agent in the *A* buffer queue according to the queuing policy (for example, Lemma 5 shows that the waiting list with declines is equivalent to a buffer queue with the FCFS queuing policy). If an *A* item arrives when the *A*buffer queue is empty, the mechanism approaches new agents (skipping agents in the *B* buffer queue, if there are any) and offers them a choice between taking the immediate *A* item or joining the *B* buffer-queue. Arrivals of *B* items are treated symmetrically.

The buffer-queue mechanism representation offers several benefits. First, this representation provides a natural state space that is used in the analysis to track the stochastic evolution of the system and calculate the stationary distribution. Second, this representation allows us to generate different schedules of expected wait by specifying a queuing policy φ that determines the relative priority among agents who declined items.

Last, the representation allows us to specify a direct mechanism, in which agents make their choice by reporting their type. The mechanism determines in each state (subject to incentive constraints) whether a mismatched agent is to join the buffer queue to wait for the preferred item, or is to be assigned an immediate mismatched item. We consider a simple parameterization, under which mismatched α agents are to join the *A* buffer queue for their preferred item if it holds less than K^A agents. We refer to K^A as the maximal buffer-queue size and say the *A* buffer queue is full if it holds K^A agents. Similarly, the mechanism specifies K^B for the *B* buffer queue.

DEFINITION 2: *A buffer-queue mechanism* $\mathcal{M} = (K^A, \varphi^A, K^B, \varphi^B)$ *is a dynamic mechanism parameterized by two buffer-queue policies:* (K^A, φ^A) *for the buffer-queue holding agents waiting for A items, and* (K^B, φ^B) *for the buffer-queue holding agents waiting for B items.*

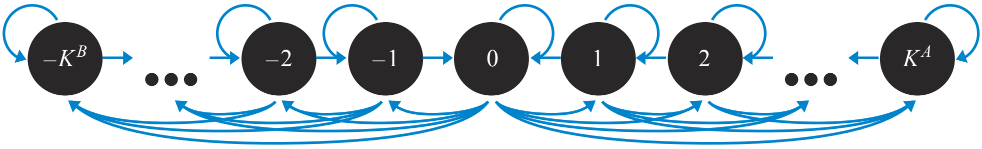

FIGURE 3. ILLUSTRATION OF THE MARKOV CHAIN S

Notes: Transitions toward state 0 correspond to an assignment of the current item to an agent on the respective buffer-queue. Transitions away from 0 correspond to the mechanism offering the current item to (potentially multiple) new agents.

To simplify notation, we restrict attention to queuing policies that track and prioritize agents based on their position in the buffer queue. Agents who join the buffer queue take the first empty position, and move forward when an agent ahead of them is assigned. This class is sufficiently general to capture mechanisms such as the waiting list with declines, and the SIRO buffer queue.

DEFINITION 3: *A buffer-queue policy (K,φ) is given by the maximal buffer-queue size $K \in \mathbb{N}$, and nonnegative assignment probabilities $\varphi = \{\varphi(k,i)\}_{1\leq i\leq k\leq K}$ such that $\sum_{i=1}^{k}\varphi(k,i) = 1$ for all $1 \leq k \leq K$.*

That is, if an item arrives when there are k agents in the A buffer queue, it will be allocated to the agent in position i with probability $\varphi(k,i)$. This class includes common queuing policies. The FCFS queuing policy is equivalent to

$$\varphi^{\text{FCFS}}(k,i) = \begin{cases} 1, & \text{if } i = 1; \\ 0, & \text{if } i \neq 1; \end{cases}.$$

The last-come-first-served (LCFS) queuing policy is equivalent to

$$\varphi^{\text{LCFS}}(k,i) = \begin{cases} 1, & \text{if } i = k; \\ 0, & \text{if } i \neq k; \end{cases}.$$

A. *Dynamics and Welfare*

The evolution of a buffer-queue mechanism is a stochastic process due to the random arrival of items and agent types. The following analysis describes the dynamics, assuming all agents report their type truthfully, and calculates the implied misallocation rate.

LEMMA 4: *The evolution of a buffer-queue mechanism $\mathcal{M} = (K^A,\varphi^A,K^B,\varphi^B)$ is a stochastic process that is generated by an ergodic Markov chain over the state space*

$$S = \{-K^B,\ldots,-1,0,1,2,\ldots,K^A\},$$

where $k \geq 0$ corresponds to k agents of type α waiting in the A buffer queue and $k \leq 0$ corresponds to $|k|$ agents of type β waiting in the B buffer queue. At most one buffer queue is nonempty at any given time. Each transition of the Markov chain corresponds to one period and one assigned item.

The state of the system can be thought of as the imbalance between the supply of arriving items and the demand from approached agents. The maximal sizes of the buffer queues give the range of imbalances the mechanism can sustain, and the mechanism is forced to misallocate items when the imbalance exceeds the range $\{-K^B, \ldots, K^A\}$. Within this range, imbalance randomly fluctuates.

Transition probabilities are calculated in Appendix A from the arrival probabilities of items and choices of agents. To calculate the stationary distribution, we use the Markov chain $\hat{S}$, which includes all the original states of S as well as two additional sets of states S^A and S^B. This construction builds on the Markov chain introduced by Caldentey, Kaplan, and Weiss (2009). A state $(k, B) \in S^B$ indicates k agents of type α are in the A buffer queue, and a current B item is about to be offered to a new agent. Similarly, a state $(-k, A) \in S^A$ indicates k agents of type β are in the B buffer queue, and a current A item is about to be offered. The original states are relabeled as $(k, \phi) \in S^\phi \cong S$. Each period starts and ends in a state in S^ϕ.

The Markov chain on $\hat{S}$ is depicted in Figure 4. Appendix A contains the full analysis of the Markov chain and related proofs. It allows us to calculate the stationary distribution over $\hat{S}$ and the misallocation rate.

THEOREM 3: *Let $\mathcal{M} = (K^A, \varphi^A, K^B, \varphi^B)$ be a buffer-queue mechanism. If $p_\alpha \neq p_A$, the misallocation rate under $\mathcal{M}$ when agents are truthful is equal to*

$$\xi = (p_A - p_\alpha)\frac{(p_\beta/p_B)^{K^B+1} + (p_\alpha/p_A)^{K^A+1}}{(p_\beta/p_B)^{K^B+1} - (p_\alpha/p_A)^{K^A+1}}.$$

If $p_\alpha = p_A = p$, the misallocation rate ξ is

$$\xi = \frac{2p(1-p)}{(1-p)K^A + pK^B + 1}.$$

Moreover, ξ is monotonically decreasing in K^A, K^B, and

$$\lim_{K^A \to \infty} \xi = \lim_{K^B \to \infty} \xi = |p_A - p_\alpha|.$$

The misallocation rate for $p_\alpha = p_A = p$ has an intuitive interpretation. If an item arrives and the respective buffer queue is full, the mechanism assigns the item to the next approached agent regardless of their type. The numerator $2p(1-p)$ captures the probability this assignment results in misallocation, and it is equal to the misallocation rate in the sequential assignment without choice mechanism. If the respective buffer queue is not full, misallocation is avoided by having a mismatched agent join the buffer queue. When K^A, K^B are larger it is less likely that the buffer queue is full, which is captured by the denominator $(1-p)K^A + pK^B + 1$ that is increasing in K^A, K^B.

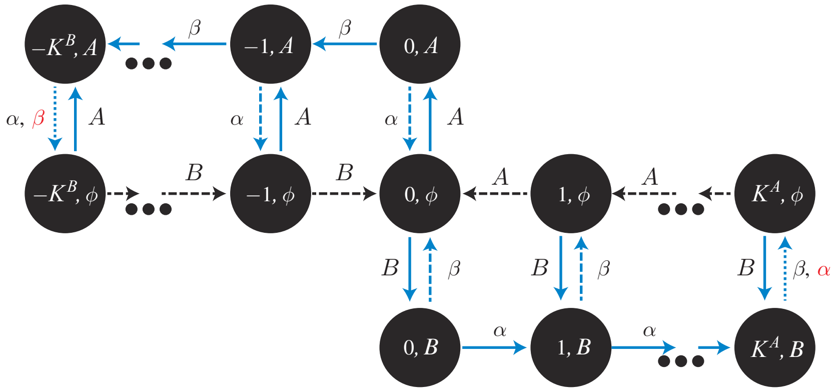

FIGURE 4

Notes: The Markov chain over state space $\hat{S}$. Arrows labeled A, B correspond to the arrival of that item. Arrows labeled α, β correspond to an offer to a new agent of that type, assuming agents are truthful. Dashed lines are transitions that assign the current item, and dotted lines indicate the current item is assigned to a potentially mismatched agent.

Theorem 1 is a direct corollary of Theorem 3 and the following lemma.

LEMMA 5: *The waiting list with declines has a unique equilibrium which is equivalent to the buffer-queue mechanism* $\mathcal{M} = \left(K^A, \varphi^{FCFS}, K^B, \varphi^{FCFS}\right)$ *with* $K^A = \lfloor p_A \bar{w} \rfloor$ *and* $K^B = \lfloor p_B \bar{w} \rfloor$.

V. Controlling Expected Wait Fluctuations via Queuing Policy Design

The preceding analysis shows that maximizing welfare is equivalent to minimizing misallocation. To incentivize agents to avoid misallocation, the mechanism needs to offer an acceptable expected wait. Figure 1 shows that using a different queue design can allow the mechanism to control the expected wait offered to agents to maintain an acceptable expected wait under a wider range of fluctuations and reduce welfare loss. This section analyzes possible queue designs to derive an optimal design and give a characterization of the SIRO queuing policy as the robustly optimal policy.

Formally, given a mechanism $\mathcal{M} = \left(K^A, \varphi^A, K^B, \varphi^B\right)$, let w_k^A denote the implied expected wait for an agent who declines a B and joins position k in the A buffer queue.[32] Let w_k^B be defined symmetrically. To simplify notation, we assume $\bar{w} \geq \max\{1/p_A, 1/p_B\}$ throughout this section, ruling out trivial parameters in which agents are unwilling to wait for the first arrival of their preferred item.

LEMMA 6: *The expected waits* $\left\{w_k^A\right\}_{k=1}^{K^A}$ *depend only on* $\left(K^A, \varphi^A\right)$, *and* p_α, p_A *(and symmetrically for* w_k^B*).*

[32] That is, w_k^A is the expected number of periods from when the agent joins the buffer queue until he is assigned an A item, conditional on joining the A buffer queue when it holds $k - 1$ agents, and assuming the following agents truthfully report their type.

Therefore, the following is well defined.

DEFINITION 4: *A buffer-queue policy (K, φ) with expected waits $\{w_k\}_{k=1}^{K}$ is incentive compatible* (*IC*) *if $w_k \leq \bar{w}$ for all $k \leq K$. A buffer-queue mechanism $\mathcal{M} = (K^A, \varphi^A, K^B, \varphi^B)$* is *IC* if both (K^A, φ^A) *and* (K^B, φ^B) *are* IC.

Under an IC mechanism, it is an equilibrium for all agents to report their type truthfully.[33] The following lemma uses Little's law to show that the average expected wait depends only on K and is independent of the queuing policy φ.

LEMMA 7: *Let (K, φ) be a buffer-queue policy. Then, independently of φ, the average expected wait for an agent who joins the buffer-queue is*

$$W(K) = E[w_{\tilde{k}}] = \begin{cases} \dfrac{K+1}{2p}, & \text{if } p_\alpha = p_A = p; \\ \dfrac{K}{p_A} + \dfrac{1}{p_A - p_\alpha} + \dfrac{1}{p_A}\dfrac{K}{(p_\alpha/p_A)^K - 1}, & \text{if } p_\alpha \neq p_A. \end{cases}$$

Moreover, if (K, φ) is an IC buffer-queue policy, then $W(K) \leq \bar{w}$.

PROOF:

By Little's law (Little 1961), if $W = E[w_{\tilde{k}}]$ is the average time an agent spends in the buffer queue, L is equal to the average number of agents in the buffer queue conditional on the buffer queue being nonempty, and λ is the arrival rate at which agents join/leave the buffer queue, then we have that $L = \lambda W$. The expected number of agents that leave the buffer queue in any given period is $\lambda = p_A$. If $p_\alpha = p_A = p$, the buffer-queue is equally likely to hold any number of agents k for $1 \leq k \leq K$ (by Lemma 10 in Appendix A), and the average number of agents in the buffer queue is $L = (K+1)/2$. Therefore, $E[w_{\tilde{k}}] = W = L/\lambda = (K+1)/2p$, which is independent of φ. Last, for any IC buffer-queue policy (K, φ) we have $W(K) = E[w_{\tilde{k}}] \leq \max_{k \leq K} w_k \leq \bar{w}$. The case $p_\alpha \neq p_A$ is proved similarly in online Appendix D. ∎

Intuitively, Lemma 7 shows that a mechanism with higher K (that is, avoiding misallocation under states of greater imbalance) needs to offer some agents a longer expected wait. The FCFS policy offers agents the minimal feasible expected wait at the current state, which is low in low states but increases with the state. A different queue policy can distribute the expected wait more equally and offer agents an expected wait below $\bar{w}$ even under greater imbalance. But if $E[w_{\tilde{k}}] > \bar{w}$, it is not possible to redistribute the expected wait so that all agents face an expected wait below $\bar{w}$. We thus obtain a lower bound for the welfare loss of any IC buffer-queue mechanism.

[33] Every agent makes at most a single choice. An α agent will always take a current A item to attain the maximal possible utility. Because immediate assignment to B is preferable to never being assigned, the mechanism can force the agent to take the current B item. If the current item is a B item and the A buffer queue is not full, the agent can choose whether to truthfully reveal he is mismatched and wait for a future A, or misreport his type and take the current B item. Given expected wait w, the α agent prefers joining the A buffer queue if $v - c \times w \geq 0$ or $w \leq v/c = \bar{w}$.

COROLLARY 3: *Suppose that $p_\alpha = p_A = p$. Then the welfare loss under any IC buffer-queue mechanism is at least*

$$v\frac{2p(1-p)}{(1-p)\lfloor 2p\bar{w}\rfloor + p\lfloor 2(1-p)\bar{w}\rfloor}.$$

A. *The Robust SIRO Queuing Policy*

The SIRO policy is a simple alternative to FCFS. By giving all agents in the buffer queue equal probability of receiving an arriving item, the SIRO policy reduces the variation in expected wait. Figure 5 provides an illustration. Intuitively, the expected wait w_K for the agent joining the last position K is lower when the last agent is given equal priority. But while there is less variation in w_k under SIRO than under FCFS, the expected waits w_k under SIRO are strictly increasing in k because the probability of getting assigned is lower when there are more agents in the buffer queue.

This section gives a characterization showing that the SIRO policy maximizes welfare subject to a robust incentive compatibility constraint. A rough intuition is that the policy wants to give higher assignment priority to agents in the last positions to reduce w_k for high k. This requires reducing the priority of agents who entered the buffer queue from the first positions, and increases the expected wait of such agents in the event that many agents join the buffer queue behind them.[34] On the other hand, agents who join the first positions may believe that additional agents will join after them, and to be robustly incentive compatible the mechanism needs to ensure that such pessimistic agents are willing to join the first positions. Giving agents in all positions equal assignment priority balances the two considerations.

Formally, given a mechanism $\mathcal{M} = (K^A, \varphi^A, K^B, \varphi^B)$, let $w^A_{k,\sigma}$ denote the *subjective expected wait* of an agent with belief σ who declines a B and joins position k in the A buffer queue ($w^B_{k,\sigma}$ is defined symmetrically). Define a general belief σ as follows.[35] Label the following agents by the order in which they are approached and offered the option to join the A buffer queue. The agent's belief $\sigma : \mathbb{N} \times \mathbb{N} \to [0,1]$ specifies the probability $\sigma_\ell(k)$ that the ℓth agent will report to be of type α and join the A buffer queue conditional on being offered position k.[36] The belief that corresponds to all agents being truthful is given by $\sigma_\ell(k) \equiv p_\alpha$ for all $\ell \geq 1$ and $k \leq K^A$. The subjective expected wait $w^A_{k,\sigma}$ is calculated by drawing future agents independently according to σ.

The following lemma generalizes Lemma 6.

[34] For example, the first agent to join an LCFS queue faces a long expected wait if many subsequent agents join.

[35] We maintain that agents have correct beliefs about the item arrival rates, as different beliefs about item arrival rates scale the expected waits w_k.

[36] That is, $\sigma_1(k)$ gives the probability that the next agent approached will report to be an α agent and join the buffer queue conditional on seeing the state where $k-1$ agents are in the buffer queue. This formulation indexes agents by the order in which they are approached (instead of referring to agents by name) allowing a more tractable formulation of belief updates.

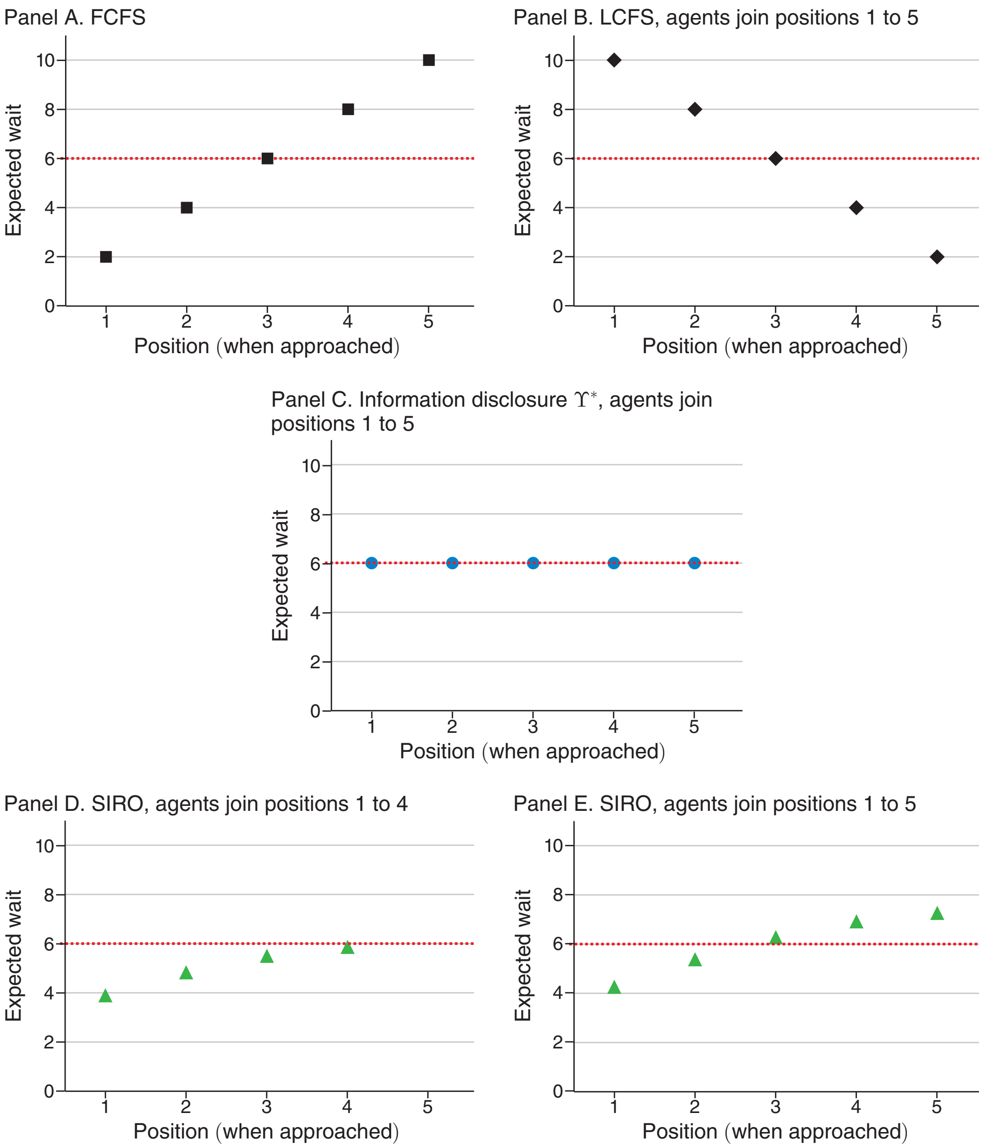

Figure 5

Notes: Expected wait for agents who join the buffer queue to wait for their preferred item under various queuing policies for $p_\alpha = p_A = 1/2$. The dotted line indicates $\bar{w} = 6$.

LEMMA 6’: *The expected waits* $\left\{w^A_{k,\sigma}\right\}_{k=1}^{K^A}$ *depend only on* $\left(K^A, \varphi^A\right)$, *the belief* σ, *and* p_A (*and symmetrically for* $w^B_{k,\sigma}$).

Therefore, the following is well defined.

DEFINITION 5: *A buffer-queue policy* (K, φ) *with expected waits* $\{w_k\}_{k=1}^{K}$ *is belief-free incentive compatible* (*BF-IC*) *if* $w_{k,\sigma} \leq \bar{w}$ *for all* $k \leq K$ *and any belief* σ. *A buffer-queue mechanism* $\mathcal{M} = \left(K^A, \varphi^A, K^B, \varphi^B\right)$ *is BF-IC if both* $\left(K^A, \varphi^A\right)$ *and* $\left(K^B, \varphi^B\right)$ *are BF-IC.*

In other words, a mechanism is BF-IC if agents who wait for their preferred item face an acceptable subjective expected wait regardless of their beliefs. In particular, BF-IC implies that an agent should not regret joining the buffer queue even if subsequent agents join the buffer queue after him. The following theorem shows that the simple SIRO buffer-queue mechanism obtains the lowest possible welfare loss of any BF-IC mechanism. We denote $\kappa^*(\bar{w}, p_\alpha, p_A) = \sup\{K' \,|\, W(K') \leq \bar{w}\}$ using the average expected wait function $W(\cdot)$ defined in Lemma 7.

THEOREM 4: *Let $\mathcal{M}^\star = (K^A, \varphi^{SIRO}, K^B, \varphi^{SIRO})$ be the SIRO buffer-queue mechanism given by $\varphi(k,i) = 1/k$ for any $i \leq k$, $K^A = \kappa^*(\bar{w}, 1, p_A)$, and $K^B = \kappa^*(\bar{w}, 1, p_B)$. Then, $\mathcal{M}^\star$ is BF-IC and achieves a weakly lower welfare loss than any BF-IC buffer-queue mechanism.*

As stated above, the intuition for the result is that SIRO balances the priority of agents. Agents may hold a belief $\sigma \equiv 1$, which is equivalent to the belief $\hat{p}_\alpha = 1$. Agents with this belief who enter position i believe that a large number of agents will join the buffer queue after them, and they will end up in position i in a buffer queue that holds the maximal number of agents K. To maintain BF-IC, it must be that such an agent does not regret joining the buffer queue. On the other hand, prioritizing agents that joined earlier over the agents that join later prevents the policy from offering a low expected wait to agents who join an almost full buffer queue. The SIRO policy balances these two goals by giving agents who join the buffer queue the maximal priority such that none of the agents already in the buffer queue regret joining, which is to give them all equal priority.

Figure 5 presents the expected waits under SIRO and other policies for $p_\alpha = p_A = 1/2$. It illustrates how the SIRO policy improves upon the FCFS policy. Any IC FCFS policy is also BF-IC, because expected waits under FCFS are independent of whether future agents join the buffer queue. But SIRO reduces the variation in w_k and maintains an acceptable expected wait under a larger range of states than FCFS. In addition, the SIRO policy is simpler than the FCFS policy in that it does not require tracking positions of agents within the buffer queue. Figure 5 also shows that w_k increases with k.

The variance of realized wait is higher under SIRO than under FCFS, and some agents can wait significantly longer than w_k before being assigned.[37] In the context of our model, any agent waiting in the buffer queue will prefer to keep waiting for their preferred item over taking an immediate mismatched item, because past waiting costs are sunk and the expected wait of any agent in the buffer queue is at most $w_K \leq \bar{w}$. However, the increased variance of waiting times may be undesirable for agents who wish to plan ahead. Additionally, although agents are offered more equitable expected waits when they make their choice, the realized wait may be less equitable. Section VI presents simulation results showing that a modification of the SIRO policy can mitigate this concern.

[37] Vasicek (1977) shows that FCFS minimizes the variance of waiting times while LCFS maximizes it.

B. *Indirect Parameter-Free SIRO Mechanism*

The SIRO mechanism $\mathcal{M}^\star = (K^A, \varphi^{SIRO}, K^B, \varphi^{SIRO})$ is almost a prior-free mechanism. The queuing policy φ^{SIRO} is prior free, but the planner needs information on the agent's valuations to appropriately set K^A, K^B to determine in which states agents decline mismatched items. If K^A, K^B are too high, the mechanism $\mathcal{M}^\star$ is not IC. If K^A, K^B are too low, the mechanism $\mathcal{M}^\star$ forces mismatched agents approached in the states K^A and $-K^B$ to take the current item even though they would have preferred to wait for their preferred item.[38] This section presents a prior-free indirect mechanism in which K^A, K^B are determined in equilibrium. This mechanism is a simple indirect mechanism that improves upon the waiting list with declines.

Consider $\mathcal{M}^\circ = (\infty, \varphi^{SIRO}, \infty, \varphi^{SIRO})$ as an indirect mechanism. In contrast to a direct mechanism $\mathcal{M}^\star = (K^A, \varphi^{SIRO}, K^B, \varphi^{SIRO})$ with finite K^A, K^B, all agents are offered the option to join the buffer queue and wait for their preferred item (that is, the planner does not impose a cap on the size of the buffer queues). The mechanism $\mathcal{M}^\circ$ differs from the waiting list with declines only in that all agents who declined items are equally likely to be assigned a future arrival of their preferred item.

Restrict attention to strategies in which agents take an immediate matching item.[39] Denote a mixed strategy of an α agent under $\mathcal{M}^\circ$ by $s : \mathbb{N} \rightarrow [0,1]$, where $s(k)$ is the probability that the α agent in position k declines a mismatched B item and waits for an A. Because of the SIRO queuing policy, the expected waits $\{w_k\}$ depend[40] on s (for an illustration, compare Figure 5, panels D and E). A strategy s constitutes a Nash equilibrium if for the corresponding $\{w_k\}$ it holds that $s(k) > 0 \Rightarrow w_k \leq \bar{w}$ and $s(k) < 1 \Rightarrow w_k \geq \bar{w}$.

LEMMA 8: *If s^* is an equilibrium of the indirect mechanism $\mathcal{M}^\circ = (\infty, \varphi^{SIRO}, \infty, \varphi^{SIRO})$, then $s^*(k) = 1$ for $1 \leq k \leq \kappa^*(\bar{w}, 1, p_A)$.*

In other words, if agents will decline a mismatched item in a state k under a BF-IC direct SIRO mechanism $\mathcal{M}^\star$, agents will decline a mismatched item in a state k under an equilibrium s^* of the indirect SIRO mechanism $\mathcal{M}^\circ$. An immediate corollary is that equilibrium welfare under the indirect mechanism is higher than under the optimal BF-IC SIRO mechanism $\mathcal{M}^\star$, and therefore also higher than the welfare under the waiting list with declines.

Under the indirect mechanism $\mathcal{M}^\circ$, any strategy s that declines mismatched items in some position k is not a dominant strategy, because an agent who joins the SIRO buffer queue faces arbitrarily long expected wait if sufficiently many subsequent agents join after him. However, the expected wait of any agent in the SIRO buffer queue equals the expected wait of the last agent who joins the buffer queue in the same period. If all α agents share identical preferences and beliefs, the decision of

[38] An extreme example is the sequential assignment without choice mechanism, which is equivalent to $K^A = K^B = 0$.

[39] That is, we rule out strictly dominated strategies in which agents decline immediate assignment to their preferred item.

[40] By Lemma 6', the expected waits $\{w_k\}$ are independent of the strategy chosen by β agents if under any strategy β agents always take an immediate B item.

the last agent to join indicates that all agents in the buffer queue also prefer to have joined.

LEMMA 9: *Let s^* be an equilibrium of the indirect mechanism $\mathcal{M}^\circ = (\infty, \varphi^{SIRO}, \infty, \varphi^{SIRO})$. Then at the end of each period, any agent in the buffer queue prefers staying in the buffer queue to being immediately assigned a mismatched item.*

The assumption that all agents in the buffer queue have identical preferences is necessary for Lemma 9. For example, if agents have heterogeneous mismatch values, an agent with a high mismatch value may regret joining the buffer queue if many agents with low mismatch values (who are willing to wait longer for their preferred item) join the buffer queue after him.

Under the indirect mechanism, the burden of deciding in which states to decline mismatched items falls on the agents. If agents are provided with historical expected wait estimates, their decision is simple, because they need only to decide whether the offered expected wait is acceptable. Altman and Shimkin (1998) prove simple learning dynamics converge to equilibrium for a similar SIRO queuing game.

Figure 6 depicts welfare loss under different mechanisms for varying values of v. The figure shows the equilibrium welfare loss under the waiting list with declines (FCFS) and the indirect SIRO buffer-queue mechanism. Figure 6 also shows the minimal welfare loss under any BF-IC buffer-queue mechanism (labeled SIRO BF-IC) which is achieved by a direct SIRO buffer-queue mechanism that appropriately sets K^A, K^B to restrict entry to the buffer queue. In addition, Figure 6 also depicts the lower bound for the welfare loss under any IC buffer-queue mechanism given by Corollary 3. Note that SIRO captures more than half of the difference between FCFS and the lower bound.

VI. Limiting Realized Envy

A potential challenge in implementing the SIRO buffer-queue mechanism is that the random assignment can cause some agents to wait significantly longer than expected. Moreover, agents who keep waiting and see others assigned before them experience realized envy and may be understandably aggravated. This section uses simulation to quantify realized envy and evaluate heuristics that limit realized envy.

To quantify realized envy, define the overtaking count of an agent who joins position k in the buffer queue[41] and assigned the ℓth item to arrive[42] to be $\max\{\ell - k, 0\}$. That is, the overtaking count measures whether an agent experiences a longer wait than the agent would have faced under FCFS. Under FCFS, the overtaking count of all agents is 0. Figure 7 shows the distribution of overtaking counts for agents who join a SIRO buffer queue with $K \in \{4, 10\}$, for $p_\alpha = p_A = 1/2$. The figure presents the distribution for all agents, as well as the distribution for agents who join an empty buffer queue. Both distributions show a small but nonnegligible probability that agents experience significant realized envy.

[41] That is, there were $k - 1$ agents in the buffer queue just before the agent joined.

[42] That is, ℓ items arrived while the agent was waiting in the buffer queue.

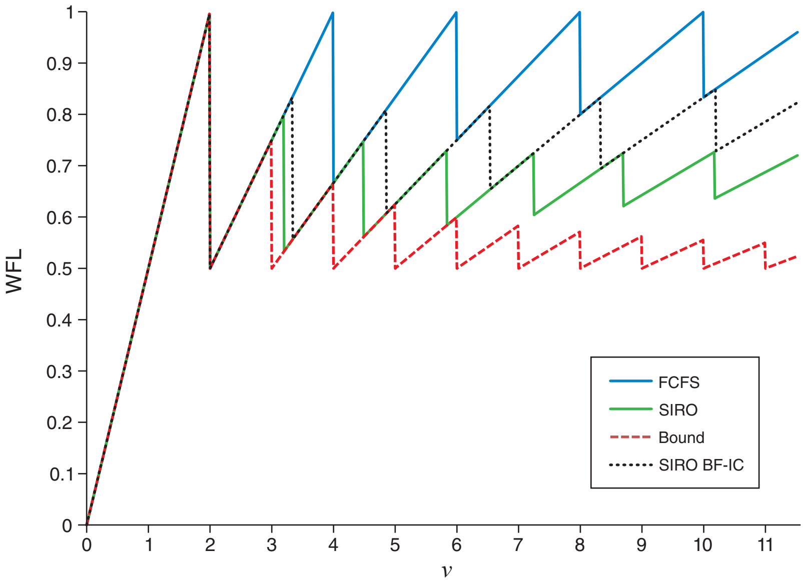

Figure 6. Welfare Loss for Varying Values of v and $p_\alpha = p_A = 1/2$, $c = 1$.

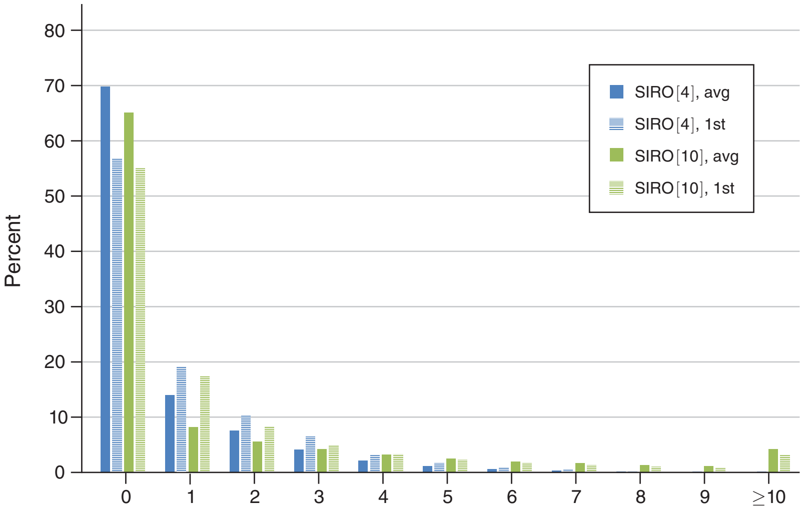

Figure 7

Notes: The distribution of the overtaking counts for agents who join a SIRO buffer queue. Parameters used are $K \in \{4, 10\}$ and $p_\alpha = p_A = 1/2$.

We simulate a simple heuristic that deviates from the SIRO policy by prioritizing agents that reach a specified overtaking limit. This simple heuristic bounds the possible overtaking counts an agent can experience, and thus avoids extreme realized envy. Setting an overtaking limit equal to 0 is equivalent to using a FCFS queuing priority. Setting an overtaking limit equal to ∞ is equivalent to using a SIRO queuing priority. By choosing an appropriate intermediate overtaking limit the planner

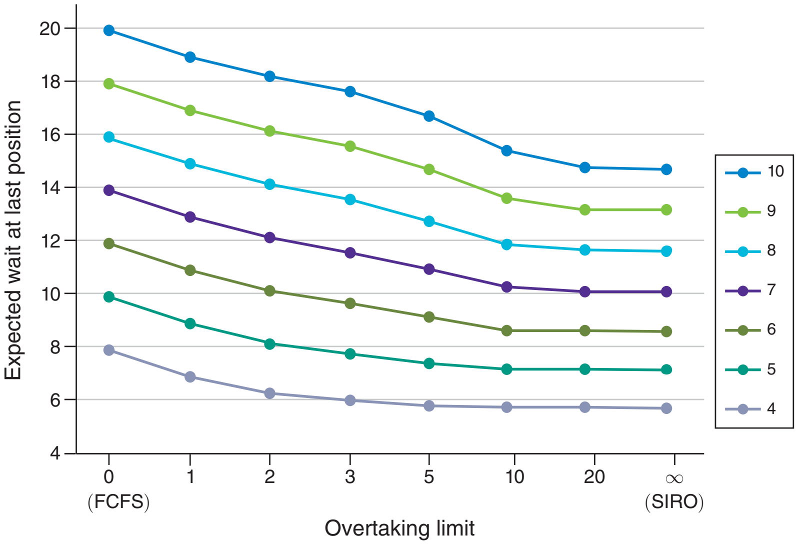

Figure 8

Notes: Expected wait for an agent who joins the last position of a SIRO buffer queue with an overtaking limit for $p_\alpha = p_A = 1/2$. Each line corresponds to a different buffer-queue size K and various overtaking limits. The FCFS policy corresponds to an overtaking limit of 0.

can strike a compromise between equalizing the expected waits and avoiding realized envy.

Figure 8 shows the performance of the SIRO with limited overtaking policy for various buffer-queue sizes and overtaking limits. The figure shows the expected wait w_K for an agent joining the last position in the buffer-queue given the maximal buffer-queue size K and the overtaking limit. Each buffer-queue policy is IC if $\bar{w} = v/c$ is higher than the depicted w_K. The figure shows that even a mild overtaking limit can significantly reduce the expected wait w_K relative to FCFS (or an overtaking limit of 0), which enables the planner to implement a larger IC buffer-queue size and lower misallocation. The figure also shows that a moderately high overtaking limit does not hinder the performance of the SIRO policy. An overtaking limit of 20 is unlikely to bind and therefore yields essentially the same w_K as SIRO.

Finally, we evaluate FCFS, SIRO, and SIRO with limited overtaking in a more elaborate setting with heterogeneous values in which the value of the preferred item v is drawn from $U[0,2]$ independently across agents.[43] In this setting, the welfare-maximizing assignment assigns all items to matching agents and attains an average assigned value of $E[v] = 1$. Because an agent assigned to a mismatched item decreases total welfare by v, the mechanism can reduce welfare loss either by reducing the misallocation rate or by replacing misallocations of agents with high v with misallocations of agents with low v.

[43] The value of the mismatched item is zero.

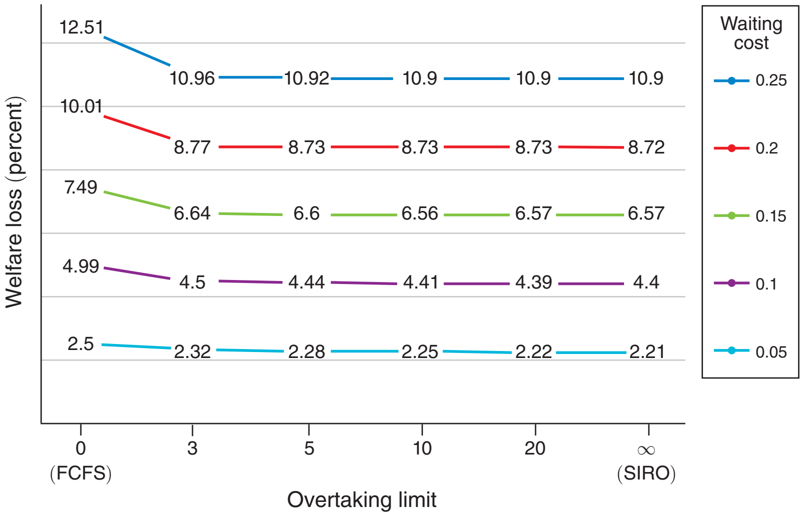

Figure 9

Note: Welfare loss under SIRO with various overtaking limits for an economy with $p_\alpha = p_A = 0.5$, v drawn from $U[0,2]$ independently for each agent and $c \in \{0.05, 0.1, 0.15, 0.2, 0.25\}$.

Figure 9 depicts the equilibrium welfare loss from misallocation given each policy for various delay costs $c \in \{0.05, 0.1, \ldots, 0.25\}$ and $p_\alpha = p_A = 0.5$. Equilibrium welfare was calculated by initializing estimates for $\{\hat{w}_k\}$, simulating the system with agents following the strategy of joining position k to wait for the preferred item if and only if $v - c \times \hat{w}_k > 0$ and calculating new expected wait estimates $\{\hat{w}_k\}$ from the simulation, and iterating until a fixed point is reached. This process converged to a fixed point quickly. Once a fixed point was found, an additional simulation used the equilibrium $\{w_k\}$ to calculate welfare loss.

Figure 9 shows that SIRO reduces welfare loss in this setting as well. The FCFS policy (given by an overtaking limit of 0) eliminates all realized envy but results in higher welfare loss. SIRO with limited overtaking reduces welfare loss, even with a small overtaking limit; in fact, most of the welfare gains can be attained by allowing a small overtaking limit.

VII. Concluding Remarks

SIRO may raise concerns of fairness, in that agents are not assigned in order. First, we note that SIRO is more fair than FCFS in that agents are offered a more consistent expected wait for their preferred items. Second, the ordering of agents on a waiting list may be arbitrary, and agents who sign up earlier may not have higher assignment value. For example, local parents may be able to register for daycare centers years in advance, whereas advanced registration is not possible for recently moved parents who may have a greater need for daycare. Constraining the mechanism to prioritize agents who made their choice earlier is equivalent to requiring the FCFS policy, which generates lower welfare than SIRO.

Under our assumptions, agents exert a positive externality when declining items because they are essentially letting other agents pass them in line. However, waiting-list policies commonly discourage agents from declining items.[44] One possible justification is that agents who decline items are an administrative burden. Although our analysis does not explicitly account for the time and costs required to administer the offers, under buffer-queue mechanisms, each agent is approached only once, limiting the administrative burden. Furthermore, these mechanisms require storing preference information for relatively few agents. We therefore believe overloaded waiting lists should encourage rather than discourage agents to decline mismatched items.

Put together, the results in this paper show that welfare in waiting list mechanism depends on the mechanism's ability to offer agents a choice between items and appropriate associated waiting times. When the system is overloaded, total waiting costs are constant and can only be transferred between agents. Expected waiting times serve a similar role to prices in guiding the assignment, but these prices fluctuate, potentially leading to misallocations. The SIRO buffer-queue mechanism and partial information mechanisms can reduce expected wait fluctuations and improve welfare.

Appendix A. The Buffer-Queue Markov Chain

In this Appendix, we describe the details of the Markov chain used to analyze the dynamics of the buffer-queue mechanism. The Markov chain captures changes in the buffer-queues from one period to another. Its states are $S = \left\{-K^B, \ldots, -1, 0, 1, 2, \ldots, K^A\right\}$, where $k \geq 0$ indicates k agents of type α waiting in the A buffer queue and $k \leq 0$ indicates $|k|$ agents of type β waiting in the B buffer queue. To see no other possible states of the system are possible, notice that at any time one of the buffer queues must be empty.

Recall that the period starts when the mechanism learns the type of the current period's item. If matching agents are in the buffer queue for the current item,[45] the mechanism assigns the item to the first agent in that buffer queue; the period ends and the next period starts with one less agent in that buffer queue.[46] If the buffer queue of the current item is empty,[47] the mechanism starts offering the current item to new agents. The mechanism continues to approach new agents until either a matching agent is found or a buffer queue reaches its maximal size. If the buffer queue reaches its maximal size, the mechanism assigns the item to the next new agent, regardless of his type. If the period started with $|k|$ agents in the buffer queue, the next period starts with $|\ell|$ agents in the buffer queue, where $|\ell - k|$ is the number of agents who declined the current item and joined the buffer queue. The possible transitions between s_t and s_{t+1} are depicted in Figure 3.

[44] We surveyed the waiting-list policies of the New York City Housing Authority, Newark Housing Authority, Boston Housing Authority, Atlanta Housing Authority, Philadelphia Housing Authority, the Housing Authority of Los Angeles, Miami-Dade County Public Housing, the Housing Authority of Baltimore City, and the Chicago Housing Authority. All of these authorities penalize agents if they decline apartments. In several authorities, agents who decline an apartment will not be offered another one.

[45] That is, an A item arrived and there are $s_t = k > 0$ agents of type α, or a B item arrived and $s_t = k < 0$.

[46] That is, if $k > 0$, then the next period starts with state $s_{t+1} = k - 1$, and if $k < 0$, the next period starts with state $s_{t+1} = k + 1$.

[47] That is, the current item is B and $k \geq 0$, or the current item is A and $k \leq 0$.

Each random transition of the Markov chain corresponds to a period, and transition probabilities are given by

$$\Pr(s_t = \ell \mid s_{t-1} = k) = \begin{cases} p_A, & \text{if } k > 0, \ell = k-1; \\ p_B p_\alpha^{\ell-k} p_\beta, & \text{if } k \geq 0, k \leq \ell < K^A, \ell \neq 0; \\ p_B p_\alpha^{\ell-k}(p_\alpha + p_\beta), & \text{if } k \geq 0, \ell = K^A; \\ p_A p_\alpha + p_B p_\beta, & \text{if } k = \ell = 0; \\ p_B, & \text{if } k < 0, \ell = k+1; \\ p_A p_\beta^{|\ell-k|} p_\alpha, & \text{if } k \leq 0, k \geq \ell > -K^B, \ell \neq 0; \\ p_A p_\beta^{|\ell-k|}(p_\beta + p_\alpha), & \text{if } k \leq 0, \ell = -K^B. \end{cases}$$

For example, if the period begins in state $s_t = 1$, that is, one α agent is waiting in the buffer queue, the probability that the period will end in state $s_{t+1} = 2 < K^A$ is $p_B p_\alpha p_\beta$. To calculate this probability, observe that the system accumulates another agent in the buffer queue when the following occurs. First, a B item arrives, which occurs with probability p_B. The B item is offered to a new agent, who declines the item and joins the buffer queue, which occurs if the agent is of type α with probability p_α. The item is then offered to the following new agent, who accepts the item, which occurs if that agent is of type β with probability p_β. If $K^A = 2$, the last agent will accept the item even if he is of type α, and so the probability of this transition becomes $p_B p_\alpha(p_\alpha + p_\beta) = p_B p_\alpha$.

We refine this Markov chain to describe the mechanism within a period using the extended set of states $\hat{S}$ defined by

$$\begin{aligned} \hat{S} &= S^\phi \cup S^A \cup S^B \\ &= \{(k,\phi) \mid -K^B \leq k \leq K^A\} \\ &\quad \cup \{(-k,A) \mid 0 \leq k \leq K^B\} \\ &\quad \cup \{(k,B) \mid 0 \leq k \leq K^A\}. \end{aligned}$$

The Markov chain on $\hat{S}$ is depicted in Figure 10. Using $\hat{S}$, we can generate the transitions over S by restricting attention to visits to S^ϕ states. Every period begins and ends in a state in S^ϕ. We move from a state in S^ϕ when an item arrives. For example, suppose the initial state is (k,ϕ) for $0 < k < K^A$. If an A item arrives, it is assigned to the first agent in the A buffer queue, the system transitions to state $(k-1,\phi)$, and the period ends. If a B item arrives, the system transitions to state (k,B). The transitions from a state (k,B) correspond to the mechanism approaching a new agent, and the following state depends on the type of the agent. If the new agent is of type β, he takes the current item, the system transitions to state (k,ϕ), and the period ends. If the new agent is of type α, he declines the item and joins the A buffer queue, the system transitions to state $(k+1,B)$, and the period continues.

For calculation purposes, the state space $\hat{S}$ has the advantage that all transitions are between adjacent states. Using the state space $\hat{S}$ and the flow equations of the

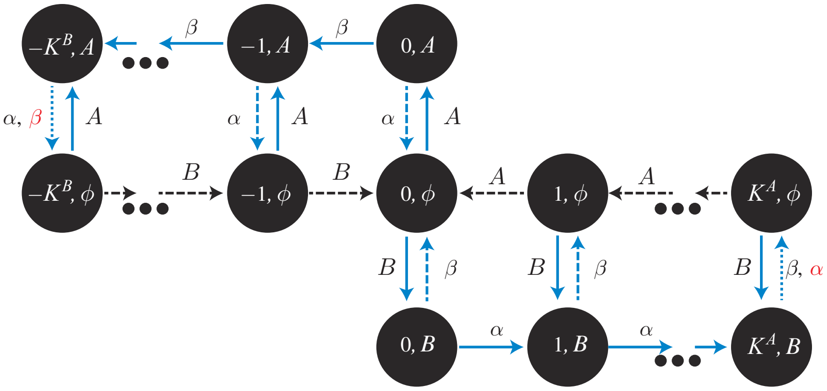

FIGURE 10. THE MARKOV CHAIN FOR THE STATE SPACE $\hat{S}$

Markov chain, we derive the stationary distribution that describes the long-run behavior of the system. Note that only two transitions in the chain imply misallocation: the transition from (K^A, B) to (K^A, ϕ) when an α agent is drawn, and the symmetric transition from $(-K^B, A)$ to $(-K^B, \phi)$ when a β is drawn. Transition probabilities are given by

$$\Pr\big(s \,|\, (k,\phi)\big) = \begin{cases} p_A, & \text{if } s = (k-1,\phi); \\ p_B, & \text{if } s = (k,B); \end{cases} \qquad k > 0,$$

$$\Pr\big(s \,|\, (k,\phi)\big) = \begin{cases} p_A, & \text{if } s = (k,A); \\ p_B, & \text{if } s = (k+1,\phi); \end{cases} \qquad k < 0,$$

$$\Pr\big(s \,|\, (0,\phi)\big) = \begin{cases} p_A, & \text{if } s = (0,A); \\ p_B, & \text{if } s = (0,B); \end{cases}$$

$$\Pr\big(s \,|\, (k,B)\big) = \begin{cases} p_\alpha, & \text{if } s = (k+1,B); \\ p_\beta, & \text{if } s = (k,\phi); \end{cases} \qquad 0 \le k < K^A,$$

$$\Pr\big(s \,|\, (K^A,B)\big) = \big\{ p_\alpha + p_\beta, \quad \text{if } s = (K^A,\phi);$$

$$\Pr\big(s \,|\, (k,A)\big) = \begin{cases} p_\alpha, & \text{if } s = (k,\phi); \\ p_\beta, & \text{if } s = (k-1,A); \end{cases} \qquad -K^B < k \le 0,$$

$$\Pr\big(s \,|\, (-K^B,A)\big) = \big\{ p_\alpha + p_\beta, \quad \text{if } s = (-K^B,\phi).$$

Using the Markov chain on $\hat{S}$, we can calculate the stationary distribution for both S and $\hat{S}$.

LEMMA 10: *The Markov chain is ergodic and its stationary distribution* π *over* $\hat{S}$ *is*

$$\pi(k,\phi) = \begin{cases} \left(\frac{p_\alpha}{p_A}\right)^k p_B \pi(0,\phi), & \text{if } k > 0; \\ \left(\frac{p_\beta}{p_B}\right)^{|k|} p_A \pi(0,\phi), & \text{if } k < 0; \end{cases}$$

and

$$\pi(k,B) = \begin{cases} \pi(k,\phi), & \text{if } k > 0; \\ p_B \pi(0,\phi), & \text{if } k = 0; \end{cases} \qquad \pi(k,A) = \begin{cases} \pi(k,\phi), & \text{if } k < 0; \\ p_A \pi(0,\phi), & \text{if } k = 0; \end{cases}$$

with

$$\pi(0,\phi) = \begin{cases} \frac{1}{2} \frac{1}{p_B K^A + p_A K^B + 1}, & \text{if } p_A = p_\alpha; \\ \frac{1}{2} \frac{p_A - p_\alpha}{p_A p_\beta \left(\frac{p_\beta}{p_B}\right)^{K^B} - p_B p_\alpha \left(\frac{p_\alpha}{p_A}\right)^{K^A}}, & \text{if } p_A \neq p_\alpha. \end{cases}$$

PROOF OF LEMMA 10:

It is clear that all states in the Markov chain on $\hat{S}$ are recurrent. Let us denote the stationary distribution by π, where $\pi(k)$ is the stationary probability of state (k,ϕ), and $\pi^B(k)$ is the stationary probability of (k,B) (and likewise for π^A). The balance equations for $0 < k < K^A$ are

$$\pi(k) = p_A \pi(k+1) + p_\beta \pi^B(k),$$

$$\pi^B(k) = p_\alpha \pi^B(k-1) + p_B \pi(k),$$

and balance equation for $(0,B)$ yields

$$\pi^B(0) = p_B \pi(0).$$

For any $k \geq 0$, the flow through the cut between $s \leq k$ and $s \geq k+1$ must be zero (see Figure 11), and therefore[48]

$$p_A \pi(k+1) = p_\alpha \pi^B(k).$$

We get that for $k = 0$ we have

$$\pi(1) = \frac{p_\alpha}{p_A} \pi^B(0) = p_B \frac{p_\alpha}{p_A} \pi(0),$$

and for $0 < k < K^A$ we have

$$\pi(k) = p_A \pi(k+1) + p_\beta \pi^B(k) = p_A \pi(k+1) + p_\beta \frac{p_A}{p_\alpha} \pi(k+1)$$

$$= p_A \frac{p_\alpha + p_\beta}{p_\alpha} \pi(k+1) = \frac{p_A}{p_\alpha} \pi(k+1),$$

[48] See Caldentey Kaplan, and Weiss (2009).

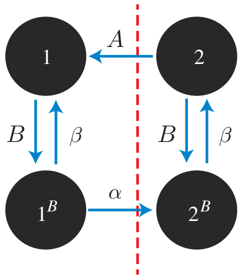

Figure 11. The Cut between $s \leq 1$ and $s \geq 2$

and

$$\pi^B(k) = \frac{p_A}{p_\alpha}\pi(k+1) = \pi(k).$$

By induction, for $0 < k \leq K^A$ we have

$$\pi(k) = p_B\left(\frac{p_\alpha}{p_A}\right)^k \pi(0).$$

The balance equations for (K^A, ϕ) yields

$$\pi(K^A) = (p_\alpha + p_\beta) \times \pi^B(K^A) = \pi^B(K^A),$$

and thus for $0 \leq k \leq K^A$ we have

$$\pi^B(k) = p_B\left(\frac{p_\alpha}{p_A}\right)^k \pi(0).$$

Finally, we calculate $\pi(0)$ by equating the total probability to 1. When $p_A = p_\alpha = p$, we have

$$\begin{aligned} 1 &= \sum_{k=1}^{K^A}\left[\pi(k) + \pi^B(k)\right] + \sum_{k=1}^{K^B}\left[\pi(-k) + \pi^A(-k)\right] + \pi(0) + \pi^B(0) + \pi^A(0) \\ &= 2\sum_{k=1}^{K^A} p_B\pi(0) + 2\sum_{k=1}^{K^B} p_A\pi(0) + \pi(0) + p_B\pi(0) + p_A\pi(0) \\ &= 2\pi(0)\left(p_B K^A + p_A K^B + 1\right), \end{aligned}$$

implying that

$$\pi(0) = \frac{1}{2}\frac{1}{(1-p)K^A + pK^B + 1}.$$

When $p_A \neq p_\alpha$,

$$\begin{aligned} 1 &= \sum_{k=1}^{K^A} \left[\pi(k) + \pi^B(k)\right] + \sum_{k=1}^{K^B} \left[\pi(-k) + \pi^A(-k)\right] + \pi(0) + \pi^B(0) + \pi^A(0) \\ &= 2\sum_{k=0}^{K^A} p_B \left(\frac{p_\alpha}{p_A}\right)^k \pi(0) + 2\sum_{k=0}^{K^B} p_A \left(\frac{p_\beta}{p_B}\right)^k \pi(0) \\ &= 2\pi(0)\left[p_B \frac{p_\alpha \left(\frac{p_\alpha}{p_A}\right)^{K^A} - p_A}{p_\alpha - p_A} + p_A \frac{p_\beta \left(\frac{p_\beta}{p_B}\right)^{K^B} - p_B}{p_\beta - p_B}\right] \\ &= 2\pi(0) \frac{p_A p_\beta \left(\frac{p_\beta}{p_B}\right)^{K^B} - p_B p_\alpha \left(\frac{p_\alpha}{p_A}\right)^{K^A}}{p_A - p_\alpha}, \end{aligned}$$

implying

$$\pi(0) = \frac{1}{2} \frac{p_A - p_\alpha}{p_A p_\beta \left(\frac{p_\beta}{p_B}\right)^{K^B} - p_B p_\alpha \left(\frac{p_\alpha}{p_A}\right)^{K^A}},$$

which converges to the former expression when $p_\alpha \to p_A$. ∎

REFERENCES

Adan, Ivo, Ana Bušić, Jean Mairesse, and Gideon Weiss. 2018. "Reversibility and Further Properties of FCFS Infinite Bipartite Matching." *Mathematics of Operations Research* 43 (2): 598–621.

Adan, Ivo, and Gideon Weiss. 2012. "Exact FCFS Matching Rates for Two Infinite Multitype Sequences." *Operations Research* 60 (2): 475–89.

Agarwal, Nikhil, Itai Ashlagi, Michael A. Rees, Paulo J. Somaini, and Daniel C. Waldinger. 2019. "Equilibrium Allocations under Alternative Waitlist Designs: Evidence from Deceased Donor Kidneys." NBER Working Paper 25607.

Akbarpour, Mohammad, Shengwu Li, and Shayan Oveis Gharan. 2020. "Thickness and Information in Dynamic Matching Markets." *Journal of Political Economy* 128 (3): 783–815.

Altman, Eitan, and Nahum Shimkin. 1998. "Individual Equilibrium and Learning in Processor Sharing Systems." *Operations Research* 46 (6): 776–84.

Anderson, Ross, Itai Ashlagi, David Gamarnik, and Yash Kanoria. 2017. "Efficient Dynamic Barter Exchange." *Operations Research* 65 (6): 1446–59.

Arnosti, Nick, and Peng Shi. 2017. "Design of Lotteries and Waitlists for Affordable Housing Allocation." Unpublished.

Ashlagi, Itai, Maximilien Burq, Patrick Jaillet, and Vahideh Manshadi. 2019. "On Matching and Thickness in Heterogeneous Dynamic Markets." *Operations Research* 67 (4): 927–49.

Asker, John, Allan Collard-Wexler, and Jan De Loecker. 2014. "Dynamic Inputs and Resource (Mis) Allocation." *Journal of Political Economy* 122 (5): 1013–63.

Baccara, Mariagiovanna, SangMok Lee, and Leeat Yariv. 2020. "Optimal Dynamic Matching." *Theoretical Economics* 15 (3): 1221–78.

Barzel, Yoram. 1974. "A Theory of Rationing by Waiting." *Journal of Law and Economics* 17 (1): 73–95.

Bergemann, Dirk, and Stephen Morris. 2005. "Robust Mechanism Design." *Econometrica* 73 (6): 1771–1813.

Bergemann, Dirk, and Maher Said. 2011. "Dynamic Auctions." *Wiley Encyclopedia of Operations Research and Management Science*, edited by James J. Cochran, Louis A. Cox Jr., Pinar Keskinocak, Jeffrey P. Kharoufeh, and J. Cole Smith. New York: John Wiley & Sons. https://doi.org/10.1002/9780470400531.eorms0270.

Bloch, Francis, and David Cantala. 2017. "Dynamic Assignment of Objects to Queuing Agents." *American Economic Journal: Microeconomics* 9 (1): 88–122.

Caldentey, René, Edward H. Kaplan, and Gideon Weiss. 2009. “FCFS Infinite Bipartite Matching of Servers and Customers.” *Advances in Applied Probability* 41 (3): 695–730.

Carlton, Dennis W. 1977. “Peak Load Pricing with Stochastic Demand.” *American Economic Review* 67 (5): 1006–10.

Carlton, Dennis W. 1978. “Market Behavior with Demand Uncertainty and Price Inflexibility.” *American Economic Review* 68 (4): 571–87.

Che, Yeon-Koo, and Olivier Tercieux. 2020. “Optimal Queue Design.” Unpublished.

Chicago Housing Authority. 2016. “Frequently Asked Questions.” https://www.thecha.org/help/faqs.

Das, Sanmay, John P. Dickerson, Zhuoshu Li, and Tuomas Sandholm. 2015. “Competing Dynamic Matching Markets.” In *AMMA: Auctions, Market Mechanisms, and Their Applications,* Vol. 112, edited by Scott Duke Kominers and Lirong Xia, 2–11. Red Hook, NY: Curran Associates, Inc.

De Vany, Arthur. 1976. “Uncertainty, Waiting Time, and Capacity Utilization: A Stochastic Theory of Product Quality.” *Journal of Political Economy* 84 (3): 523–41.

Doval, Laura. 2015. “A Theory of Stability in Dynamic Matching Markets.” Unpublished.

Doval, Laura, and Balázs Szentes. 2018. “On the Efficiency of Queueing in Dynamic Matching Markets.” Unpublished.

Forbes. 2007. “Toughest NFL Waiting Lists.” *Forbes,* September 7. https://www.forbes.com/2007/09/07/nfl-football-tickets-forbeslife-cx_ls_0907tickets.html?sh=60edc49424c8.

Hassin, Refael, and Moshe Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems.* Vol. 59, New York: Springer Science and Business Media.

Kaplan, Edward H. 1984. “Managing the Demand for Public Housing.” PhD diss. Massachusetts Institute of Technology.

Kaplan, Edward H. 1986. “Tenant Assignment Models.” *Operations Research* 34 (6): 832–43.

Kaplan, Edward H. 1987. “Tenant Assignment Policies with Time-Dependent Priorities.” *Socio-Economic Planning Sciences* 21 (5): 305–10.

Kaplan, Edward H. 1988. “A Public Housing Queue with Reneging and Task-Specific Servers.” *Decision Sciences* 19 (2): 383–91.

Kessler, Judd B., and Alvin E. Roth. 2014. “Getting More Organs for Transplantation.” *American Economic Review* 104 (5): 425–30.

Leshno, Jacob D. 2022. “Replication Data for: Dynamic Matching in Overloaded Waiting Lists.” American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E178142V1.

Levin, Jonathan. 2003. “Relational Incentive Contracts.” *American Economic Review* 93 (3): 835–57.

Lindsay, Coton M., and Bernard Feigenbaum. 1984. “Rationing by Waiting Lists.” *American Economic Review* 74 (3): 404–17.

Little, John D. C. 1961. “A Proof for the Queuing Formula: L = (lambda)W.” *Operations Research* 9 (3): 383–87.

Martin, Stephen, and Peter C. Smith. 1999. “Rationing by Waiting Lists: An Empirical Investigation.” *Journal of Public Economics* 71 (1): 141–64.

Naor, P. 1969. “The Regulation of Queue Size by Levying Tolls.” *Econometrica* 37 (1): 15–24.

New York City Public Housing Authority. 2015. “New York City Housing Authority’s Tenant Selection and Assignment Plan.” https://eshare.nycha.info/RFQ/Property%20Management%20RFP%20Documents/Forms/AllItems.aspx?RootFolder=%2fRFQ%2fProperty%20Management%20RFP%20Documents%2fRFP-66734%2fRFP%2066734%20-%20Exhibit%20D%2fTenant%20Selection%20and%20Assignment%20Plan (October 25, 2022).

Schummer, James. 2021. “Influencing Waiting Lists.” *Journal of Economic Theory*: 105263.

Su, Xuanming, and Stefanos A. Zenios. 2004. “Patient Choice in Kidney Allocation: The Role of the Queueing Discipline.” *Manufacturing and Service Operations Management* 6 (4): 280–301.

Su, Xuanming, and Stefanos A. Zenios. 2005. “Patient Choice in Kidney Allocation: A Sequential Stochastic Assignment Model.” *Operations Research* 53 (3): 443–55.

Su, Xuanming, and Stefanos A. Zenios. 2006. “Recipient Choice Can Address the Efficiency-Equity Trade-Off in Kidney Transplantation: A Mechanism Design Model.” *Management Science* 52 (11): 1647–60.

Thakral, Neil. 2016. “The Public-Housing Allocation Problem.” Unpublished.

United Network for Organ Sharing (UNOS). 2014. *Annual Report of the US Organ Procurement and Transplantation Network and the Scientific Registry of Transplant Recipients*. Richmond, VA: United Network for Organ Sharing.

United Network for Organ Sharing (UNOS). 2016. “UNOS Data.” Organ Procurement and Transplantation Network. https://optn.transplant.hrsa.gov/data/view-data-reports/national-data/# (accessed December 2016).

Ünver, M. Utku. 2010. “Dynamic Kidney Exchange.” *Review of Economic Studies* 77 (1): 372–414.

van Dijk, Winnie. 2019. “The Socio-Economic Consequences of Housing Assistance.” Unpublished.

Vasicek, Oldrich A. 1977. “An Inequality for the Variance of Waiting Time under a General Queuing Discipline.” *Operations Research* 25 (5): 879–84.

Verdier, Valentin, and Carson Reeling. 2022. “Welfare Effects of Dynamic Matching: An Empirical Analysis.” *Review of Economic Studies* 89 (2): 1008–37.

Waldinger, Daniel. 2021. “Targeting In-Kind Transfers through Market Design: A Revealed Preference Analysis of Public Housing Allocation.” *American Economic Review* 111 (8): 2660–96.

Wiesner, Russell, Erick Edwards, Richard Freeman, Ann Harper, Ray Kim, Patrick Kamath, Walter Kremers et al. 2003. “Model for End-Stage Liver Disease (MELD) and Allocation of Donor Livers.” *Gastroenterology* 124 (1): 91–96.