Improving Affordability and Accessibility for Socially-Beneficial Services via Government Incentives

Xiaoyan Zhao Venus Lo Stephen Shum[1]

[1] *Department of Decision Analytics and Operations, City University of Hong Kong*

April 29, 2026

Abstract

Many socially-beneficial services, like dental care, suffer from a dual challenge: low affordability for citizens and long waiting times due to insufficient provider capacity. We model a government's problem of designing subsidy policies to address this issue. We analyze three financial interventions: consumer vouchers, which are on the citizen side; and two provider-side subsidies aimed at increasing capacity: fee-for-service subsidies and downtime rebates. Using a continuous-time Markov chain model, we capture the dynamic interactions between citizens and the service provider in response to these policies. Our findings yield several structural insights into subsidy design. We prove that subsidizing idle service capacity (a form of risk mitigation) always outperforms fee-for-service subsidies (a form of reward enhancement) in terms of cost-effectiveness. However, the choice between citizen-side and provider-side policies depends critically on the primary system bottleneck. To further improve the government's cost efficiency, we propose a mixed-subsidy policy. Although optimizing this policy is intractable, we develop an algorithm to find near-optimal solutions. Numerical experiments demonstrate that a mixed policy combining consumer vouchers with idle-time rebates can offer substantial cost savings compared to the best single-subsidy approach, highlighting the potential efficiency gains of a more integrated subsidy structure.

Keywords: Accessibility and affordability of services; Dynamic model; Government policy; Government subsidies

1 Introduction

Many socially beneficial services are delayed because of a dual challenge: affordability issues for citizens and insufficient service capacity. On the demand side, about 41% of UK citizens delay dental visits due to price (Foundation, 2024). On the supply side, even those who can afford services face long queues; for instance, pediatric dental surgery wait times in some regions can reach 80 weeks (Green et al., 2023). This delay is often caused by low service capacity due to high operating costs for providers, such as dentists in Northern Ireland who increasingly found involvement with the National Health Service (NHS) financially unsustainable (Telford, 2023). Similar challenges plague other social services, such as respite care for children with special needs, where capacity shortages leave families waiting even after securing funding (Kemp, 2024). Improving affordability and shortening waiting times for these essential services is thus a critical public policy challenge.

To tackle these issues, governments typically intervene using various subsidy designs, targeting either the demand or supply side. Some governments target citizens directly to lower financial barriers, an approach we refer to as the *consumer-voucher policy*. Examples include Singapore's Community Health Assist Scheme for dental care, Hong Kong's Elderly Health Care Vouchers for private healthcare, and funding for respite care in Ontario, Canada (Ministry of Health, Singapore, 2024; Department of Health, HKSAR, 2024; Government of Ontario, 2024). Similarly, the U.S. has moved to cap daycare costs for low-income families and those with children with disabilities (U.S. Department of Health and Human Services, 2024).

Furthermore, some governments may offer incentives to service providers to expand capacity by increasing their profit margins, which we define as the *fee-for-service policy*. For instance, the UK's NHS offers additional payments to dentists for treating new patients, and Western Australia provides extra funding to private schools enrolling students with special needs (Department of Health and Social Care, 2024; Government of Western Australia, 2024). Other examples include grants to improve elderly care capacity (Government Grants Management Function, UK, 2024), or subsidizing the cost of fuel for eligible local bus services in order to encourage bus operators to offer services in unprofitable locations (Department of Health and Social Care and Whately, 2023).

Alternatively, governments can mitigate the provider's operational risks, particularly the cost of idle capacity, through what we refer to as the *downtime-rebate policy*. This was evident during the COVID-19 pandemic when governments paid to guarantee occupancy rates for quarantine hotels to ensure capacity was available (The Standard, 2020). It is also a key proposal for sustaining rural dentistry, where small patient populations make practices

financially precarious (Dent-Line of Canada Inc., 2024). In such cases, the fee-for-service policy may fail due to insufficient demand.

While these subsidy policies are common, it remains unclear when and why one approach is more cost-effective than another. Subsidizing citizens may boost demand without solving the underlying capacity shortage, whereas subsidizing providers may not be sufficient if services remain unaffordable. Most existing research focuses on one-time consumption, but many social services involve recurring needs where citizens cycle between being served and needing care again. This paper addresses this gap by analyzing the trade-offs between citizen-side and provider-side subsidies in a system with recurring service needs. We aim to provide insights on when it is better to subsidize citizens versus providers. We also explore whether combining two types of incentives can produce sufficient cost-savings to justify the use of a more complex policy structure.

To analyze how government incentives can effectively resolve the bottlenecks of high prices and insufficient service capacity, we develop a queuing-game model where both demand and supply are endogenous. Recognizing that social services are rarely one-off events, we employ a closed system where citizens cycle through a natural lifecycle: they are *satisfied* after receiving care, but inevitably transition back to a *needy* state over time. In this framework, needy citizens strategically decide when to seek service and join the queue as *waiting*, based on out-of-pocket prices. Simultaneously, the service provider acts as a profit-maximizing agent, strategically choosing her service rate (e.g., by investing in staff or equipment) in response to the government subsidies and the resulting demand. To ensure our policy recommendations remain grounded in reality, we extend the baseline model to account for practical complexities such as non-linear capacity costs, moral hazard in elective demand, and the risk of subsidy misreporting.

Our analysis provides a strategic roadmap for policymakers and service managers. A key finding is that the downtime-rebate policy, by mitigating provider risks, consistently outperforms the fee-for-service policy in terms of cost-effectiveness for the government. However, we find that neither the consumer-voucher policy nor the two provider-based policies is universally optimal. Instead, the choice of incentives should be guided by system characteristics such as the provider's gross profit margin, the citizens' responsiveness to price reduction, and the government's target on the fraction of satisfied citizens over the population. We offer a diagnostic framework for policy selection. When a service is unaffordable but the provider is profitable, vouchers are the most effective tool. However, when the provider is unprofitable and capacity is low, downtime rebates are superior. Moreover, we find that as the government's target satisfaction level becomes aggressive, the consumer-voucher policy may be preferred since it can incentivize both the citizens and provider.

We further propose and analyze a mixed-subsidy policy as a strategic tool for governments seeking to simultaneously improve both affordability and accessibility. This integrated approach directly addresses the limitations of single-subsidy policies, offering a more comprehensive solution. For instance, by strategically allocating funds to both parties, i.e., citizens via vouchers and providers via capacity or risk-sharing incentives, it can overcome both price barriers and supply-side bottlenecks. While the optimal design of such a bi-level policy presents computational challenges due to its complexity, we develop an algorithm to identify near-optimal solutions with a provable performance guarantee. This analytical framework enables us to identify conditions under which a mixed policy delivers substantial cost savings compared to the best single-subsidy policy. Crucially, our findings suggest that when both affordability and capacity issues are significant, or the target satisfaction level is high, a carefully balanced mixed subsidy is not only feasible but also demonstrably more efficient than a single instrument. This provides a strategy for policymakers to enhance social welfare in essential services.

Importantly, we show that the policy findings from our baseline model remain valid across several practical challenges that often complicate social service management. For instance, considering a non-linear capacity cost, our policy recommendations hold when expanding capacity becomes increasingly expensive due to limited resources like medical staff. We also consider situations where low prices might lead to an increase in demand for elective care. Furthermore, we address the possibility that providers might misreport their idle time to claim higher rebates by proposing a mixed policy that helps manage these reporting risks. This ensures that our framework offers a reliable basis for policy design in complex social service settings.

The remainder of this paper is organized as follows. In Section 2, we present our literature review. Section 3 presents the model and Section 4 analyzes the single-subsidy policies. Section 5 focuses on the mixed-subsidy policy and discusses the potential cost savings. Section 6 discusses several extensions. Finally, Section 7 provides concluding remarks.

2 Literature Review

Our paper is closely related to the stream of operations management literature that investigates subsidy policies to generate positive social outcomes. We focus our review on works that utilize models to study subsidies for improving accessibility and affordability of social services. We divide these works based on the recipients of the subsidies.

The first stream of literature studies policies that provide subsidies directly to citizens. Andritsos and Aflaki (2015) use an M/G/1 queue to study the competition between a for-

profit hospital and a not-for-profit hospital when patients may receive a reimbursement from an external source of fund. Qian and Zhuang (2017) and Qian et al. (2017) use M/M/1 queues to study subsidy policies which encourage citizens to use private hospitals in order to reduce the waiting time in public hospitals. Siddiq et al. (2022) examine two ways to implement subsidy programs for public transportation systems when the government wants to meet a target on the program's effectiveness while minimizing its cost. They propose a direct scheme which funds the subsidies by charging a road congestion fee from drivers, but this program can be costly and may hurt commuters' welfare. As a less costly alternative, they propose an indirect scheme in which the government works with a for-profit service provider who funds the subsidy in return for higher demand and revenue.

The second stream focuses on policies that provide subsidies to the service provider in order to resolve supply-side issues. Hua et al. (2016) study a model where customers choose between a for-profit service provider and a not-for-profit service provider based on differences in waiting time and service quality. They model the system by two parallel M/M/1 queues and they show that the government can tax the for-profit service provider and subsidize the not-for-profit service provider to maximize the total social welfare. Çakıcı and Mills (2025) examine the effect of pay-parity policies for a service provider who provides both in-person and virtual medical appointments. The service provider is reimbursed by insurance at different rates for her service in the two channels and can choose to allocate her capacity across the channels. They show that offering the service provider a fixed reimbursement across channels may be detrimental to citizens' welfare.

Our model differs from these works by focusing on a setting where citizens return to the service provider after a period of time. Guo et al. (2019) study returning patients in a healthcare context, where they compare the social welfare from implementing a fee-for-service policy versus a bundled-subsidy policy. These policies affect the service provider's capacity and quality, and hence they also affect the rate at which citizens return for follow-up visits. In contrast to their M/M/1 model, we investigate our policies using a closed queue, which can be applied to a wide range of real-world problems (De Véricourt and Jennings, 2008). Moreover, we do not focus on maximizing welfare. We choose to focus on minimizing the government's cost subject to a desired level of citizens' satisfaction because government budgets tend to focus on reporting the cost to achieve a desired goal. We also consider the use of a mixed-subsidy policy and show that consumer vouchers can complement provider-based incentives to bring substantial benefits over a single-subsidy policy.

There are works that study mixed-subsidy policies in service settings. Mehrotra and Natarajan (2020) analyze the use of subsidies to improve access to high-quality healthcare services. Using an M/M/1 queue, they compare the use of supply-side and demand-side

single-subsidy policies to impact service quality and arrival rate, the latter of which also depends on the quality of service. They show that a mixed-subsidy policy always performs better than single-subsidy policies, and that the performance gap widens as the budget increases. In contrast to their M/M/1 queue, we use a two-dimensional closed queue so that we can model citizens who transition among three states and who need to return for services periodically. We allow the total arrival (demand) rate to fluctuate based on the price of service and the proportion of citizens in each state. Even after we collapse our two-dimensional queue into a one-dimensional queue which focuses on waiting citizens, our model remains challenging and more complex than traditional M/M/1 models. There are other examples of works which study mixed-subsidy policies in the service industry. Using a multi-stage decision model, Arora et al. (2021) analyze mixed-subsidy policies under a limited budget to increase equitable access to childcare services among heterogeneous consumers. They focus on minimizing inequality subject to a constraint on the available budget. Olsder et al. (2023) study optimal allocation of subsidies in the context of treatment for rare diseases. They consider the use of subsidies in the presence of exogenous pricing, similar to regulated pricing, and outcome-based payment for drug manufacturers. They use a game-theoretical model and focus on maximizing welfare subject to a budget constraint. We differ from these works by using a queuing model by which we can model citizens in different states. Furthermore, we reverse the objective and the constraint by focusing on minimizing the government's cost subject to achieving a desired level of satisfaction in the system.

Our optimization problem under a closed queue model is difficult to analyze. Even when we assume that there is one service provider, the problem of finding the optimal mixed-subsidy policy could lead to an intractable optimization problem. As a result, we study the structure of near-optimal policies which have good theoretical performance guarantee. Our algorithm for constructing these near-optimal policies uses a geometric discretization on the space of voucher values, and this technique is more commonly seen in discrete optimization problems (see, for example, Désir et al., 2022). We show that techniques in discrete optimization can be modified for a non-convex optimization problem so that we can find solutions which are guaranteed to be within a desired factor of optimality.

While our paper focuses on recurring services, a vast body of operations management literature investigates subsidy policies for manufactured goods. Many of these works focus on increasing consumers' purchases of green technologies (Cohen et al., 2016; Chemama et al., 2019), health products (Taylor and Xiao, 2014), household goods (Xiao et al., 2020).

More importantly, by analyzing provider-side policies that target the capacity and risk, our work resonates with the growing literature on food and agriculture supply chains, which explores subsidy design and risk-sharing mechanisms for small producers. For instance,

Alizamir et al. (2019) compare price subsidies versus revenue protection in the agriculture industry. They find that protecting farmers against downside risk can be more effective than simple price supports. Tang et al. (2024) analyze the effects of input and output cost-reduction subsidies. They show that low-yield farmers prefer input subsidies to lower upfront costs, whereas high-yield farmers benefit more from output subsidies that improve processing margins. While these studies focus on physical goods, our paper extends these insights to the service context.

There are also a number of papers which compare single-subsidy and mixed-subsidy policies for increasing access to goods. Raz and Ovchinnikov (2015) compare the effectiveness of consumer vouchers versus service rate subsidies, as well as a mixed-subsidy policy, in order to maximize social welfare from increasing sales of a public interest good. Yu et al. (2018) consider similar subsidies, but they focus on two competing providers who change their selling prices in response to the government's policies. The aforementioned works focus on the one-time consumption of a physical good, whereas our model focuses on repeated consumption of a service with waiting time when the service provider is operating at a low service rate.

Since our model focuses on a government which provides subsidies to motivate two sides of a market, it is also related to works in operations management which focus on a "leader" who allocates a limited budget between two parties to increase access to a service. Previous works have explored resource allocation for non-profit organizations which operates different programs and departments. Kotsi et al. (2023) study budget allocation between programs for citizens and administrative departments which help with program efficiency. McCoy and Lee (2014) study resource allocation among different outreach efforts in a healthcare application. Arora et al. (2022) study resource allocation in multi-stage services. Wei et al. (2024) explore resource allocation problem among several service providers to improve accessibility and equity. These works typically consider a two-party system involving either a service provider and citizens, or a funding agency and a service provider. In contrast, our model incorporates three parties: the funding agency (government), the service provider, and the citizens receiving the service. Including the decisions of three parties adds another level of complexity to our model.

3 Model: Citizens, Service Provider, and Government

In this section, we begin with an overview of the interactions between citizens, the service provider, and the government. We then explain how these interactions are modeled as each party's optimal response to the actions of the other parties in our continuous-time Markov chain model (CTMC).

3.1 Overview of Model

We consider a system with N citizens, a service provider, and a government. Citizens need to meet with the service provider on a recurring basis. The standard price of service charged by the service provider is fixed at r per visit. The fixed price reflects the structure of some partially subsidized services; in England, NHS dental treatments are assigned to three groups with pre-determined prices (National Health Service, 2024). The service provider operates a single service station and decides on her service rate μ, which can be interpreted as her capacity. The marginal cost of capacity is c. Then the total cost of operating at a service rate μ is given by $c\mu$, which must be paid regardless of utilization. We use a linear cost structure for tractability. We further consider a convex cost in Appendix C1.

The government may offer a subsidy to the citizens, the service provider, or both parties simultaneously. There are three ways that the government can offer a subsidy. First, the government can run a consumer-voucher policy and offer a voucher of $\kappa \leq r$ to needy citizens for each visit. The voucher allows the citizens to pay $r - \kappa$ for each complete service while ensuring that the service provider continues to earn the full price r. Second, the government can run a fee-for-service policy and pay the service provider η for each citizen served. The fee-for-service policy increases the profitability of the service provider so that she will increase her service rate to see more citizens in a timely manner. Finally, the government can implement a downtime-rebate policy and pay the service provider ζ for each unit of downtime. The purpose of the downtime rebate policy is to reduce the service provider's cost during her downtime, which is her risk of doing business. As such, the service provider should be willing to increase her service rate. We require $\zeta < c$ to avoid a risk-free operation. Let $\theta = (\kappa, \eta, \zeta)$ denote the government's policy. If exactly one of these values is non-zero, then it is a single-subsidy policy. If two or more of these values are non-zero, then it is a mixed-subsidy policy. The base case without any government intervention is represented by a policy with all three values being zero.

3.2 Citizens' Response to the Price of Service

We now focus on one citizen and his transitions among the three states: satisfied, needy, and waiting. We model the transitions of a citizen across the three states using three independent exponential distributions.

A satisfied citizen becomes needy when he requires a visit with the service provider. For example, a citizen could develop a toothache after his last dental visit which took place several months ago. A citizen's satisfaction duration is exponentially distributed with an exogenous rate ϕ, which represents the average frequency of service visits necessitated by

status deterioration.

A needy citizen may delay registering for a visit with the service provider because of the high service price. We assume in the baseline model that citizens' decisions are driven entirely by the service price, but we generalize this assumption in Section 6.2 by considering the joint effect of price and waiting time. Given the government's policy θ, the time until a needy citizen registers for his next visit follows an exponential distribution with demand rate $\lambda_\theta = f(\kappa)$, where $f(\kappa) : [0, r] \to [0, \infty)$ is a strictly increasing function. This implies that the citizen is more willing to access service if he receives a larger voucher. Returning to dental services under NHS, a study showed that 23% of patients avoid the dentist due to the cost despite the subsidized prices, with some patients taking drastic actions to postpone visits such as doing their own dental work at home with household items (Morris, 2023). Our model captures a citizen's postponement of his dental visit with demand rate $\lambda_\theta = f(\kappa)$ to recognize the final price paid by him. While our demand function uses the voucher value κ as input, this is mathematically equivalent to a price-dependent model since the base price r is fixed.

A waiting citizen waits for his turn to be served after he registers for a visit. When it is his turn, his service time follows an exponential distribution with service rate μ. In the aforementioned example of the UK dental services, waiting time for the NHS's dental services could be several months, despite the services being financially inaccessible to some citizens (Green et al., 2023). Our model captures the need to wait longer as the subsidy increases and more citizens seek service in a timely manner via the queue of waiting citizens. In addition, we allow the service provider to increase her service rate in response to the subsidy and the increased demand, as discussed in the next subsection.

A citizen becomes satisfied again after the service. However, issues will occur over time and it is natural for a citizen to require another visit. This is modeled by the deterioration rate and the citizen's transition from being satisfied to needy.

This system of N citizens can be modelled as a CTMC. To make the notations cleaner in this discussion, let $\lambda = \lambda_\theta$. At time t, let $S(t)$ denote the number of satisfied citizens, $D(t)$ denote the number of needy citizens, and $W(t)$ denote the number of waiting citizens. The state of the system is denoted by the number of satisfied citizens and waiting citizens, (s, w), so that the state space is $\mathcal{S} = \{(s, w) : s+w \leq N;\ s, w = 0, 1, \ldots, N\}$. Since there are exactly N citizens, we must have $D(t) = N - S(t) - W(t)$. At a state $(S(t), W(t)) = (s, w)$, one of three events could occur when the system transitions at time $t' > t$. First, if $S(t) > 0$, then one of the s satisfied citizens could develop a toothache and become needy; the system goes to $(S(t'), W(t')) = (s - 1, w)$ according to an exponential distribution with rate $s\phi$. Second, if $S(t) + W(t) < N$, then one of the $N - s - w$ needy citizens visits the service provider and

queues up behind other waiting citizens; the system goes to $(S(t'), W(t')) = (s, w+1)$ with rate $(N-s-w)\lambda$. Finally, if $W(t) > 0$, then one of the w waiting citizens could be served by the service provider and becomes satisfied; the system goes to $(S(t'), W(t')) = (s+1, w-1)$ with rate μ. The state-transition process of the system is illustrated in Figure 1.

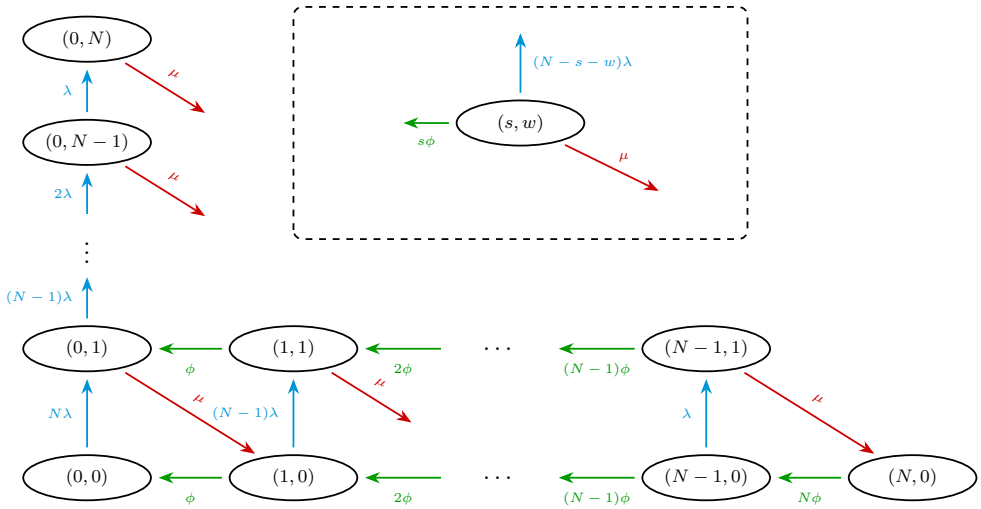


Figure 1: State-transition process. The state pair (s, w) denotes the number of satisfied and waiting citizens.

3.3 Service Provider's Response to the Price, Cost, and Demand Rate of Citizens

The service provider chooses her service rate in response to the citizens' overall demand, her revenue from each visit, and her cost. The total demand rate observed by the service provider depends on the proportion of citizens in each state. The provider decides on a service rate μ to maximize her profit in the steady state. We drop the reference to t in $S(t), W(t)$, and $D(t)$ in future discussions.

For a given pair of service rate μ and demand rate λ, let $Q_{s,w}(\mu, \lambda)$ denote the steady-state probability of being in (s, w). The service provider earns revenue r at a rate of $\mu \cdot (1 - \tilde{Q}_0(\mu, \lambda))$, where $\tilde{Q}_w(\mu, \lambda) := \sum_{s=0}^{N-w} Q_{s,w}(\mu, \lambda)$ is the probability that there are w waiting citizens. Her profit function is $r\mu \cdot (1 - \tilde{Q}_0(\mu, \lambda)) - c\mu$ when there are no subsidies from the government.

Recall that the government could subsidize the service provider via its policy $\theta = (\kappa, \eta, \zeta)$. The first policy affects the service provider via the citizens' demand λ_θ. If the government implements the fee-for-service policy, then she earns η in addition to r with every visit at a rate of $\mu \cdot (1 - \tilde{Q}_0(\mu, \lambda_\theta))$. If the government implements the downtime-rebate policy, then

the provider earns ζ at the rate that she *does not* earn revenue, or $\mu \cdot \tilde{Q}_0(\mu, \lambda_\theta)$. We denote $\Pi_\theta(\mu)$ as the service provider's profit function if she operates at a service rate of μ when the government implements policy θ. The service provider's decision of choosing a service rate to maximize her profit function, in response to the citizens' demand rate and the government's policy, is:

$$\max_{\mu} \ \Pi_\theta(\mu) = \max_{\mu} \{(r+\eta) \cdot \mu \cdot \Big(1 - \tilde{Q}_0(\mu, \lambda_\theta)\Big) + \zeta\mu \cdot \tilde{Q}_0(\mu, \lambda_\theta) - c\mu\}. \tag{1}$$

Let μ_θ be the optimal solution to Problem (1), which corresponds to the service provider's best response to policy θ.

Our model considers a service provider who serves one citizen at a time, or a single-server model. It can be argued that the service provider could increase the number of service stations in the long-run. In the NHS dental service example, however, each dentist operates independently and can choose the number of subsidized patients that she treats (LDC Confederation, 2024). The low price and high cost of practicing have caused dentists in Northern Ireland to refrain from accepting new NHS patients (Telford, 2023), which we represent as a lower service rate in the model. One way that a dentist could increase her service rate would be to reduce turnover time between patients. For example, the dentist could reduce the time to clean tools and complete paperwork, and c would be the additional staffing and technology cost that would allow for faster turnover on one service station. As a more extreme example, it may be too expensive for a hospital to purchase an extra MRI machine, but the service rate can be increased by having additional staff to prepare the next waiting patient to improve turnover. The single-server model is also popular in the literature due to its simpler structure. While the main analysis is conducted on the single-server model, we will extend the model to the multi-server setting in Section 6.1, which is more appropriate if we could hire additional service providers.

3.4 Government's Choice of Policy: Cost and Target

Consistent with typical public policy mandates, the government seeks to meet a target outcome with minimum cost. Specifically, it ensures that at least a fraction $B \in [0, 1]$ of citizens over the population is satisfied, where B is an exogenous and pre-determined target. Mathematically, the goal of the government is to ensure that $\mathbb{E}[S|\ \mu_\theta, \lambda_\theta]/N \geq B$.

Let $C(\theta)$ denote the cost incurred by the government under policy $\theta = (\kappa, \eta, \zeta)$. We consider how each of the three subsidies adds to the government's cost. First, under the consumer-voucher policy, the cost to the government is $\kappa\mu_\theta \cdot (1 - \tilde{Q}_0(\mu_\theta, \lambda_\theta))$. Under the fee-for-service policy, the government pays the service provider η after each visit, and gov-

ernment's cost is equal to the service provider's additional revenue of $\eta\mu_\theta \cdot (1 - \tilde{Q}_0(\mu_\theta, \lambda_\theta))$. In contrast, under the downtime-rebate policy, the government provides a rebate of ζ for each unit of idle time when there are no waiting citizens. The cost to the government is $\zeta\mu_\theta \cdot \tilde{Q}_0(\mu_\theta, \lambda_\theta)$.

Based on the above three cost components, the government's total cost under policy θ is

$$C(\theta) = (\kappa + \eta) \cdot \mu_\theta \cdot \Big(1 - \tilde{Q}_0(\mu_\theta, \lambda_\theta)\Big) + \zeta\mu_\theta \cdot \tilde{Q}_0(\mu_\theta, \lambda_\theta). \tag{2}$$

The government's problem is to choose θ to minimize its cost subject to achieving the desired fraction of satisfied citizens:

$$\min \left\{ C(\theta) \;\middle|\; \mathbb{E}[S|\ \mu_\theta, \lambda_\theta]/N \geq B \right\}. \tag{3}$$

By setting κ, η, and/or ζ to 0, the above model is able to describe both single-subsidy policies and mixed-subsidy policies. We consider the single-subsidy policies in Section 4 and the mixed-subsidy policy in Section 5.

4 Model Analysis and Single-Subsidy Policies

In this section, we first analyze the steady state of the system and determine the service provider's best response to citizens' demand rate and government subsidies. Then we proceed to study the effectiveness of single-subsidy policies.

4.1 Analysis of Model: Citizens and Service Provider

In order to solve the government's optimal policy in Problem (3), we need to solve for $\tilde{Q}_0(\mu_\theta, \lambda_\theta)$ and $\mathbb{E}[S|\ \mu_\theta, \lambda_\theta]$ in closed forms and to understand the service provider's optimal decision. Proposition 1 characterizes the steady state in closed forms, as well as the expected number of citizens in each state when μ and λ are given.

Proposition 1. *For any fixed μ and λ, the stationary probabilities at $(s, w) \in \mathcal{S}$ can be expressed as $Q_{s,w}(\mu, \lambda) = n_{s,w}(\mu, \lambda)/\Gamma(\mu, \lambda)$, where*

$$n_{s,w}(\mu, \lambda) = \frac{N!}{s!(N - s - w)!} \cdot \frac{\lambda^{s+w}}{\phi^s \mu^w}$$

and $\Gamma(\mu, \lambda) = \sum_{s=0}^{N} \sum_{w=0}^{N-s} n_{s,w}(\mu, \lambda)$.

Let $\nu = \frac{\mu}{\phi} + \frac{\mu}{\lambda}$. Then, the probability that there are w waiting citizens can be simplified as

$$\tilde{Q}_w(\mu, \lambda) = \tilde{Q}_w(\nu) = \frac{\nu^{N-w}/(N-w)!}{\sum_{j=0}^{N} \nu^{N-j}/(N-j)!},$$

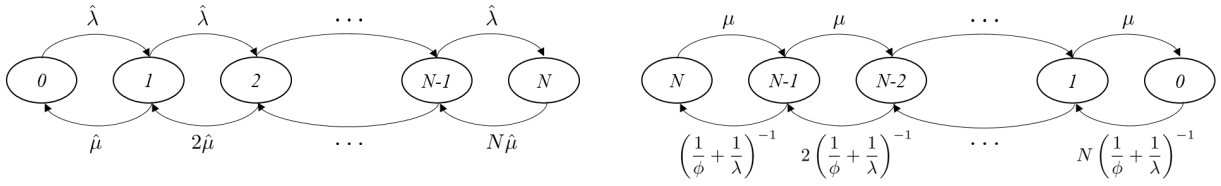
and the expected number of citizens in each state are: $\mathbb{E}[S|\, \mu, \lambda] = \frac{\mu}{\phi} \cdot \left(1 - \tilde{Q}_0(\nu)\right)$, $\mathbb{E}[D|\, \mu, \lambda] = \frac{\mu}{\lambda} \cdot \left(1 - \tilde{Q}_0(\nu)\right)$ *and* $\mathbb{E}[W|\, \mu, \lambda] = N - \nu \cdot \left(1 - \tilde{Q}_0(\nu)\right)$.

Before proceeding, we make a few observations from Proposition 1 regarding the value of ν and its implications on the state of the system. First, the probability that there are w waiting citizens depends on μ and λ only through ν. We write $\tilde{Q}_0(\nu)$ to show that this probability remains constant as long as ν remains unchanged even when λ and μ change. As the expected number of waiting citizens is also affected by μ and λ only through ν, we could write it as $\mathbb{E}[W|\, \nu]$.

Second, the value of $\tilde{Q}_0(\nu)$ can be rewritten as $\tilde{Q}_0(\nu) = \left(\sum_{s=0}^{N} \frac{N!}{(N-s)!} \cdot \nu^{-s}\right)^{-1}$, which increases when ν increases. The formula for $\tilde{Q}_w(\nu)$ is reminiscent of the M/M/N/N queue, or the Erlang-B loss model, except that we have $N - w$ in place of w in the probabilities. This implies that our model in Figure 1 can be collapsed into an M/M/N/N queue if we focus on the waiting citizens. In a standard M/M/N/N queue with N call center staff, callers arrive at a rate of $\hat{\lambda}$ and staff members complete calls at a rate of $\hat{\mu}$. Callers balk if all N staff members are busy. In place of the call centre staff, we have N citizens. A waiting citizen corresponds to a non-busy staff member: instead of a caller arriving to take up the time of a call-centre staff, we have the service provider arrive to serve a waiting citizen, so $\hat{\lambda}$ corresponds to μ. A non-waiting citizen corresponds to a busy staff member: instead of completing a call, a "toothache" develops in each non-waiting citizen. The time between his last visit and completion of his current visit corresponds to the call time, so that $\hat{\mu}$ corresponds to $\left(\frac{1}{\phi} + \frac{1}{\lambda}\right)^{-1}$. The relationship between the two models is shown in Figure 2. The relationship implies that ν corresponds to the utilization in a queuing model, which drives whether a system is typically overloaded or idle. This correspondence explains the importance of ν in studying the probability of observing w waiting citizens as well as the expected fraction of waiting citizens in the system.

The above observations can help us prove that Problem (1) has a unique optimal service rate. Given policy θ, the demand rate λ_θ is fixed from the perspective of the service provider. A change of variable allows us to consider a maximization problem over ν instead of μ:

$$\max_{\mu}\ \Pi_\theta(\mu) = \left(\frac{1}{\phi} + \frac{1}{\lambda_\theta}\right)^{-1} \cdot \max_{\nu}\ \left\{(r + \eta - \zeta) \cdot \nu \cdot \left(1 - \tilde{Q}_0(\nu)\right) - (c - \zeta) \cdot \nu\right\}, \quad (4)$$



(a) State denotes # of busy call-center staff. **(b)** State denote # of waiting citizens.

Figure 2: Our system collapses into an M/M/N/N queue. The busy staff members correspond to the non-waiting citizens. Caller arrival corresponds to the service provider completing service on a citizen, whereas call completion corresponds to a "toothache" afflicting a citizen.

which is a concave maximization problem following a result by Krishnan (1990) for the Erlang-B loss model. Since there exists a unique optimal solution ν_θ, we can recover the optimal service rate μ_θ by reversing the change of variable. Lemma 2 formally proves the existence of a unique ν_θ and μ_θ and identifies a clean optimality condition.

Lemma 2. *Suppose we are given policy θ. There is a unique value of ν_θ which satisfies*

$$\frac{r+\eta-c}{r+\eta-\zeta} = \tilde{Q}_0(\nu_\theta) \cdot (1 + \mathbb{E}[W|\ \nu_\theta]), \tag{5}$$

and the service provider's best response μ_θ can be recovered as $\mu_\theta = \nu_\theta \cdot \left(\frac{1}{\phi} + \frac{1}{\lambda_\theta}\right)^{-1}$. Moreover, ν_θ is increasing with the ratio $\frac{r+\eta-c}{r+\eta-\zeta}$.

Since $\zeta < c$, Lemma 2 implies that the utilization rate ν_θ as well as the service provider's optimal service rate increases when η or ζ increases. Furthermore, if the government implements a pure consumer-voucher policy, κ does not affect the optimality condition in Eq. (5). This implies that the provider's optimal utilization ν_θ remains constant at the baseline level ν_0. However, as we will show later in Lemma 3, this does not mean the provider is inactive; rather, she scales up her service rate μ_θ to perfectly match the increased demand, thereby maintaining her optimal utilization.

Using this optimality condition, we are now ready to compare the three single-subsidy policies. As we will compare various policies against the no-subsidy policy, $\theta_0 = (0, 0, 0)$, we use subscript of 0 to denote this benchmark case with λ_0, μ_0 and ν_0.

4.2 Analysis of Single-Subsidy Policies

We begin by understanding the effect of the consumer-voucher policy. This policy is different from the provider-based policies because it has an impact on the citizens' decision, and at the same time, it has an indirect effect on the service provider's decision. Then, we study

and compare the two provider-based policies, before finally comparing the consumer-voucher policy against the provider-based policies.

Consumer-Voucher Policy

Under the consumer-voucher policy, the government implements $\theta = (\kappa, 0, 0)$. When $\lambda_\theta = f(\kappa)$ increases due to larger vouchers, the expected number of needy citizens should decrease because they will register for their next visit at a faster rate. In response, we should expect the service provider to increase her service rate in order to meet the higher demand. Nevertheless, it is not obvious whether the higher service rate will be sufficient to reduce the number of waiting citizens. In fact, as we will establish in Lemma 3, the effects of a higher demand and a higher capacity perfectly offset each other, such that the system utilization and the resulting queue length remain unchanged.

Lemma 3. *Suppose the government implements the consumer-voucher policy $\theta = (\kappa, 0, 0)$. As κ increases, the expected number of satisfied citizens, $\mathbb{E}[S|\ \mu_\theta, \lambda_\theta]$, increases. The expected number of needy citizens, $\mathbb{E}[D|\ \mu_\theta, \lambda_\theta]$, decreases. The expected number of waiting citizens, $\mathbb{E}[W|\ \mu_\theta, \lambda_\theta]$, does not change.*

The consumer-voucher policy directly reduces the expected number of needy citizens by encouraging them to take action. However, as noted earlier, the optimality condition in Lemma 2 tells us that ν_θ remains constant. From a managerial standpoint, the service provider's profit depends on her utilization. Since her gross profit margin does not change under the consumer-voucher policy, she is motivated to increase her service rate only to maintain the same probability of downtime. As a result, the expected number of waiting citizens remains the same. Thus, the consumer-voucher policy should be applied only if there are many needy citizens who cannot afford to visit the service provider on a timely basis, but it cannot be expected to motivate the service provider to work at a sufficiently high service rate and resolve supply-side issues.

Observe that the government's cost under the consumer-voucher policy can be simplified as $C(\theta) = \kappa\phi \cdot \mathbb{E}[S|\ \mu_\theta, \lambda_\theta]$. Using this observation and Lemma 3, we obtain the next corollary, which gives us a simple way to characterize the optimal consumer voucher.

Corollary 4. *As the voucher κ increases, the government's cost increases. Hence, either the consumer-voucher policy is infeasible, i.e., $\mathbb{E}[S|\ \mu_\theta, \lambda_\theta]/N < B$ for all $\theta = (\kappa, 0, 0)$ when $\kappa \leq r$, or the optimal consumer-voucher policy θ^* satisfies $\mathbb{E}[S|\ \mu_{\theta^*}, \lambda_{\theta^*}]/N = B$.*

Building on these results, we examine how the government's cost changes when aiming to increase the fraction of satisfied citizens through larger vouchers. Consider a baseline

policy $\theta = (\kappa, 0, 0)$, possibly with $\kappa = 0$. For some $\psi > 1$, suppose that the government wants to estimate the additional funds needed to increase the number of satisfied citizens to $\psi \cdot \mathbb{E}[S|\ \mu_\theta, \lambda_\theta]$. The next proposition shows that the government would need to ensure that the demand rate increases by a factor of at least ψ, and the goal may be unattainable if the government is aggressive and sets a large ψ.

Proposition 5. *Suppose the government is currently implementing a policy $\theta = (\kappa, 0, 0)$. For some $\psi > 1$, the government wants to consider an alternative policy $\theta' = (\kappa', 0, 0)$ such that $\mathbb{E}[S|\ \mu_{\theta'}, \lambda_{\theta'}] = \psi \cdot \mathbb{E}[S|\ \mu_\theta, \lambda_\theta]$. If we define $\gamma = \frac{\psi}{1-(\psi-1)\cdot\frac{\lambda_\theta}{\phi}}$, then:*

1. *$\lambda_{\theta'} = \gamma\lambda_\theta$, and the government's new target is infeasible if $\psi > 1 + \frac{\phi}{\lambda_\theta}$,*
2. *If $\kappa > 0$, then the cost of the new policy is at least $C(\theta') = \psi \cdot C(\theta) \cdot \frac{f^{-1}(\gamma\lambda_\theta)}{f^{-1}(\lambda_\theta)}$.*

Proposition 5 highlights two practical implications of the consumer-voucher policy. First, the policy becomes progressively less effective. To achieve a ψ-fold increase in satisfaction, the government must stimulate a disproportionately larger increase in demand ($\gamma > \psi$ and is convexly increasing in ψ). This gap widens as the target becomes more ambitious. In fact, γ is only well-defined if $\psi < 1 + \frac{\phi}{\lambda_\theta}$. It indicates that if the current demand rate is already high, then it may be infeasible to achieve the new goal by using consumer vouchers alone.

Second, the policy becomes disproportionately more expensive. Any improvement is at least ψ times more costly than the baseline. This cost escalation is driven by how consumers react to the total price paid for service. Specifically, if the voucher has decreasing marginal attractiveness ($f(\kappa)$ is concave), the required subsidy cost grows much faster than the government's targeted improvement ψ.

Provider-Based Policies

Next, we investigate the fee-for-service policy $(0, \eta, 0)$ and the downtime-rebate policy $(0, 0, \zeta)$. Since $\kappa = 0$, both policies observe the same demand rate $\lambda_\theta = f(0)$. The following lemma shows the affect of the provider-based policies on citizens in different states.

Lemma 6. *Suppose the government implements the fee-for-service policy $\theta = (0, \eta, 0)$ (*resp. *the downtime-rebate policy $\theta = (0, 0, \zeta)$). As η (*resp. *ζ) increases, the expected number of satisfied and needy citizens, $\mathbb{E}[S|\ \mu_\theta, \lambda_\theta]$ and $\mathbb{E}[D|\ \mu_\theta, \lambda_\theta]$, increase. The expected number of waiting citizens, $\mathbb{E}[W|\ \mu_\theta, \lambda_\theta]$, decreases.*

Lemma 6 tells us that both policies directly stimulate the capacity investment thereby reducing the number of waiting citizens. However, since the out-of-pocket price for citizens

remains high, satisfied citizens still transition to the needy state at the same rate as before and continue to delay visits. Consequently, the proportion of needy citizens in the system actually increases.

Similar to Corollary 4, we can show that both policies increase in cost as the size of the subsidy increases.

Corollary 7. *As the fee for service η (*resp. *rebate ζ) increases, the government's cost increases. Hence, either $\mathbb{E}[S|\ \mu_\theta, \lambda_\theta]/N < B$ for all $\theta = (0, \eta, 0)$ (*resp. *$\theta = (0, 0, \zeta)$ with $\zeta < c$), or the optimal fee-for-service policy θ^* (*resp. *downtime-rebate policy) satisfies $\mathbb{E}[S|\ \mu_{\theta^*}, \lambda_{\theta^*}]/N = B$.*

Quantifying the cost change for provider-based policies is more complex, as these incentives endogenously influence the service rate μ, which often lacks a closed-form expression. Nevertheless, we demonstrate that different provider-based policies exhibit a similar structural property. Specifically, by invoking the optimality condition from Lemma 2, we define $g_\eta(\nu) = \frac{c}{1-\tilde{Q}_0(\nu)\cdot(1+\mathbb{E}[W|\ \nu])} - r$ as the fee-for-service subsidy required to achieve ν when $\zeta = 0$. Similarly, we define $g_\zeta(\nu) = r - \frac{r-c}{\tilde{Q}_0(\nu)\cdot(1+\mathbb{E}[W|\ \nu])}$ as the downtime rebate required to achieve ν when $\eta = 0$. This allows us to obtain a result analogous to Proposition 5.

Proposition 8. *Suppose the government is currently implementing a policy $\theta = (0, \eta, 0)$ (*resp. *$(0, 0, \zeta)$). For some $\psi > 1$, the government wants to consider an alternative policy $\theta' = (0, \eta', 0)$ (*resp. *$(0, 0, \zeta')$) such that $\mathbb{E}[S|\ \mu_{\theta'}, \lambda_{\theta'}] = \psi \cdot \mathbb{E}[S|\ \mu_\theta, \lambda_\theta]$. Define $\alpha = \frac{\psi\cdot\left(1-\tilde{Q}_0(\nu_\theta)\right)}{1-\psi\cdot\tilde{Q}_0(\nu_\theta)}$. Then:*

1. *$\mu_{\theta'} > \alpha \cdot \mu_\theta$, and the government's target is infeasible if $\psi \geq \left(\tilde{Q}_0(\nu_\theta)\right)^{-1}$;*
2. *If the fee-for-service policy is chosen and $\eta > 0$, then $C(\theta') > \psi \cdot C(\theta) \cdot g_\eta(\alpha\nu_\theta)/g_\eta(\nu_\theta)$;*
3. *If the downtime-rebate policy is chosen and $\zeta > 0$, then $C(\theta') > \alpha\psi \cdot C(\theta) \cdot g_\zeta(\alpha\nu_\theta)/g_\zeta(\nu_\theta)$.*

Proposition 8 reveals that the efficiency of provider-based policies depends heavily on the provider's current state. When the system is already very busy ($\tilde{Q}_0(\nu_\theta) \to 0$), a provider subsidy is highly effective. Achieving a target ψ-fold increase in satisfied citizens requires only a roughly ψ-fold increase in the service rate. The policy works efficiently because the new capacity is immediately utilized. Conversely, when the system is often idle ($\tilde{Q}_0(\nu_\theta) \to \psi^{-1}$), a provider subsidy becomes wasteful. To achieve the same ψ-fold increase in satisfied citizens, the government must incentivize a disproportionately larger (α-fold where $\alpha\mu_\theta \to \infty$ as $\tilde{Q}_0(\nu_\theta) \to \psi^{-1}$) increase in the service rate that often remains unused, leading to significant fiscal inefficiency.

Moreover, provider-based subsidies face the same challenge of diminishing returns as consumer vouchers. In fact, if we consider α as a function $\alpha = \alpha(\psi)$, then $\alpha(\psi) > \psi$ and $\alpha(\psi)$ is a convex and strictly increasing function on $1 < \psi < \left(\tilde{Q}_0(\nu_\theta)\right)^{-1}$. In general, part 1 of the proposition tells us that the service rate needs to increase by a factor which is significantly larger than ψ in order to get a corresponding increase in the number of satisfied citizens.

Given that both provider-based policy share similar qualitative effects, which policy is less expensive to implement? The latter parts of Proposition 8 say that the downtime-rebate policy increases the government's cost by a factor of exceeding ψ^2, which might initially appear more expensive and less attractive than the fee-for-service policy. However, the following lemma shows the interesting result that the downtime-rebate policy is guaranteed to result in a lower cost to the government.

Lemma 9. *Suppose policies $\theta_\eta = (0, \eta, 0)$ and $\theta_\zeta = (0, 0, \zeta)$ can both achieve $\mathbb{E}[S|\ \mu_{\theta_\eta}, \lambda_{\theta_\eta}] = \mathbb{E}[S|\ \mu_{\theta_\zeta}, \lambda_{\theta_\zeta}] = \psi \cdot \mathbb{E}[S|\ \mu_0, \lambda_0]$ for some $\psi > 1$. Then the fee-for-service policy will incur higher cost than the downtime-rebate policy for the government: $C(\theta_\eta) \geq C(\theta_\zeta)$. Furthermore, the values of the subsidies satisfy $\zeta = c \cdot \frac{\eta}{(r-c)+\eta}$, and ζ grows sublinearly in η.*

Lemma 9 establishes a powerful and universally applicable principle for designing supply-side subsidies: mitigating downside risk is more cost-effective for the government than enhancing upside reward. Fundamentally, the downtime-rebate policy addresses the core drivers of under-investment in service capacity by shifting the strategic focus toward risk mitigation. Unlike a fee-for-service subsidy, which merely increases the upside of being busy, the downtime-rebate functions as a risk-sharing mechanism that insures the provider against the financial losses of underutilization. This framing of idle capacity as a direct loss makes its avoidance a primary focus for managers, a behavior consistent with findings in behavioral operations management (Kai-Ineman et al., 1979; Bendoly et al., 2006). Furthermore, this targeted approach aligns with the public finance principle of marginal subsidization. It avoids wasteful spending on services the provider would have profitably offered anyway, thereby directing government funds exclusively to the undesirable states where behavior needs to change.

While the downtime-rebate policy is theoretically superior, it presents challenges from an implementation perspective in practice. It requires precise knowledge of the provider's true service rate and idle time, which are susceptible to misreporting. In contrast, the fee-for-service policy is easier to audit and more common in practice. The suitability of the downtime-rebate policy depends on the application, and it has been successfully applied in the context of the government's guarantee on minimum occupancy for COVID quarantine

hotels in Hong Kong (The Standard, 2020). Thus, we recommend downtime rebates primarily in contexts where suitable monitoring controls can mitigate the risk of fraudulent reporting.

Comparison of Consumer-Voucher Policy and Provider-Based Policies

The choice between incentivizing citizens or providers depends on which party's response most effectively resolves the system bottleneck. To compare these approaches formally, we say that policy θ_1 *dominates* policy θ_2 if it achieves at least the same satisfaction level at a lower cost; i.e., $\mathbb{E}[S|\ \mu_{\theta_1}, \lambda_{\theta_1}] \geq \mathbb{E}[S|\ \mu_{\theta_2}, \lambda_{\theta_2}]$ and $C(\theta_1) \leq C(\theta_2)$.

Suppose the government seeks to increase the fraction of satisfied citizens by a factor of $\psi > 1$ relative to the base case of $(0, 0, 0)$ with λ_0 and μ_0. Let $\theta_\kappa = (\kappa, 0, 0)$ be the specific consumer-voucher policy which achieves this target. As established in Proposition 5, the demand rate under θ_κ must satisfy $\lambda_{\theta_\kappa} = \gamma\lambda_0$, where γ was previously defined. We the investigate whether a fee-for-service policy, $\theta_\eta = (0, \eta, 0)$, can dominate θ_κ by considering the subsidy level $\eta = g_\eta(\gamma\mu_0)$ required to scale the service rate by the same factor. The following proposition identifies the conditions under which such dominance occurs.

Proposition 10. *Let $\theta_\kappa = (\kappa, 0, 0)$ be a consumer-voucher policy that achieves a demand rate of $\lambda_{\theta_\kappa} = \gamma\lambda_0$. Let $\theta_\eta = (0, \eta, 0)$ be a fee-for-service policy that achieves a service rate of $\mu_{\theta_\eta} = \gamma\mu_0$.*

1. *If $\tilde{Q}_0(\mu_0, \lambda_0) \in \left[\frac{\lambda_0}{\lambda_0+\phi}, 1\right]$ and $\kappa \leq \eta \cdot \frac{\gamma\lambda_0+\phi}{\lambda_0+\phi} \cdot \frac{1}{(\gamma^N-1)\tilde{Q}_0(\nu_0)+1}$, then no fee-for-service policy can dominate the consumer-voucher policy θ_κ.*
2. *If $\tilde{Q}_0(\mu_0, \lambda_0) \in \left[0, \frac{\lambda_0}{\lambda_0+\phi} \cdot \frac{\gamma-1}{\gamma^N-1}\right]$ and $\kappa \geq \eta \cdot \frac{\gamma\lambda_0+\phi}{\lambda_0+\phi} \cdot \frac{1}{(\gamma-1)\tilde{Q}_0(\nu_0)+1}$, then there exists a fee-for-service policy which dominates the consumer-voucher policy θ_κ.*

Part 1 of Proposition 10 describes a scenario where the provider's downtime probability $\tilde{Q}_0(\nu_0)$ is close to 1, signifying a system where citizens rarely face long queues upon registration. Such a state typically arises when the service provider operates with a high gross profit margin, incentivizing a high service rate even without government intervention. Under these conditions, the second part of the result suggests that a relatively smaller voucher value is sufficient to stimulate demand and meet the target. Consequently, when the provider is already profitable, a consumer-voucher policy dominates the fee-for-service policy because the cost of further incentivizing the provider outweighs the gains from incremental capacity.

In contrast, Part 2 of the proposition identifies cases where $\tilde{Q}_0(\nu_0)$ is close to 0, suggesting that citizens are likely to encounter high waiting times. This situation corresponds to a low-profit environment where the provider is unwilling to sustain any downtime, thereby limiting service capacity. Here, the voucher required to overcome extreme unaffordability

becomes more expensive for the government than a fee-for-service payment that directly expands capacity. Thus, when a system is plagued by low profitability of the provider, the government can achieve its targets more efficiently though a supply-side subsidy.

These conditions define the clear-cut boundaries for optimal policy selection. The 'gray area' not covered by the proposition represents intermediate scenarios where the system suffers from a mix of both affordability and capacity issues. Interestingly, our analysis shows that this gray area shrinks as the government's target for improvement is small (γ is small). In other words, for incremental improvements, the choice of policy is less critical. However, for ambitious, large-scale interventions, choosing the right policy tool becomes paramount to avoid significant waste of public funds.

We obtain analogous results for comparing the downtime-rebate policy with the consumer-voucher policy in Proposition 11.

Proposition 11. *Given $\gamma > 1$, let $\theta_\kappa = (\kappa, 0, 0)$ be a consumer-voucher policy that achieves a demand rate of $\lambda_{\theta_\kappa} = \gamma\lambda_0$. Let $\theta_\zeta = (0, 0, \zeta)$ be a downtime-rebate policy that achieves a service rate of $\mu_{\theta_\zeta} = \gamma\mu_0$.*

1. *If $\tilde{Q}_0(\mu_0, \lambda_0) \in \left[\frac{\lambda_0}{\lambda_0+\phi}, 1\right]$ and $\kappa \leq \zeta \cdot \frac{\gamma\lambda_0+\phi}{\lambda_0+\phi} \cdot \frac{1}{(\gamma-1)\tilde{Q}_0(\nu_0)+1} \cdot \frac{\gamma\lambda_0}{\phi}$, then no downtime-rebate policy can dominate the consumer-voucher policy θ_κ.*
2. *If $\tilde{Q}_0(\mu_0, \lambda_0) \in \left[0, \frac{\lambda_0}{\lambda_0+\phi} \cdot \frac{\gamma-1}{\gamma^N-1}\right]$ and $\kappa \geq \zeta \cdot \frac{\gamma\lambda_0+\phi}{\lambda_0+\phi} \cdot \frac{1}{(\gamma^N-1)\tilde{Q}_0(\nu_0)+1} \cdot \frac{\gamma\lambda_0}{\phi}$, then there exists a downtime-rebate policy which dominates the consumer-voucher policy θ_κ.*

Although the analytical structure remains similar to Proposition 10, the introduction of the factor $\gamma\lambda_0/\phi$ in the conditions suggests a revised dependence on the scaling factor γ. This nuanced difference implies that the range of scenarios where either policy has a clear advantage expands when downtime rebates are used as the provider-side instrument, further highlighting the strategic value of risk-sharing incentives.

To conclude, although each single-subsidy policy has their dominance under some conditions, our analysis reveals a fundamental structural limitation shared by all three policy instruments: the total subsidy cost exhibits a convex and increasing pattern relative to the social target. Whether the government intervenes on the demand side via consumer vouchers (Proposition 5) or on the supply side via provider subsidies (Proposition 8), it eventually encounters diminishing marginal returns. This implies that achieving more ambitious satisfaction targets through any single policy becomes prohibitively expensive, as the marginal cost of improvement escalates rapidly. Such a universal inefficiency of single subsidies provides the primary economic justification for the mixed-subsidy policy discussed in the next section. By balancing multiple tools, the government can potentially navigate around these individual cost explosions to achieve social targets more efficiently.

5 Exploring Mixed-Subsidy Policies: Challenges and Insights

In this section, we propose and analyze mixed-subsidy policies that combine the advantages of both provider-based and citizens-based incentives. From an optimization perspective, since a mixed policy generalizes the single incentives discussed earlier, it represents a relaxation of the restricted problems analyzed in Section 4.2. Consequently, while a mixed policy is mathematically guaranteed to perform at least weakly better than any single-subsidy alternative, our focus is on identifying scenarios where it delivers substantial cost savings. We develop an algorithm to find near-optimal policies and use numerical experiments to quantify these efficiency gains and draw insights into policy structure.

5.1 Intractability of Finding Optimal Mixed-Subsidy Policies

A mixed-subsidy policy provides a two-pronged approach by simultaneously addressing the financial barriers to affordability and the supply-side capacity shortages that lead to long waiting times. By encouraging citizens to visit the service provider in a timely manner and rewarding providers for maintaining a sufficiently high service rate, the government can resolve bottlenecks more effectively than with a single instrument. From hereon, we consider mixed-subsidy policies in two forms: $\theta = (\kappa, \eta, 0)$ and $\theta = (\kappa, 0, \zeta)$. We do not consider policies which apply both the fee-for-service and the rebate simultaneously, specifically $\eta, \zeta > 0$, because the downtime-rebate policy achieves the same target at a lower cost. However, we study either form of provider-based incentives in the mixed policy because the fee-for-service policy is easier to implement in practice. In Section 6.4, we propose a mixed policy that combines these two provider-side instruments to leverage their respective strengths: the implementation simplicity of the fee-for-service policy and the cost-efficiency of the downtime-rebate policy. This combined approach serves as a strategic tool to resolve the challenges of information asymmetry.

To implement the mixed-subsidy policy, we need to solve problem (3) for θ with either $\eta = 0$ or $\zeta = 0$. While we could discuss single-subsidy policies for a general demand rate function $\lambda_\theta = f(\kappa)$, this is not possible under the mixed-subsidy policy. From hereon, we assume that the demand rate depends on the voucher value via the logistic function: $\lambda_\theta = f(\kappa) = \frac{1}{1+e^{-\frac{\kappa-\tau}{\sigma}}}$, where τ, σ are location and scale parameters. We choose the logistic function because it can portray a common phenomenon where small discounts are ineffective and large discounts have decreasing marginal effectiveness. Unfortunately, the government's cost function is not jointly convex in its inputs. We provide counterexamples in Appendix

B.

5.2 Deriving Insights from Numerical Experiments

For the purpose of understanding the structure of an optimal policy θ^*, we construct an algorithm which finds a policy $\hat{\theta}$ such that $C(\hat{\theta}) \leq (1+\epsilon) \cdot C(\theta^*)$, where $\epsilon \in (0,1)$ is a parameter that we can change to control the accuracy of the algorithm at the expense of running time. We stress that the algorithm can be used regardless of the structure of $f(\kappa)$. As the algorithm is not the highlight of this section, we defer Algorithm 1 to Appendix A. By generating instances of our problem and computing the corresponding $\hat{\theta}$, we study the structure of the proposed policies to draw insights on applying mixed-subsidy policies.

Setup: Our numerical experiments are conceptually motivated by the UK's NHS dental care system. We select parameter values based on official NHS reports and clinical guidelines to reflect real-world dynamics. First, we set the population size to $N = 1000$. This number represents the typical patient list size for a full-time NHS dentist. According to NHS Dental Statistics for England 2023/24, approximately 24,200 dentists served 24.6 million patients, yielding an average of 1,017 patients per dentist a year (NHS Business Services Authority, 2024). Second, we set the deterioration rate to $\phi = 1$. This implies that a citizen needs dental care (e.g., check-up or treatment) once a year on average. This setting aligns with the National Institute for Health and Care Excellence (NICE) guidelines, which recommend a one-year recall interval for typical adult patients (Scott et al., 2022). Third, we model the demand function as $f(\kappa) = 0.375 + \frac{1}{1+e^{-\frac{\kappa-2}{0.8}}}$. This structure captures two key behaviors observed in the UK dental market: 1) Price-insensitive demand: We include a constant term so that $\lambda_0 > 0.375$. The baseline rate 0.375 ensures that without any subsidy, the steady-state fraction of satisfied citizens averages 25% across tested profit margins, which aligns with the fraction of fully self-funded citizens in the UK (The Guardian, 2026). 2) Price-sensitive demand: A 2024 survey by the Oral Health Foundation shows that 41% of UK adults delay dental visits due to cost (Foundation, 2024). We use a logistic function to model this sensitivity. It captures the reality that subsidies become effective only after they reach a certain threshold. Next, we model the crucial parameter, the provider's gross profit margin, $(r-c)/r$. The NHS England Dental Earnings and Expenses Estimates 2023/24 report indicates an average pre-tax profit margin of 49.5% for dentists in England (Service, 2025). However, this figure includes income from private services. The marginal profitability of NHS-specific services is much lower. Reports confirm that government subsidies often fail to cover the full cost of NHS treatments (Hill and Yhnell, 2024). Therefore, centering around the profit margin of 50%, we examine two scenarios. We define a "low-profitability zone"

(1-15%) to simulate the financial reality of NHS providers. This allows us to test which policies effectively prevent provider exit. We also test a "high-profitability zone" (over 50%) to generalize our insights to a broader range of essential service sectors, such as education or child care, that are characterized by recurring needs while having different profit margins from the healthcare sector. Finally, we vary the government's target B by considering a percentage increase of the fraction of satisfied citizens over the base case. Specifically, we set the government's target as $B = (1+\delta) \cdot \frac{\mathbb{E}[S|\,\mu_0,\lambda_0]}{N}$, where δ is target increase of satisfied citizens and is set between 2% to 20%.

Comparisons of Single-Subsidy Policies

We first compare the consumer-voucher policy against the two provider-based policies. Figure 3 compares the cost of the consumer-voucher policy against the fee-for-service policy as we vary the profit margin and δ, and we use the cell colors to indicate the less expensive policy. We repeat the same process for comparing the cost of the consumer-voucher policy against the downtime-rebate policy in Figure 4.

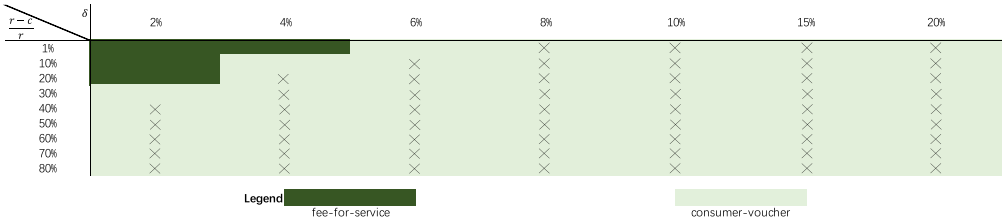


Figure 3: Comparison of the consumer-voucher policy and the fee-for-service policy. The darker cells indicate that the fee-for-service policy has lower cost than the consumer-voucher policy for the corresponding δ and $\frac{r-c}{r}$. An $\times$ indicates that the fee-for-service policy is infeasible.

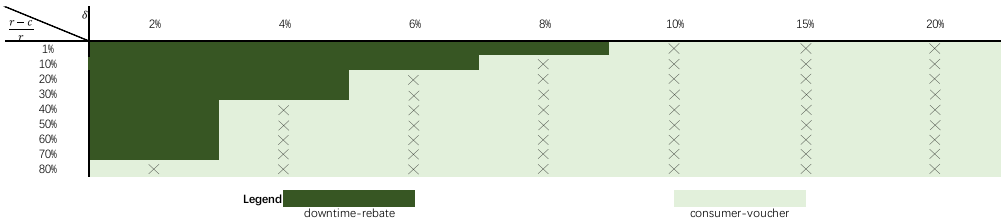


Figure 4: Comparison of the consumer-voucher policy and the downtime-rebate policy. The darker cells indicate that the downtime-rebate policy has lower cost than the consumer-voucher policy for the corresponding δ and $\frac{r-c}{r}$. An $\times$ indicates that the downtime-rebate policy is infeasible.

In line with Propositions 10 and 11, the provider-based policies incur lower costs than the consumer-voucher policy if the service provider's gross profit margin is low and the

government sets a low target. When the service provider's gross profit margin is low, handing out consumer vouchers will cause needy citizens to seek services in a timely manner, but the service provider only increases her service rate slightly and the new service rate is not enough to shorten the current queue. Since the consumer-voucher policy is less effective in decreasing the fraction of waiting citizens, the government would need to spend more money to achieve the desired target. In contrast, when the gross profit margin is high, then the bottleneck lies with needy citizens' delay in registering for their next visit, and the consumer-voucher policy becomes less expensive than the provider-based policies.

Since the downtime-rebate policy is always less expensive than the fee-for-service policy by Lemma 9, we know that the downtime-rebate policy will be less expensive than the consumer-voucher policy under more situations. This is observed in Figures 3 and 4.

When the government sets an aggressive target δ, then it becomes even more important to address the issue of affordability. Figures 3 and 4 suggest that when δ is large, the provider-based policies incur higher costs than the consumer-voucher policy. Furthermore, there are many instances where the provider-based policies are unable to achieve the government's targeted fraction of satisfied citizens.

We also report the subsidies received by each party in the three single-subsidy policies in Figures 5, 6, and 7. The value of the incentives under both provider-based incentives increase quickly as the gross profit margin increases, which implies that the cost of provider-based policies increases quickly when the system bottleneck does not lie with the service provider. In fact, if the government gives the service provider an excessively high fee for service or downtime rebate, then the service provider would be tempted to increase her service rate to an unnecessary level because the rewards are high and the risks are low. However, as the citizens are not receiving any financial assistance, they continue to stay away from the service provider and the government's spending is wasted on providing a high service rate without corresponding demand.

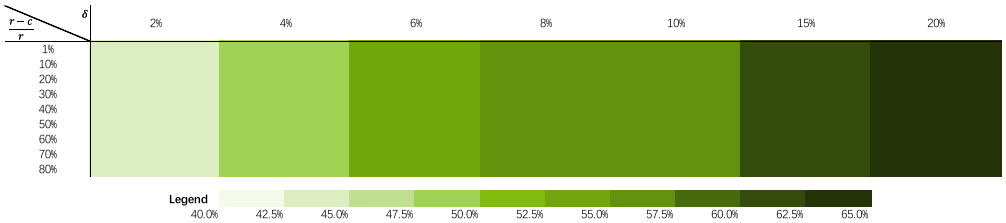


Figure 5: The voucher received by citizens under the consumer-voucher policy, as measured by κ/r.

On the other hand, we observe in Figure 5 that the value of the voucher remains nearly unchanged across different gross profit margins. This contrasts starkly with provider-based

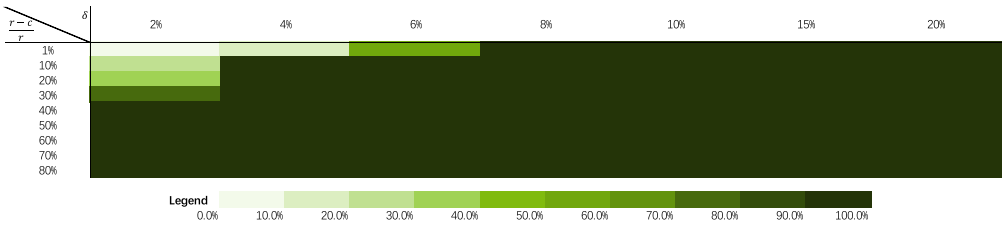


Figure 6: The subsidy received by the provider under the fee-for-service policy, as measured by η/r.

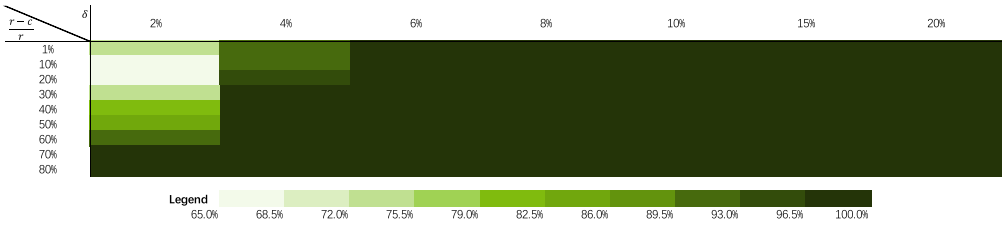


Figure 7: The rebate received by the provider under the downtime-rebate policy, as measured by ζ/c.

policies, where a higher gross profit margin quickly increases the policy's cost since it needs to provide higher subsidies. This is in line with the intuition that citizens do not care about the service provider's cost, and the consumer-voucher policy should have similar performances regardless of the gross profit margin.

Structure of Near-Optimal Mixed-Subsidy Policies

Next, we show that a mixed-subsidy policy is able to achieve cost savings in many scenarios. We turn to investigating the cost of the approximately-optimal policy, $\hat{\theta}$, which we find using Algorithm 1 with accuracy $\epsilon = 0.1$. Let θ_κ, θ_η, and θ_ζ denote the optimal single-incentive policies when each of κ, η, and ζ are non-zero respectively.

To measure the savings from applying $\hat{\theta}$, we report on the relative reduction in cost against the best of the two corresponding single-incentive policies. In particular, when we consider $\hat{\theta} = (\kappa, \eta, 0)$, we report on the savings $\Delta_\eta = \frac{\min\{C(\theta_\kappa), C(\theta_\eta)\} - C(\hat{\theta})}{\min\{C(\theta_\kappa), C(\theta_\eta)\}}$. Similarly, when we consider $\hat{\theta} = (\kappa, 0, \zeta)$, we report on $\Delta_\zeta = \frac{\min\{C(\theta_\kappa), C(\theta_\zeta)\} - C(\hat{\theta})}{\min\{C(\theta_\kappa), C(\theta_\zeta)\}}$. Since the mixed-subsidy policy includes single-incentive policies as special cases, Δ_η and Δ_ζ are non-negative. Hence, we say that the mixed-subsidy policy has lower cost only when $\Delta_\eta > 0$ or $\Delta_\zeta > 0$.

First, we consider the mixed-subsidy policy $\hat{\theta} = (\kappa, \eta, 0)$ in Figure 8, where we use a tri-colour graph to indicate the policy with the lowest cost. The mixed-subsidy policy excels when the gross profit margin is extremely low and the government only plans to increase the fraction of satisfied citizens by a small rate. As quantified in Figure 9, the

mixed-subsidy policy can reduce the government's cost by up to $\Delta_\eta = 50\%$ compared to single-subsidy policies. While the precise magnitude of these gains would naturally vary depending on localized operational complexities, the result underscores the fundamental economic advantage of integrating demand- and supply-side incentives. Also, we observe that the mixed-subsidy policy removes the entire region where the fee-for-service policy dominates the consumer-voucher policy, which indicates that the mixed policy does mitigate the inherent inefficiencies of single subsidies.

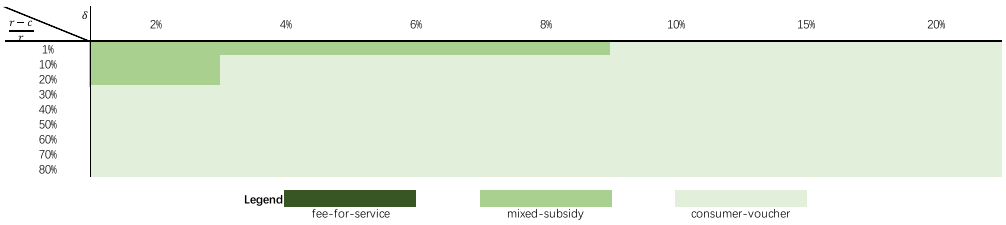


Figure 8: Comparison of the mixed-subsidy policy $\hat{\theta} = (\kappa, \eta, 0)$ against the optimal single-subsidy policies distributing either κ or η

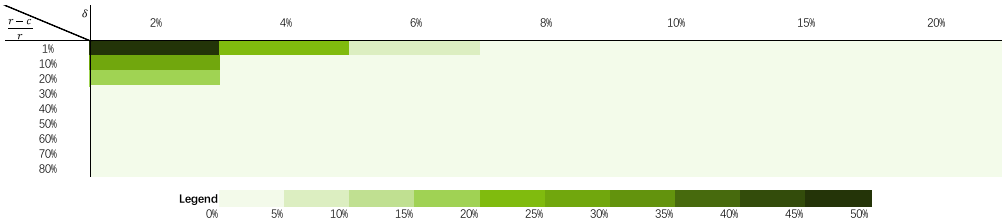


Figure 9: Cost-savings from the mixed-subsidy policy $\hat{\theta} = (\kappa, \eta, 0)$ against the optimal single-subsidy policy distributing either κ or η, as measured by $\Delta_\eta = \frac{\min\{C(\theta_\kappa), C(\theta_\eta)\} - C(\hat{\theta})}{\min\{C(\theta_\kappa), C(\theta_\eta)\}}$.

Next, we consider the mixed-subsidy policy $\hat{\theta} = (\kappa, 0, \zeta)$. From Figure 10 we see that the mixed-subsidy policy out-performs the single-subsidy policies in a wide range of scenarios, particularly along the off-diagonal where we see the optimal single-subsidy choice shifts. Notably, the timing of the downtime-rebate and consumer-voucher payouts complements each other. These advantages enable the mixed-subsidy policy to integrate both incentives effectively, resulting in cost savings across more scenarios. We quantify the cost savings in Figure 11. Compared to the single best alternative, the mixed-subsidy policy demonstrates substantial cost efficiencies with cost reductions reaching up to 25%. The darker off-diagonal, which indicates more savings from the mix-subsidy policy, matches the switchover between the two single-subsidy policies in Figure 4, confirming that the mixed policy effectively bridges the performance gap between single subsidies.

We also examine the fraction of the funds used on citizens versus the service provider in Figure 12. This is a key indicator for public budget reporting, as directing money toward

citizens rather than corporations often makes a policy easier to promote in practice. In many scenarios, we find that the mixed-subsidy policy indeed prioritizes citizens. The government spends more on rebates only when the target δ is small, or the provider's gross profit margin is low. In these cases, it is more effective to directly increase the service rate rather than relying on the indirect effects of consumer vouchers.

The strength of this mixed policy lies in its efficient re-allocation of funds. Even when the downtime-rebate policy is less costly than the consumer-voucher policy (e.g., at a 70% and 2% target), the mixed policy still re-allocates about 10% of the budget to vouchers, which leads to a 20% saving in total cost. Conversely, when the consumer-voucher policy is less expensive, allocating just a small fraction to the service provider is sufficient to encourage her to serve more citizens. These results demonstrate a strong complementarity of the two subsidies. By combining both incentives, the government can utilize the unique advantages of each to improve overall policy performance and reduce total spending.

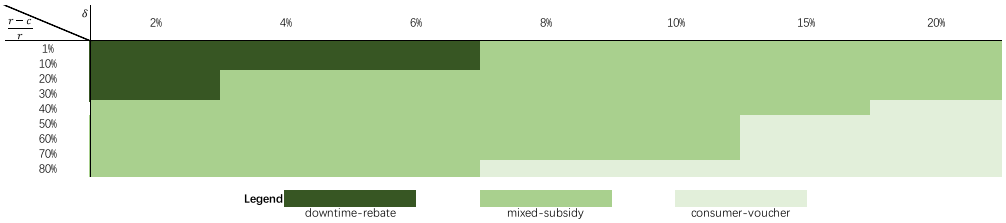


Figure 10: Comparison of the mixed policy $(\kappa, 0, \zeta)$ against the optimal single-incentive policy distributing either κ or ζ.

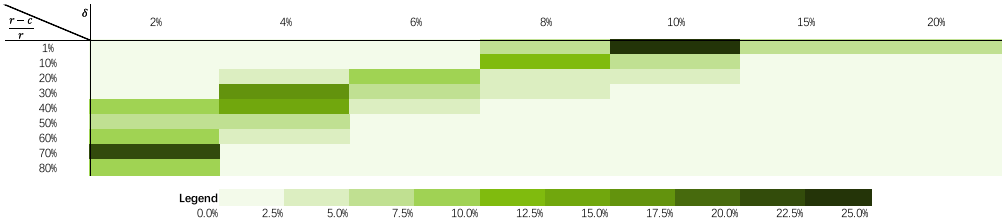


Figure 11: Cost saved by the mixed policy $(\kappa, 0, \zeta)$ against the optimal single-incentive policy distributing either κ or ζ, as measured by $\Delta_\zeta = \frac{\min\{C(\theta_\kappa), C(\theta_\zeta)\} - C(\hat{\theta})}{\min\{C(\theta_\kappa), C(\theta_\zeta)\}}$.

6 Extensions

In this section, we discuss several extensions to our main model that could account for more complex service settings.

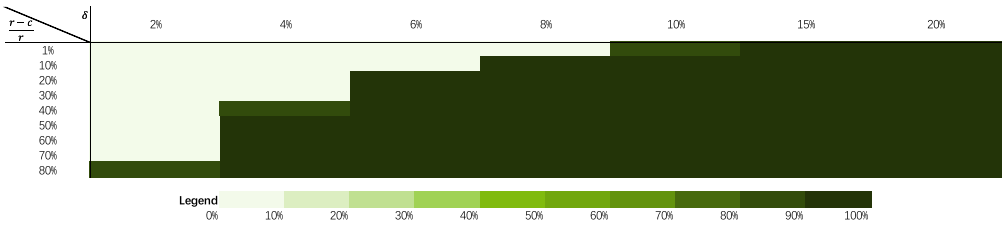


Figure 12: Total spending on consumer vouchers as a fraction of the cost of the mixed-subsidy policy $\hat{\theta} = (\kappa, 0, \zeta)$: $\frac{\kappa \cdot \mu_{\hat{\theta}} \cdot \left(1 - \tilde{Q}_0(\mu_{\hat{\theta}}, \lambda_{\hat{\theta}})\right)}{C(\hat{\theta})}$.

6.1 The Multi-Server Setting

Throughout this work, we assumed that the service provider has a single service station and can change her service rate. Studying systems with a single-server model is common because these systems are easier to analyze, whereas there are fewer theoretical results for the multi-server setting.

From our discussion in Section 3, the single-server model and the multi-server model are applicable under different situations. The single-server model is more appropriate when expensive machines are involved and there are less expensive ways to increase service rate, and the multi-server model is more appropriate when it is feasible to increase the workforce. In this section, we extend our model to the multi-server setting.

6.1.1 Structure of the Multi-Server Model

Suppose the government implements policy $\theta = (\kappa, \eta, \zeta)$. Under the multi-server setting, the citizens' demand rate does not change and depends only on the voucher value: $\lambda_\theta = f(\kappa)$. Each citizen's transition among the three states remains the same as in the single-server model. The main difference is the service provider's profit function and her ability to decide on the number of service stations to use instead of the long-run service rate. The service time for each citizen at a service station follows an exponential distribution with a fixed rate μ. The service provider can decide on the number of service stations that she wants to operate, M, in order to maximize her profit. We assume that she pays a fixed operating cost of c per unit time for each service station, resulting in a total cost rate of cM in the long run.

If the number of waiting citizens is $w \geq M$, then all of the service stations are serving citizens and producing revenue. However, if $w < M$, then only some of the service stations are producing revenue. Let $\tilde{Q}_w(M, \lambda)$ denote the probability that there are w waiting citizens when needy citizens book with rate λ and she operates M service stations, and let $\mathbb{E}[W|\ M, \lambda]$ denote the expected number of waiting citizens. We can apply similar notations for the expected number of satisfied and needy citizens. Under policy θ, the service provider's subsi-

dized revenue is $\sum_{w=0}^{N}(r+\eta)\cdot\mu\cdot\min\{w,M\}\cdot\tilde{Q}_w(M,\lambda_\theta) = (r+\eta)\cdot\mu\cdot\mathbb{E}[\min\{W,M\}|\ M,\lambda_\theta]$. On the other hand, she receives the downtime rebate whenever one or more of her service stations are not serving citizens, and her expected rebate is $\sum_{w=0}^{N}\zeta\mu\cdot\max\{M-w,0\}\cdot\tilde{Q}_w(M,\lambda_\theta) = \zeta\mu M - \zeta\mu\cdot\mathbb{E}[\min\{W,M\}|\ M,\lambda_\theta]$. Hence, in the multi-server setting, the service provider's problem is:

$$\max_{\substack{1\le M\le N\\ M\in\mathbb{Z}}} \{(r+\eta-\zeta)\cdot\mu\cdot\mathbb{E}[\min\{W,M\}|\ M,\lambda_\theta] - (c-\zeta\mu)\cdot M\}.$$

We require $r+\eta-\zeta\ge 0$ and $c-\zeta\mu\ge 0$ because the government does not offer rebates that are large enough to cover the cost of a permanently idle machine. Let M_θ denote the service provider's best response to Problem (6).

When citizens and the service provider operate under their best responses, the government's cost has a similar form to the cost function in Section 3: $C(\theta) = (\kappa+\eta-\zeta)\cdot\mu\cdot\mathbb{E}[\min\{W,M_\theta\}|\ M_\theta,\lambda_\theta] + \zeta\mu M_\theta$. The government's problem can be updated with the new cost function: $\min\left\{C(\theta)\ \middle|\ \mathbb{E}[S|\ M_\theta,\lambda_\theta]/N \ge B\right\}$.

6.1.2 Analysis of the Multi-Server Model

To what extent can we analyze the multi-server setting? By computing the stationary probabilities (see Appendix A), we can determine the steady-state probability of observing w waiting citizens and the expected number of citizens in each state:

Proposition 12. *For any fixed M and λ, the probability that there are w citizens waiting in steady state is*

$$\tilde{Q}_w(M,\lambda) = \begin{cases} \tilde{Q}_0(M,\lambda)\cdot\binom{N}{w}\cdot\nu^{-w} & \text{if } 0\le w\le M-1 \\ \tilde{Q}_0(M,\lambda)\cdot\binom{N}{w}\cdot\nu^{-w}\cdot\frac{w!}{M!\,M^{w-M}} & \text{if } M\le w\le N \end{cases},$$

where $\tilde{Q}_0(M,\lambda) = \left[\sum_{w=0}^{M-1}\binom{N}{w}\cdot\nu^{-w} + \sum_{w=M}^{N}\binom{N}{w}\cdot\nu^{-w}\cdot\frac{w!}{M!\,M^{w-M}}\right]^{-1}$.

The expected numbers of citizens in each state are: $\mathbb{E}[W|\ M,\lambda] = N-\nu\cdot\mathbb{E}[\min\{W,M\}|\ M,\lambda]$, $\mathbb{E}[S|\ M,\lambda] = \frac{\mu}{\phi}\cdot\mathbb{E}[\min\{W,M\}|\ M,\lambda]$ *and* $\mathbb{E}[D|\ M,\lambda] = \frac{\mu}{\lambda}\cdot\mathbb{E}[\min\{W,M\}|\ M,\lambda]$.

In order to obtain analytical results, we need to determine the service provider's best response. Although the number of service stations to operate is a discrete decision, we can linearize the profit function over M. Let $\Pi_\theta(M)$ be a function such that $\Pi_\theta(M) = (r+\eta-\zeta)\cdot\mu\cdot\mathbb{E}[\min\{W,M\}|\ M,\lambda_\theta] - (c-\zeta\mu)\cdot M$ when $M\in\mathbb{Z}$. If $M\notin\mathbb{Z}$ such that $M = \alpha\cdot\lfloor M\rfloor + (1-\alpha)\cdot\lceil M\rceil$ for $0<\alpha<1$, then set $\Pi_\theta(M) = \alpha\cdot\Pi_\theta(\lfloor M\rfloor) + (1-\alpha)\cdot\Pi_\theta(\lceil M\rceil)$.

Problem (6) is equivalent to $\max_{1\leq M\leq N}\Pi_\theta(M)$ because $\Pi_\theta(M)\leq\max\{\Pi_\theta(\lfloor M\rfloor),\Pi_\theta(\lceil M\rceil)\}$, and the main challenge is to prove that $\Pi(M)$ is concave.

Showing that $\Pi_\theta(M)$ is concave is equivalent to showing that $\mathbb{E}[\min\{W,M\}|\ M,\lambda]$ is increasing and concave in M, or that $\mathbb{E}[W|\ M,\lambda]$ is decreasing and convex by Proposition 12. As one may expect, adding service stations provides a similar effect to increasing the service rate of a single service station and reduces the expected number of waiting citizens.

Lemma 13. *Suppose λ is fixed. As the provider increases the number of operating service stations M, the expected number of satisfied and needy citizens, $\mathbb{E}[S|\ M,\lambda]$ and $\mathbb{E}[D|\ M,\lambda]$, increase. The expected number of waiting citizens, $E[W|\ M,\lambda]$, decreases.*

Unfortunately, it is very difficult to prove the concavity of $\Pi_\theta(M)$. In the single-server model, we were able to show that the service provider's best response is the optimal solution to a concave maximization problem by collapsing the model to a M/M/N/N queue, or the Erlang-B loss model. In the multi-server model, the analysis on the number of waiting citizens collapses into the machine repairmen model (Buckley and Jowers, 2006). Under the machine repairmen model, there are n machines serving customers, but they break down with rate $(\frac{1}{\lambda}+\frac{1}{\phi})^{-1}$. There are also m repairmen who fix the machines with rate μ. The event of a machine failing corresponds to the event that a satisfied citizen developing a toothache and registering to see the service provider, so that he becomes a waiting citizen. The event of a machine being repaired corresponds to a waiting citizen receiving service to become satisfied again. However, there are insufficient theoretical results behind the machine repairmen model that would allow us to obtain corresponding results to Lemma 2.

Due to the difficulty of giving a closed-form optimality condition for the service-provider's best response, we run numerical experiments to determine whether the insights in the single-server setting are robust and applicable to the multi-server setting.

6.1.3 Numerical Experiments

Setup: Let $N=1000$. The service provider charges a price of $r=10$ to each customer she serves. The service rate of each station is $\mu=800$. The deterioration rate is set as $\phi=600$. Similar with Section 5.2, we consider a logistic demand rate where $\lambda_\theta=f(\kappa)=375+1000\cdot\frac{1}{1+e^{-\frac{\kappa-2}{0.8}}}$.

Unlike the single server setting, we find that provider-based incentives are not necessary when $r\mu>c$. This setting implies that a service station generates positive income when there is a continuous stream of citizens waiting for service. In contrast, if $r\mu\leq c$, then a service station is not profitable even if it is permanently busy. Our numerical study shows that these two scenarios result in very different strategies for the government and is one of

the main drivers of policy decisions, and we present both settings in Figures 13 and 14. We vary $r\mu/c$ between 80% to 130% and set the cost of a service station accordingly.

We also change the way that we vary the government's target. Whereas we do not see sharp jumps in the service provider's best response in the single-server setting, fixed costs would lead to sharp jumps in the number of service stations in the current setting. In particular, if $r\mu \leq c$, then the service provider is unprofitable and would prefer to run the minimum number of service stations, which results in almost no satisfied citizens. We set the minimum to $M = 1$ for the purpose of ensuring that there is some minimal level of service in the system. On the other hand, if $r\mu > c$, then the number of service stations and the expected fraction of satisfied citizens increase substantially. As a result, it becomes unfair to consider a relative improvement on the expected number of satisfied citizens and we turn to imposing absolute targets on the fraction of satisfied citizens. We vary the government's target B between 20% to 50% of the population being satisfied in expectation.

Results: We list the results of the mixed-subsidy policy using fee-for-service subsidies and consumer vouchers. The insights of the mixed-subsidy policy using downtime rebates and consumer vouchers are similar in the M-server model. In Figure 13, we identify the less expensive single-subsidy policy under different settings. We include settings where the government's target has been achieved without offering any subsidies, and these settings are marked as "no policy". This option was not necessary under the single-server setting because we focused on relative performance to the base case without government incentives and any improvements would require positive subsidy. We also extend the tests to situations where all policies are infeasible because the government has chosen an aggressive target, and these settings are marked as "infeasible". Similar to the single-server setting, the consumer-voucher policy and the fee-for-service policy are each suited to different scenarios which depend on the service provider's profitability and the government's target. When $r\mu \leq c$, it is natural to implement the fee-for-service policy because the service provider cannot make a positive profit even if she is permanently busy with a single service station. As such, she has no incentive to add service station and help the government increase the fraction of satisfied citizens. However, if the government's target is too aggressive with a large B, then the target becomes infeasible because the bottleneck lies with service affordability to the citizens. On the other hand, if $r\mu > c$, then the consumer-voucher policy should be chosen if subsidies are necessary.

When we include the mixed-subsidy policy in Figure 14, we observe that the infeasible region has shrunk. When $r\mu < c$ and the government has an aggressive target B, combining consumer vouchers with fee-for-service subsidies would help generate the demand necessary to achieve the target.

These results are consistent with the single-server model. Single-subsidy policies suffice when the service provider earns a reasonable profit or when citizens are not hampered by excessively high prices. When both issues are observed, a mixed-subsidy policy is the most cost-effective method of achieving the government's target.

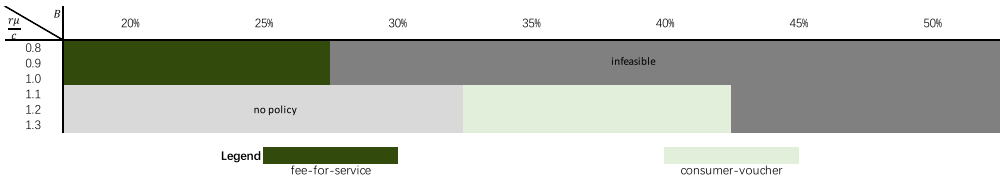


Figure 13: Comparison of the fee-for-service policy and the consumer-voucher policy in the M-server model.

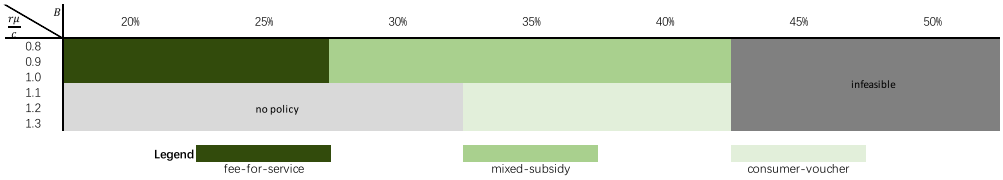


Figure 14: Comparison of the mixed-subsidy policy against the best single policies in the M-server model.

6.2 Endogenous Demand with Waiting Costs

In the main model, we assumed that the citizens' demand rate is primarily determined by the out-of-pocket price and hence influenced by the voucher value κ. However, in many service settings, long waiting times can act as a deterrent, causing citizens to delay seeking care or to balk after registering. This phenomenon implies that the demand rate should be endogenous to the congestion level of the system.

In this extension, we extend our model to incorporate waiting costs. We demonstrate that our main structural results, especially the cost-effectiveness of the downtime-rebate policy over the fee-for-service policy, and the benefits of the mixed policy, remain robust under this more complex setting.

6.2.1 Model Setup for Endogenous Demand with Waiting Costs

Let $\mathbb{E}[W]$ denote the expected number of waiting citizens in the system. We assume that citizens can observe the expected congestion level and adjust their registration rate accordingly. We model the effective demand rate λ_{eff} as a function of both the voucher value κ

and the expected queue length $\mathbb{E}[W]$:

$$\lambda_{eff}(\kappa, \mathbb{E}[W]) = f(\kappa) \cdot \exp\left(-\alpha \cdot \mathbb{E}[W]\right), \tag{6}$$

where $f(\kappa)$ is the price-dependent demand function, and $\alpha \geq 0$ represents citizens' sensitivity to delays. Notice that when $\alpha = 0$, the model degenerates to the baseline case where demand depends only on price. When $\alpha > 0$, a longer queue exerts a negative effect on demand, capturing the balking behavior.

Since the steady-state expected queue length $\mathbb{E}[W]$ itself depends on the demand rate λ_{eff} and the service rate μ, the equilibrium demand rate λ^* is the solution to a fixed-point problem:

$$\lambda^* = f(\kappa) \cdot \exp\left(-\alpha \cdot \mathbb{E}[W|\ \mu, \lambda^*]\right). \tag{7}$$

Due to the lack of closed-form solutions in this extension, analytically tracking the specific number of citizens in each state is difficult. However, our numerical results confirm that the core policy insights remain robust.

6.2.2 Numerical Experiments with Waiting Costs

We conduct numerical experiments with the delay sensitivity parameter $\alpha = 0.01$. We test a range of profit margins from 1% to 80%, and government targets from 5% to 30%. The results are summarized in Figures 15-18.

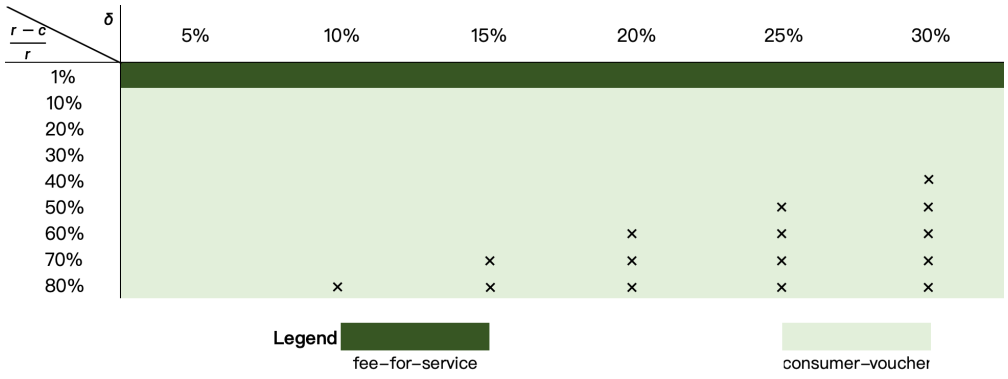


Figure 15: Comparison of the consumer-voucher policy and the fee-for-service policy with waiting costs. The darker cells indicate that the fee-for-service policy has lower costs than the consumer-voucher policy for the corresponding δ and $\frac{r-c}{r}$. An $\times$ indicates that the fee-for-service policy is infeasible.

Consistent with the main model, the downtime-rebate policy remains more cost-effective than the fee-for-service policy, particularly when the profit margin and the target are high.

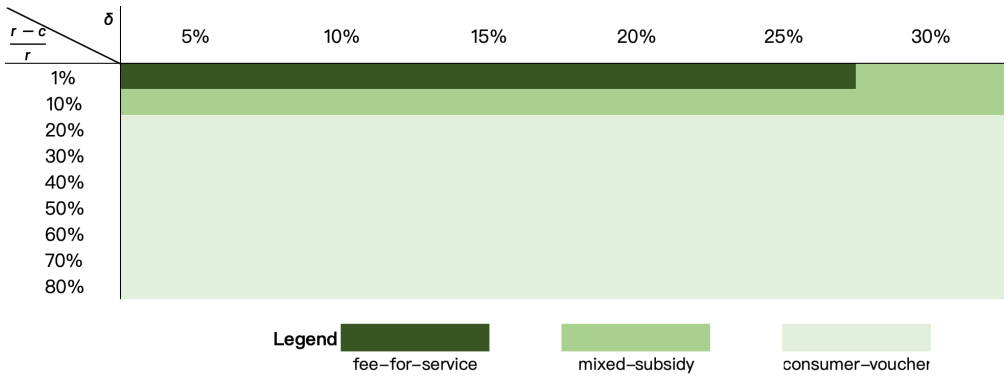


Figure 16: Comparison of the mixed policy $(\kappa, \eta, 0)$ against the optimal single-incentive policy distributing either κ or η, with waiting costs.

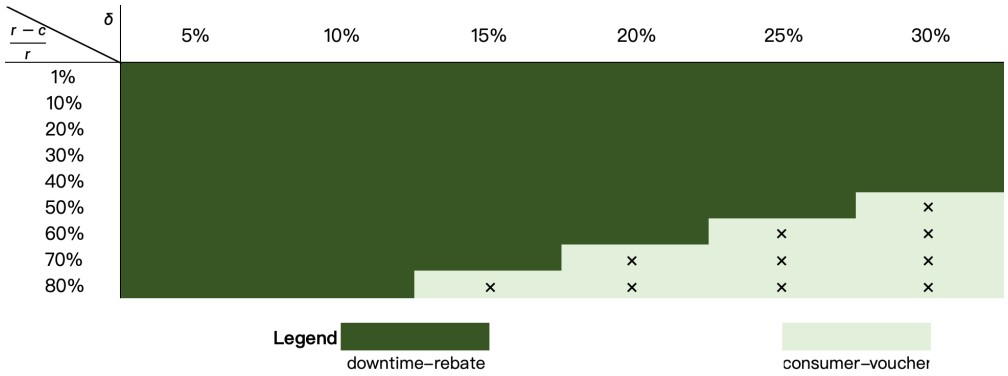


Figure 17: Comparison of the consumer-voucher policy and the downtime-rebate policy with waiting costs. The darker cells indicate that the downtime-rebate policy has lower costs than the consumer-voucher policy for the corresponding δ and $\frac{r-c}{r}$. An $\times$ indicates that the downtime-rebate policy is infeasible.

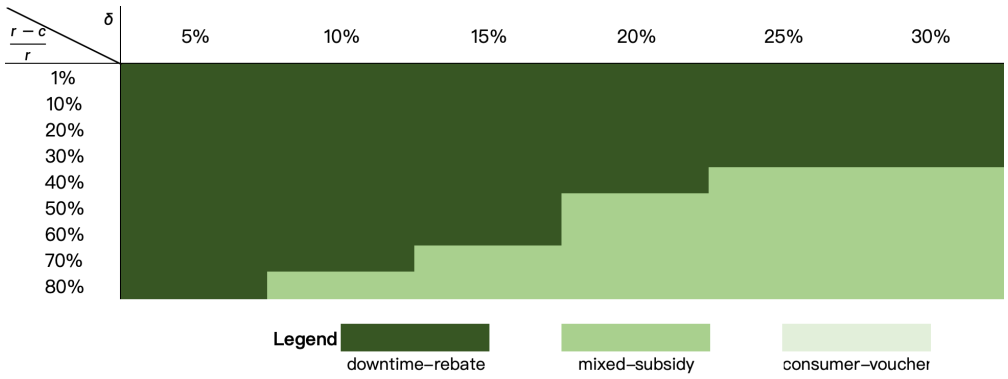


Figure 18: Comparison of the mixed policy $(\kappa, 0, \zeta)$ against the optimal single-incentive policy distributing either κ or ζ, with waiting costs.

For example, in the scenario with a profit margin of 60% and a target increase of 20%, the fee-for-service policy diminishes in effectiveness and becomes infeasible, while the downtime-rebate policy retains its dominance.

The mixed-subsidy policy continues to outperform single-subsidy policies, particularly when the provider operates at a moderate profit margin. The introduction of waiting costs does not alter the fundamental trade-offs between subsidizing the provider and citizens derived in the main text.

Notably, the inclusion of waiting costs strengthens the case for provider-based policies by expanding their effective regions. This is because provider-based policies incentivize the provider to increase the service rate μ, which reduces the expected queue length $\mathbb{E}[W]$ and has a secondary benefit that it encourages more citizens to register for services.

In contrast, policies that rely solely on stimulating demand via vouchers result in higher congestion. Under $\alpha > 0$, this congestion dampens the effective impact of the subsidy, making such policies less efficient. Thus, subsidizing the provider to reduce waiting times serves a dual purpose: it improves accessibility and sustains patient engagement.

6.3 The Impact of Moral Hazard and Elective Demand

In our baseline model, the rate at which citizens transition from satisfied to needy, i.e., the rate ϕ, is assumed to be exogenous. However, in many real-world service settings, consumer behavior is dynamically influenced by both financial incentives and operational performance factors, e.g., waiting time.

In this section, we relax the baseline assumptions to incorporate these endogenous behaviors. We analyze two distinct scenarios: (1) Price-driven moral hazard, where ϕ is exacerbated by low out-of-pocket prices; (2) Wait-time-affected demand, where both λ and ϕ are influenced by the expected waiting time.

6.3.1 Price-Driven Moral Hazard

Our baseline model implies that consumer vouchers facilitate access for citizens. However, in many social service, particularly healthcare contexts, reducing out-of-pocket costs can induce moral hazard, where citizens seek care for low-value or elective reasons simply because the price is low. For example, a citizen may visit a dentist solely for reassurance.

Model Setup. To capture the citizens' radical reaction to a price decrease, we model ϕ as an increasing function of the voucher value. This reflects the behavioral tendency of citizens to seek care more frequently when out-of-pocket costs are low, often for minor or elective issues that would otherwise be ignored.

Let the effective deterioration rate be $\phi_e(\kappa) = \phi_0 \cdot \left(1 + \alpha \cdot \frac{\kappa}{r}\right)$, where ϕ_0 is the baseline deterioration rate representing genuine medical needs. $\alpha \geq 0$ is the sensitivity parameter for price reduction. A higher α implies that citizens are more prone to seeking elective care when the price decreases. This formulation captures the induced demand. By this formulation, as the voucher κ increases, the flow of citizens entering the needy state accelerates. These citizens may then make additional visits that consume provider capacity and government funds but may contribute little to the core social target of treating genuine needs.

Numerical Experiments with Price-Driven Moral Hazard. We run the numerical experiments with a sensitivity of $\alpha = 1$ and other settings being fixed as in Section 5.2. Our results in Figures 19-22 demonstrate that the structural dominance of the policies remains similar to the baseline analysis. Specifically, the downtime-rebate policy retains its cost-effectiveness advantage over the fee-for-service policy. While the introduction of moral hazard makes the consumer-voucher policy less efficient since every dollar spent on vouchers now triggers an inflationary effect on demand volume through ϕ_e, the mixed-subsidy policy continues to outperform single-subsidy policies in a wide range of scenarios.

We do observe a marginal shrink in the dominance region of the mixed policy $(\kappa, 0, \zeta)$, especially when the provider's profit margin is high. Intuitively, when $\kappa > 0$, the induced increase in ϕ_e necessitates a higher service rate μ to maintain the same satisfaction target. This makes both the voucher and downtime-rebate component more costly relative to the baseline. Consequently, in scenarios where the provider is already profitable and the rebate's leverage on capacity is limited, the government may find it optimal to rely more heavily on the consumer-based subsidy, instead of splitting the budget over two costly subsidies.

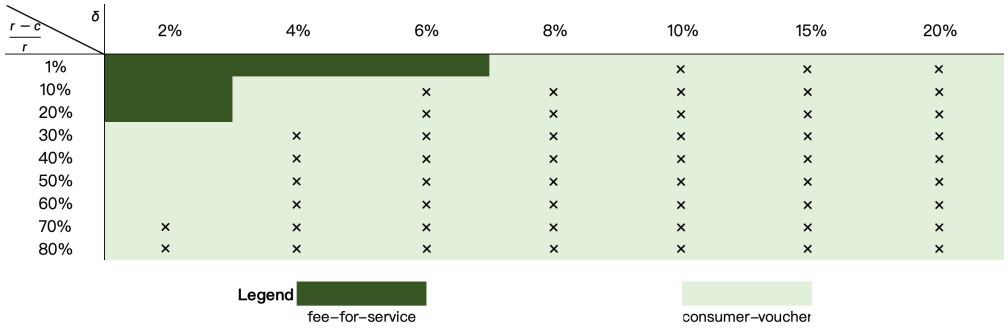


Figure 19: Comparison of the consumer-voucher policy and the fee-for-service policy with a price-dependent deterioration rate $\phi_e = \phi_0 \cdot (1 + \frac{\kappa}{r})$. The darker cells indicate that the fee-for-service policy has lower costs than the consumer-voucher policy for the corresponding δ and $\frac{r-c}{r}$. An $\times$ indicates that the fee-for-service policy is infeasible.

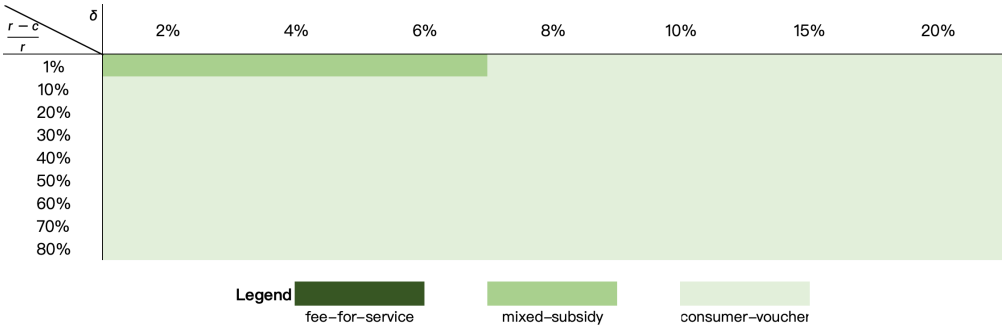


Figure 20: Comparison of the mixed policy $(\kappa, \eta, 0)$ against the optimal single-incentive policy distributing either κ or η, with a price-dependent deterioration rate $\phi_e = \phi_0 \cdot (1 + \frac{\kappa}{r})$.

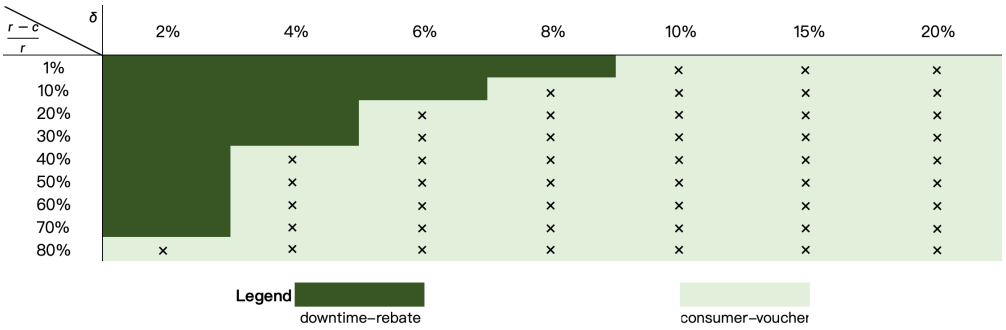


Figure 21: Comparison of the consumer-voucher policy and the downtime-rebate policy with a price-dependent deterioration rate $\phi_e = \phi_0 \cdot (1 + \frac{\kappa}{r})$. The darker cells indicate that the downtime-rebate policy has lower costs than the consumer-voucher policy for the corresponding δ and $\frac{r-c}{r}$. An $\times$ indicates that the downtime-rebate policy is infeasible.

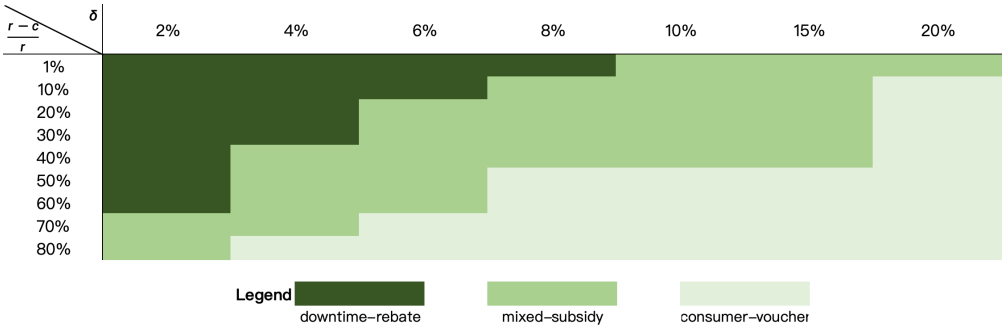


Figure 22: Comparison of the mixed policy $(\kappa, 0, \zeta)$ against the optimal single-incentive policy distributing either κ or ζ, with a price-dependent deterioration rate $\phi_e = \phi_0 \cdot (1 + \frac{\kappa}{r})$.

6.3.2 Elective Demand with Reduced Waiting Cost

We further extend the model to a setting where ϕ could be influenced by the system congestion.

Model Setup. We modify the model to allow both the demand rate and the deterioration rate to be decreasing with the expected queue length:

$$\lambda_e(\kappa, \mathbb{E}[W]) = f(\kappa) \cdot \exp(-\alpha_\lambda \cdot \mathbb{E}[W]), \quad \phi_e(\mathbb{E}[W]) = \phi_0 \cdot \exp(-\alpha_\phi \cdot \mathbb{E}[W]),$$

where $\alpha_\lambda \geq 0$ and $\alpha_\phi \geq 0$ represent the sensitivity to balking. Under this setting, satisfied citizens may transition to the needy state more frequently (e.g., for minor checks or reassurance) when they observe that the system is efficient and wait times are negligible. Conversely, high waiting costs deter citizens from seeking care for minor issues. Note that in this extension, ϕ_e is not directly affected by the voucher value κ, isolating the effect of wait-time-driven behavior.

Numerical Experiments with Waiting-Cost-Affected Demand. Similarly with Section 6.2, we solve this system numerically using a fixed-point iteration algorithm. With everything else being set as in Section 5.2, we set $\alpha_\lambda = \alpha_\phi = 0.01$ to simulate a scenario where citizens are sensitive to delays.

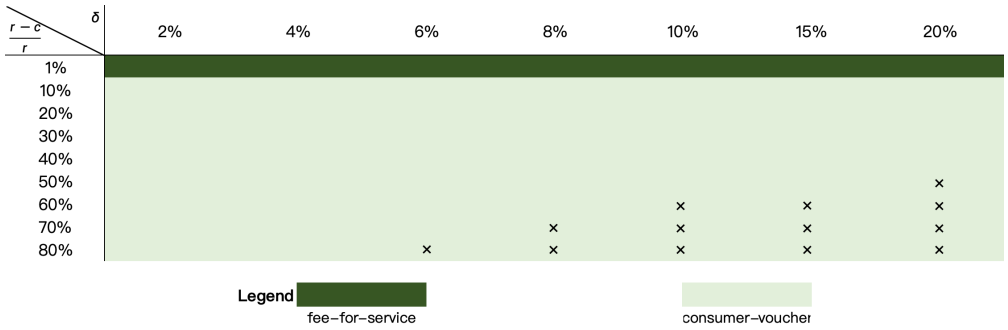


Figure 23: Comparison of the consumer-voucher policy and the fee-for-service policy with a waiting-time-affected deterioration rate $\phi_e = \phi_0 \cdot \exp(\alpha_\phi \cdot \mathbb{E}[W])$. The darker cells indicate that the fee-for-service policy has lower costs than the consumer-voucher policy for the corresponding δ and $\frac{r-c}{r}$. An $\times$ indicates that the fee-for-service policy is infeasible.

As shown in Figures 23–26, the qualitative behavior of the optimal policies is similar to the baseline model. The downtime-rebate policy remains superior to the fee-for-service policy, and the mixed-subsidy policy yields significant cost savings, particularly in the area where neither affordability nor capacity is the sole bottleneck.

These results confirm that our main findings are structurally robust to endogenous demand as well as deterioration rates. Whether the friction arises from price (moral hazard)

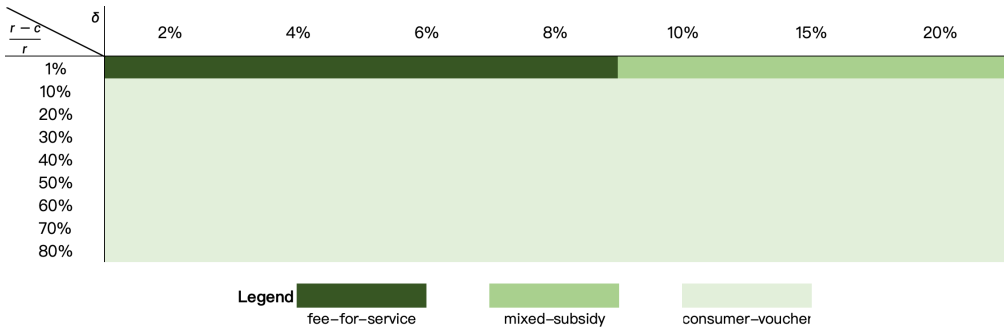


Figure 24: Comparison of the mixed policy $(\kappa, \eta, 0)$ against the optimal single-incentive policy distributing either κ or η, with a waiting-time-affected deterioration rate $\phi_e = \phi_0 \cdot \exp(\alpha_\phi \cdot \mathbb{E}[W])$.

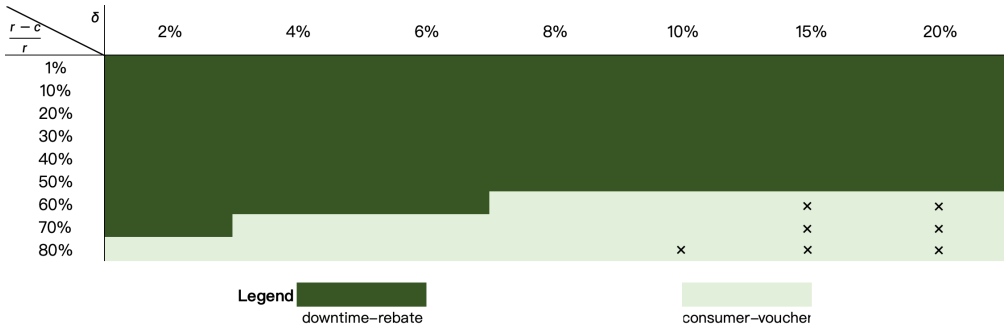


Figure 25: Comparison of the consumer-voucher policy and the downtime-rebate policy with a waiting-time-affected deterioration rate $\phi_e = \phi_0 \cdot \exp(\alpha_\phi \cdot \mathbb{E}[W])$. The darker cells indicate that the downtime-rebate policy has lower costs than the consumer-voucher policy for the corresponding δ and $\frac{r-c}{r}$. An $\times$ indicates that the downtime-rebate policy is infeasible.

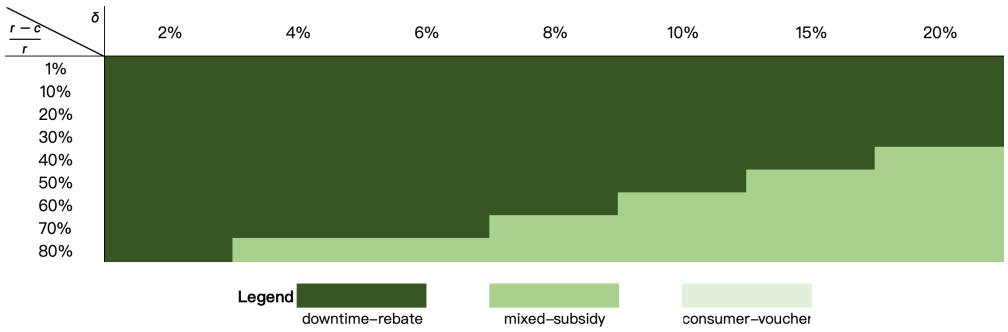


Figure 26: Comparison of the mixed policy $(\kappa, 0, \zeta)$ against the optimal single-incentive policy distributing either κ or ζ, with a waiting-time-affected deterioration rate $\phi_e = \phi_0 \cdot \exp(\alpha_\phi \cdot \mathbb{E}[W])$.

or waiting time (congestion), the principle of mitigating provider risk via downtime rebates while subsidizing consumers remains the most cost-effective approach for the government when the system bottleneck is unclear.

6.4 Policy Implementation: A Mechanism Design Approach

In Section 4, we established that the downtime-rebate policy is theoretically more cost-effective than the fee-for-service policy. However, practical implementation is complicated by information asymmetry regarding the provider's unobservable service rate μ and her true idle time. In this extension we adopt a mechanism design framework to solve this problem. We first solve for the theoretically optimal allocation of service rates and government payments. We then demonstrate that this optimum can be implemented using a specific menu of contracts based on our policy tools, the fee-for-service subsidy η and the downtime-rebate ζ, provided they are designed with a specific volume cap structure.

6.4.1 Model Setup for Policy Implementation

Consider the provider's operating cost c to be private information, taking a value of c_L (efficient) or c_H (inefficient) with probabilities β and $1-\beta$, respectively, where $c_L < c_H$. We start by considering a direct revelation mechanism where the government invites the provider to report her cost type and, based on this report, assigns a *verifiable expected throughput* $R(\mu)$ that is related to her service rate μ. For example, $R(\mu)$ may be the number of served citizens over an audit year at a provider. Assume $R(\mu)$ to be strictly concave and increasing with $R'(\mu) > 0, R''(\mu) < 0$. Let $R(0) = 0$. Theoretically, the government's problem is to design a menu $\{(\mu_L, P_L), (\mu_H, P_H)\}$ to minimize expected costs subject to incentive compatibility (IC) and individual rationality (IR) constraints. Here $P_i, i \in \{L, H\}$ is the assigned payment to type i.

The provider's problem. Let $U_i = P_i - c_i\mu_i$ be the utility of a truth-telling provider. The rent of a type-i provider mimicking to be type-$\hat{i}$, $i \neq \hat{i}$, and $i, \hat{i}\{L, H\}$, is as follows: $\Pi(\hat{i}|\ i) = P_{\hat{i}} - c_i\mu_{\hat{i}} = U_{\hat{i}} + (c_{\hat{i}} - c_i)\mu_{\hat{i}}$. Then given a menu of contracts, a type-i provider's problem is expressed as $\max\{U_i, \Pi(\hat{i}|\ i)\}$.

The government's problem. Before proceeding, we first define the IC constraint for the provider. The IC constraint is to prevent the provider from mimicking. Hence, for a type-i provider, the contract must ensure that $U_i \geq \Pi(\hat{i}|\ i) = U_{\hat{i}} + (c_{\hat{i}} - c_i)\mu_{\hat{i}}$. Since U_i has a unique mapping with (μ_i, P_i), we formulate the government's problem by (μ_i, U_i) instead

of (μ_i, P_i):

$$\min_{\{\mu_i, U_i\}_{i\in\{L,H\}}} \quad \beta(c_L\mu_L + U_L) + (1-\beta)(c_H\mu_H + U_H) \tag{8}$$

$$\text{s.t.} \quad U_L \geq U_H + (c_H - c_L)\mu_H, \tag{IC_L}$$

$$U_H \geq U_L + (c_L - c_H)\mu_L, \tag{IC_H}$$

$$U_L \geq 0, \tag{IR_L}$$

$$U_H \geq 0, , \tag{IR_H}$$

$$\beta R(\mu_L) + (1-\beta)R(\mu_H) \geq \mathcal{T}. \tag{9}$$

Solving for the optimal allocation yields the following Lemma.

Lemma 14. *At the optimal solution $\{(\mu_i^*, U_i^*), i \in \{L, H\}\}$ to Problem (8) satisfies the following properties:*

1. *The efficient provider's IC constraint IC_L and the inefficient provider's IR constraint IR_H are binding, yielding information rents $U_H^* = 0$ and $U_L^* = (c_H - c_L)\mu_H^*$.*
2. *The target constraint is binding.*
3. *The constraints IC_H and IR_L are redundant.*
4. *The optimal service rates satisfy the monotonicity condition $\mu_L^* > \mu_H^*$.*

The optimal payment P_i^* can be derived by $P_i^* = U_i^* + c_i\mu_i^*$.

6.4.2 Implementation via Policy Instruments

Having derived the theoretical optimum, we now address the implementation: How can the government achieve this outcome using the fee-for-service subsidy and the downtime rebate?

A naive linear contract of the form creates a deviation problem. If the efficient provider chooses the inefficient contract, her lower marginal cost would incentivize her to produce substantially more than μ_H^* , deviating from the government's target and disrupting the predicted costs. To solve this, we propose that the optimal mechanism can be implemented as a menu of *capped contracts*.

Contract design. The government offers a menu of contracts, where each contract i is defined by a tuple $(\eta_i, \zeta_i, \bar{R}_i), i \in \{L, H\}$. Here η_i and ζ_i represent the fee-for-service subsidy and the downtime rebate, respectively. While $\bar{R}_i$ represents the *maximum subsidized volume*. The payment rule $T_i(R)$ for a provider producing volume R is defined as $T_i(R) = (\eta_i - \zeta_i)\min\{R, \bar{R}_i\} + \zeta_i K$. Here we make the assumption that the benchmark capacity K

satisfies $K > R(\mu_L^*) + \frac{(\eta_H - \zeta_H)[R(\mu_L^*) - R(\mu_H^*)] - c_L(\mu_L^* - \mu_H^*)}{(\eta_L - \zeta_L) - (\eta_H - \zeta_H)}$. Since in the downtime-rebate context, K typically refers to the total physical capacity (e.g., maximum possible admitted patients) while $R(\mu_L^*)$ is the targeted utilization, the assumption is a mild regularity condition. It simply requires that the benchmark capacity is not set virtually identical to the efficient target volume.

Proposition 15. *The optimal allocation* $\{(\mu_i^*, P_i^*), i \in \{L, H\}\}$ *is implemented by a menu of two capped contracts* $(\eta_i, \zeta_i, \bar{R}_i), i \in \{L, H\}$ *that are designed as follows:*

1. *The government sets the volume caps equal to the optimal target volumes, i.e.,* $\bar{R}_L = R(\mu_L^*)$ *and* $\bar{R}_H = R(\mu_H^*)$.

2. *The rates satisfy* $\eta_L > \eta_H$ *and* $\zeta_L < \zeta_H$.

Proposition 15 demonstrates that the government can screen providers by offering a specific trade-off between the two policy tools. On the one hand, it offers a high-risk-high-reward contract with a high fee-for-service rate η_L but a low downtime rebate ζ_L. This incentivizes the efficient provider to leverage her lower marginal costs to produce high volume up to $\bar{R}_L$. On the other hand, it offers a low-risk contract with a low fee-for-service rate η_H but a generous downtime rebate ζ_H. This ensures participation of the inefficient provider.

The volume cap $\bar{R}_H$ is a critical instrument that prevents the efficient provider from mimicking the inefficient one. Without the cap, an efficient provider would select contract-H to enjoy the high downtime rebate ζ_H while simultaneously overproducing to capture fee-for-service revenue. By capping the subsidized volume at $\bar{R}_H = R(\mu_H^*)$, the government effectively eliminates the incentive for the efficient provider to get double benefits.

7 Conclusions

In this work, we investigate the use of government subsidies when there are both supply and demand constraints in a partially subsidized service sector. Our model incorporates the citizens' rate of pursuing service in response to the price paid and the service provider's service rate in response to the profitability of providing service. We show that single-subsidy policies, which direct subsidies exclusively to the citizens or the service provider, can only solve one side of supply and demand issues. Subsidies paid to the citizens will make access more affordable, whereas subsidies paid to the service provider will incentivize higher capacity and reduce waiting time for citizens who are already seeking service. Both types of subsidies have diminishing returns as the government targets a larger fraction of satisfied citizens.

The key takeaway for policymakers is the significant value of an integrated approach to subsidy design. Our findings can be synthesized into a practical policy selection framework. The optimal policy choice should be guided by an initial diagnosis of the system's provider profitability and its primary bottleneck. When the provider is profitable but the service is unaffordable, a consumer-voucher policy is most effective. Conversely, when the service is affordable but the provider is unprofitable and capacity is low, a provider-side subsidy, specifically a downtime rebate, is the superior choice. Moreover, we propose and analyze a mixed-subsidy policy that strategically allocates funds to both citizens and providers. Our results demonstrate that such a policy can reduce government costs significantly compared to the best single-subsidy policy. This substantial efficiency gain provides a compelling justification for adopting a more integrated policy structure. By integrating consumer vouchers with provider-based subsidies, the government can address issues of affordability and timely access simultaneously.

One challenge in this model was to incorporate multiple service stations. We present some preliminary work on the multi-server setting and numerical results. We leave the theoretical results for this analytically challenging model for future work. Another promising direction is to investigate the impact of citizen heterogeneity to design more equitable, targeted subsidy schemes. Furthermore, while our baseline assumes a fixed target, a government could endogenously adjust its objectives to balance increased service accessibility against redundant demand induced by moral hazard. Our supplemental analysis confirms that our core policy insights remain robust under such adaptive targeting.

*

REFERENCES

Alizamir, S., Iravani, F., and Mamani, H. (2019). An analysis of price vs. revenue protection: Government subsidies in the agriculture industry. *Management Science*, 65(1):32–49.

Andritsos, D. A. and Aflaki, S. (2015). Competition and the operational performance of hospitals: The role of hospital objectives. *Production and Operations Management*, 24(11):1812–1832.

Arora, P., Rahmani, M., and Ramachandran, K. (2022). Doing less to do more? Optimal service portfolio of non-profits that serve distressed individuals. *Manufacturing & Service Operations Management*, 24(2):883–901.

Arora, P., Wei, W., and Solak, S. (2021). Improving outcomes in child care subsidy voucher programs under regional asymmetries. *Production and Operations Management*, 30(12):4435–4454.

Bendoly, E., Donohue, K., and Schultz, K. L. (2006). Behavior in operations management: Assessing recent findings and revisiting old assumptions. *Journal of Operations Management*, 24(6):737–752.

Buckley, J. J. and Jowers, L. J. (2006). *The Machine/Service Queuing Model*, pages 125–131. Springer Berlin Heidelberg, Berlin, Heidelberg.

Çakıcı, Ö. E. and Mills, A. F. (2025). Telehealth in acute care: Pay parity and patient access. *Manufacturing & Service Operations Management*, 27(1):40–58.

Chemama, J., Cohen, M. C., Lobel, R., and Perakis, G. (2019). Consumer subsidies with a strategic supplier: Commitment vs. flexibility. *Management Science*, 65(2):681–713.

Cohen, M. C., Lobel, R., and Perakis, G. (2016). The impact of demand uncertainty on consumer subsidies for green technology adoption. *Management Science*, 62(5):1235–1258.

De Véricourt, F. and Jennings, O. B. (2008). Dimensioning large-scale membership services. *Operations Research*, 56(1):173–187.

Dent-Line of Canada Inc. (2024). Dental Care in Rural Areas: North American Challenges and Solutions. `https://www.dent-line.com/dental-care-in-rural-areas-north-american-challenges-and-solutions.html#Factors_leading_to_limited_availability_of_dental_care_in_rural_areas`. Accessed: December 20, 2024.

Department of Health and Social Care (2024). New payments for dentists to make more appointments available. `https://www.gov.uk/government/news/new-payments-for-dentists-to-make-more-appointments-available`. Accessed: March 1, 2024.

Department of Health and Social Care and Whately, H. (2023). £600 million social care winter workforce and capacity boost. `https://www.gov.uk/government/news/600-million-social-care-winter-workforce-and-capacity-boost`. Accessed: May 28, 2024.

Department of Health, HKSAR (2024). Elderly Healthcare Voucher Scheme. `https://www.hcv.gov.hk/en/hcvs/background.html`. Accessed: May 28, 2024.

Désir, A., Goyal, V., and Zhang, J. (2022). Capacitated assortment optimization: Hardness and approximation. *Operations Research*, 70(2):893–904.

Foundation, O. H. (2024). UK could face dental health crisis as costs soar, says Oral Health Foundation. `https://www.dentalhealth.org/news/uk-could-face-dental-health-crisis-as-costs-soar-says-oral-health-foundation`.

Government Grants Management Function, UK (2024). Bus Service Operators Grant. `https://www.find-government-grants.service.gov.uk/grants/bus-service-operators-grant-commercial-transport-operators-1#summary`. Accessed: May 28, 2024.

Government of Ontario (2024). Family support and respite for children and youth with special needs. `https://www.ontario.ca/page/family-support-and-respite-children-and-youth-special-needs`. Accessed: May 28, 2024.

Government of Western Australia (2024). Special Education Funding. `https://www.education.wa.edu.au/special-education-per-capita-funding`. Accessed: May 28, 2024.

Green, R., Fagg, J., and Hughes, D. (2023). Children waiting over a year in pain for NHS tooth removal. `https://www.bbc.com/news/health-66095984`. Accessed: July 24, 2024.

Guo, P., Tang, C. S., Wang, Y., and Zhao, M. (2019). The impact of reimbursement policy on social welfare, revisit rate, and waiting time in a public healthcare system: Fee-for-service versus bundled payment. *Manufacturing & Service Operations Management*,

21(1):154–170.

Hill, M. and Yhnell, R. (2024). NHS dental budget: More than £25m going unspent. https://www.bbc.com/news/uk-england-somerset-68413931.

Hua, Z., Chen, W., and Zhang, Z. G. (2016). Competition and coordination in two-tier public service systems under government fiscal policy. *Production and Operations Management*, 25(8):1430–1448.

Kai-Ineman, D., Tversky, A., et al. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):363–391.

Kemp, C. (2024). Mother of sons with autism says respite, other specialized services lacking in rural Manitoba. https://www.cbc.ca/news/canada/manitoba/respite-westman-autism-support-1.7110677. Accessed: August 12, 2024.

Kotsi, T. O., Aflaki, A., Aydin, G., and Pedraza-Martinez, A. J. (2023). Allocation of nonprofit funds among program, fundraising, and administration. *Manufacturing & Service Operations Management*, 25(5):1873–1889.

Krishnan, K. (1990). The convexity of loss rate in an Erlang loss system and sojourn in an Erlang delay system with respect to arrival and service rates. *IEEE Transactions on Communications*, 38(9):1314–1316.

LDC Confederation (2024). What is a UDA? https://ldc.org.uk/what-is-a-uda/. Accessed: January 23, 2024.

McCoy, J. H. and Lee, H. L. (2014). Using fairness models to improve equity in health delivery fleet management. *Production and Operations Management*, 23(6):965–977.

Mehrotra, M. and Natarajan, K. V. (2020). Value of combining patient and provider incentives in humanitarian health care service programs. *Production and Operations Management*, 29(3):571–594.

Ministry of Health, Singapore (2024). Community Health Assist Scheme (CHAS). https://www.moh.gov.sg/managing-expenses/schemes-and-subsidies/chas. Accessed: December 20, 2024.

Morris, J. (2023). One in ten Britons have performed dentistry on themselves, half in the last two years. https://yougov.co.uk/politics/articles/45450-one-ten-britons-have-performed-dentistry-themselve?redirect_from=%2Ftopics%2Fpolitics%2Farticles-reports%2F2023%2F03%2F22%2Fone-ten-britons-have-performed-dentistry-themselve. Accessed: June 14, 2024.

National Health Service (2024). How much will I pay for NHS dental treatment? https://www.nhs.uk/nhs-services/dentists/how-much-will-i-pay-for-nhs-dental-treatment/. Accessed: June 14, 2024.

NHS Business Services Authority (2024). Dental statistics - England 2023/24. https://www.bbc.com/news/uk-northern-ireland-64631019. Accessed: August 22, 2024.

Olsder, W., Martagan, T., and Tang, C. S. (2023). Improving access to rare disease treatments: Subsidy, pricing, and payment schemes. *Management Science*, 69(9):5256–5274.

Qian, Q., Guo, P., and Lindsey, R. (2017). Comparison of subsidy schemes for reducing waiting times in healthcare systems. *Production and Operations Management*, 26(11):2033–2049.

Qian, Q. and Zhuang, W. (2017). Tax/subsidy and capacity decisions in a two-tier health system with welfare redistributive objective. *European Journal of Operational Research*, 260(1):140–151.

Raz, G. and Ovchinnikov, A. (2015). Coordinating pricing and supply of public interest goods using government rebates and subsidies. *IEEE Transactions on Engineering Management*, 62(1):65–79.

Scott, H., Cope, A. L., Wood, F., Joseph-Williams, N., Karki, A., Roberts, E. M., Lovell-Smith, C., and Chestnutt, I. G. (2022). A qualitative exploration of decisions about dental recall intervals-Part 1: attitudes of NHS general dental practitioners to NICE guideline CG19 on the interval between oral health reviews. *British Dental Journal*, 232(5):327–331.

Service, N. H. (2025). Dental Earnings and Expenses, 2023/24.

Siddiq, A., Tang, C. S., and Zhang, J. (2022). Partnerships in urban mobility: Incentive mechanisms for improving public transit adoption. *Manufacturing & Service Operations Management*, 24(2):956–971.

Tang, C. S., Wang, Y., and Zhao, M. (2024). The impact of input and output farm subsidies on farmer welfare, income disparity, and consumer surplus. *Management Science*, 70(5):3144–3161.

Taylor, T. A. and Xiao, W. (2014). Subsidizing the distribution channel: Donor funding to improve the availability of malaria drugs. *Management Science*, 60(10):2461–2477.

Telford, L. (2023). Dentists in NI lose money on most NHS treatments, BDA says. `https://www.bbc.com/news/uk-northern-ireland-64631019`. Accessed: August 22, 2024.

The Guardian (2026). Almost a third of people in England use private dentists amid NHS dental crisis. `https://www.theguardian.com/society/2026/mar/09/nhs-dental-crisis-private-dentist-patient-incease-england`.

The Standard (2020). HK government guarantees 50pc occupancy for designated quarantine hotels. `https://www.thestandard.com.hk/breaking-news/section/4/160237/HK-government-guarantees-50pc-occupancy-for-designated-quarantine-hotels`. Accessed: May 29, 2024.

U.S. Department of Health and Human Services (2024). Biden-Harris Administration Announces New Rule to Reduce Costs for More than 100,000 Families Receiving Child Care Subsidies. `https://www.hhs.gov/about/news/2024/02/29/biden-harris-administration-announces-new-rule-reduce-costs-more-than-100000-families-receiving-child-care-subsidies.html`. Accessed: May 28, 2024.

Wei, W., Arora, P., and Solak, S. (2024). Allocation of funds in bilevel subsidy welfare programs. *Manufacturing & Service Operations Management*, 26(4):1435–1453.

Xiao, P., Xiao, R., Liang, Y., Chen, X., and Lu, W. (2020). The effects of a government's subsidy program: Accessibility beyond affordability. *Management Science*, 66(7):3211–3233.

Yu, J. J., Tang, C. S., and Shen, Z.-J. M. (2018). Improving consumer welfare and manufacturer profit via government subsidy programs: subsidizing consumers or manufacturers? *Manufacturing & Service Operations Management*, 20(4):752–766.

Appendix

A Proofs of Results

Proof. **Proof of Proposition 1:** For cleanliness, we fix λ and μ and denote $Q_{s,w} = Q_{s,w}(\mu, \lambda)$ in this proof. We drop the notation on λ, μ from $n_{s,w}$ and Γ too. First, we can write out the balance equations based on Figure 1 as follows:

$$
\begin{aligned}
&(N-s-w+1)\cdot\lambda\cdot Q_{s,w-1} + \mu\cdot Q_{s-1,w+1} + (s+1)\cdot\phi\cdot Q_{s+1,w} \\
&\qquad = \big(\mu + s\phi + (N-s-w)\cdot\lambda\big)\cdot Q_{s,w} && \text{if } s,w>0; s+w<N \\
&\lambda\cdot Q_{s,w-1} + \mu\cdot Q_{s-1,w+1} = (\mu+s\phi)\cdot Q_{s,w} && \text{if } s,w>0; s+w=N \\
&(N-w+1)\cdot\lambda\cdot Q_{0,w-1} + \phi\cdot Q_{1,w} = \big(\mu+(N-w)\cdot\lambda\big)\cdot Q_{0,w} && \text{if } s=0; w\neq 0,N \\
&\lambda\cdot Q_{0,N-1} = \mu\cdot Q_{0,N} && \text{if } s=0; w=N \\
&\mu\cdot Q_{s-1,1} + (s+1)\cdot\phi\cdot Q_{s+1,0} = \big(s\phi+(N-s)\cdot\lambda\big)\cdot Q_{s,0} && \text{if } w=0; s\neq 0,N \\
&\mu\cdot Q_{N-1,1} = N\phi\cdot Q_{N,0} && \text{if } w=0; s=N \\
&\phi\cdot Q_{1,0} = N\lambda\cdot Q_{0,0} && \text{if } w=0; s=0 \\
&\sum_{s=0}^{N}\sum_{w=0}^{N-s} Q_{s,w} = 1
\end{aligned}
$$

We then prove the correctness of the stationary probabilities. Since Γ is a scaling factor, it suffices to show that $n_{s,w}$ satisfies the balance equations.
Case 1: If $s, w > 0$ and $s + w < N$, then:

$$
\begin{aligned}
&(N-s-w+1)\cdot\lambda\cdot n_{s,w-1} + \mu\cdot n_{s-1,w+1} + (s+1)\cdot\phi\cdot n_{s+1,w} \\
=&\frac{N!}{s!(N-s-w)!}\cdot\frac{\lambda^{s+w}}{\phi^s\mu^{w-1}} + \frac{N!}{(s-1)!(N-s-w)!}\cdot\frac{\lambda^{s+w}}{\phi^{s-1}\mu^w} + \frac{N!}{s!(N-s-1-w)!}\cdot\frac{\lambda^{s+w+1}}{\phi^s\mu^w} \\
=&\big(\mu+s\phi+(N-s-w)\cdot\lambda\big)\cdot\frac{N!}{s!(N-s-w)!}\cdot\frac{\lambda^{s+w}}{\phi^s\mu^w} \\
=&\big(\mu+s\phi+(N-s-w)\cdot\lambda\big)\cdot n_{s,w}
\end{aligned}
$$

Case 2: If $s, w > 0$ and $s + w = N$, then:

$$
\begin{aligned}
\lambda\cdot n_{s,w-1} + \mu\cdot n_{s-1,w+1} =& \frac{N!}{s!}\cdot\frac{\lambda^N}{\phi^s\mu^{w-1}} + \frac{N!}{(s-1)!}\cdot\frac{\lambda^N}{\phi^{s-1}\mu^w} \\
=&\big(\mu+s\phi\big)\cdot\frac{N!}{s!}\cdot\frac{\lambda^N}{\phi^s\mu^w} \\
=&\big(\mu+s\phi\big)\cdot n_{s,w}
\end{aligned}
$$

Case 3: If $s = 0$ and $w \neq 0, N$, then:

$$\begin{aligned}(N-w+1)\cdot\lambda\cdot n_{0,w-1} + \phi\cdot n_{1,w} =& \frac{N!}{(N-w)!}\cdot\frac{\lambda^w}{\mu^{w-1}} + \frac{N!}{(N-1-w)!}\cdot\frac{\lambda^{w+1}}{\mu^w}\\ =&\big(\mu+(N-w)\cdot\lambda\big)\cdot\frac{N!}{(N-w)!}\cdot\frac{\lambda^w}{\mu^w}\\ =&\big(\mu+(N-w)\cdot\lambda\big)\cdot n_{0,w}\end{aligned}$$

Case 4: If $s = 0$ and $w = N$, then:

$$\lambda\cdot n_{0,N-1} = N!\cdot\frac{\lambda^N}{\mu^{N-1}} = N!\cdot\mu\cdot\frac{\lambda^N}{\mu^N} = \mu n_{0,N}$$

Case 5: If $w = 0$ and $s \neq 0, N$, then:

$$\begin{aligned}\mu\cdot n_{s-1,1} + (s+1)\cdot\phi\cdot n_{s+1,0} =& \frac{N!}{(s-1)!(N-s)!}\cdot\frac{\lambda^s}{\phi^{s-1}} + \frac{N!}{s!(N-s-1)!}\cdot\frac{\lambda^{s+1}}{\phi^s}\\ =&\big(s\phi+(N-s)\cdot\lambda\big)\cdot\frac{N!}{s!(N-s)!}\cdot\frac{\lambda^s}{\phi^s}\\ =&\big(s\phi+(N-s)\cdot\lambda\big)\cdot n_{s,0}\end{aligned}$$

Case 6: If $w = 0$ and $s = N$, then:

$$\mu\cdot n_{N-1,1} = \frac{N!}{(N-1)!}\cdot\frac{\lambda^N}{\phi^{N-1}} = N\phi\cdot\frac{\lambda^N}{\phi^N} = N\phi\cdot n_{N,0}$$

Case 7: If $w = 0$ and $s = 0$, then:

$$\phi\cdot n_{1,0} = N\lambda = N\lambda\cdot n_{0,0}$$

Observe that $n_{s,w} = \left(\frac{\lambda}{\phi}\right)^s\cdot\binom{N-w}{s}\cdot n_{0,w}$. Hence,

$$\sum_{s=0}^{N-w} n_{s,w} = n_{0,w}\cdot\sum_{s=0}^{N-w}\binom{N-w}{s}\cdot\left(\frac{\lambda}{\phi}\right)^s = \frac{N!}{(N-w)!}\cdot\left(\frac{\lambda}{\mu}\right)^w\cdot\left(\frac{\lambda}{\phi}+1\right)^{N-w} = \frac{N!}{(N-w)!}\cdot\left(\frac{\lambda}{\mu}\right)^N\cdot\nu^{N-w}$$

where the second equality applies the binomial theorem. Hence, for any w, we have:

$$\tilde{Q}_w = \frac{\sum_{s=0}^{N-w} n_{s,w}}{\sum_{w'=0}^{N}\sum_{s=0}^{N-w'} n_{s,w'}} = \frac{\nu^{N-w}/(N-w)!}{\sum_{w'=0}^{N}\nu^{N-w'}/(N-w')!}$$

and $\tilde{Q}_0$ follows by multiplying the numerator and denominator by $N!/\nu^N$.

To compute the expected number of waiting citizens, observe that $\tilde{Q}_w = \tilde{Q}_0\cdot\frac{N!}{(N-w)!}\cdot\nu^{-w}$.

Hence, we can write the following:

$$\begin{aligned}
\mathbb{E}[W] = N - \mathbb{E}[N - W] &= N - \sum_{w=0}^{N-1}(N-w)\cdot\tilde{Q}_0\cdot\frac{N!}{(N-w)!}\cdot\nu^{-w} \\
&= N - \sum_{w=0}^{N-1}\tilde{Q}_0\cdot\frac{N!}{(N-w-1)!}\cdot\frac{\nu^{-(w+1)}}{\nu^{-1}} \\
&= N - \nu\cdot\sum_{w=1}^{N}\tilde{Q}_0\cdot\frac{N!}{(N-w)!}\cdot\nu^{-w} \\
&= N - \nu\cdot(1-\tilde{Q}_0).
\end{aligned}$$

The expected number of satisfied citizens, $\mathbb{E}[S]$, can be computed in a similar manner by observing that $Q_{s,w}/Q_{0,w} = n_{s,w}/n_{0,w}$:

$$\begin{aligned}
\mathbb{E}[S] = \sum_{w=0}^{N}\sum_{s=0}^{N-w} s\cdot Q_{s,w} &= \sum_{w=0}^{N}\sum_{s=0}^{N-w} s\cdot\binom{N-w}{s}\cdot\left(\frac{\lambda}{\phi}\right)^s\cdot Q_{0,w} \\
&= \sum_{w=0}^{N}\sum_{s=1}^{N-w}\frac{(N-w)!}{(s-1)!\,(N-w-(s-1))!}\cdot(N-w-(s-1))\cdot\left(\frac{\lambda}{\phi}\right)^s\cdot Q_{0,w} \\
&= \frac{\lambda}{\phi}\cdot\sum_{w=0}^{N}\sum_{s=0}^{N-w}\binom{N-w}{s}\cdot(N-w-s)\cdot\left(\frac{\lambda}{\phi}\right)^s\cdot Q_{0,w} \\
&= \frac{\lambda}{\phi}\cdot\sum_{w=0}^{N}\sum_{s=0}^{N-w}(N-w-s)\cdot Q_{s,w} \\
&= \frac{\lambda}{\phi}\cdot(N-\mathbb{E}[W]-\mathbb{E}[S])
\end{aligned}$$

By rearranging and substituting in the value of $\mathbb{E}[W]$ from above, we obtain $\mathbb{E}[S] = \frac{\mu}{\phi}\cdot\left(1-\tilde{Q}_0(\nu)\right)$. Finally, $\mathbb{E}[D] = N - \mathbb{E}[W] - \mathbb{E}[S] = \frac{\mu}{\lambda}\cdot\left(1-\tilde{Q}_0(\nu)\right)$.

□

Proof. **Proof of Lemma 2:** Krishnan (1990) showed that $\nu\cdot\tilde{Q}_0(\nu)$ is a convex function in ν, which implies that $\Pi_\theta(\nu)$ is a concave function. Taking the first derivative of $\Pi_\theta(\nu)$ with respect to ν gives us

$$\frac{d}{d\nu}\Pi_\theta(\nu) = (r+\eta-c) - (r+\eta-\zeta)\cdot\left(\tilde{Q}_0(\nu) + \nu\cdot\frac{d}{d\nu}\tilde{Q}_0(\nu)\right).$$

We focus on the derivative of $\tilde{Q}_0(\nu)$ with respect to ν:

$$\begin{aligned}\frac{d}{d\nu}\tilde{Q}_0(\nu) &= (-1)\cdot\left(\sum_{w=0}^{N}\frac{N!}{(N-w)!}\cdot\nu^{-w}\right)^{-2}\cdot\left(\sum_{w=1}^{N}\frac{N!}{(N-w)!}\cdot(-w\nu^{-w-1})\right)\\ &= \tilde{Q}_0(\nu)\cdot\nu^{-1}\cdot\mathbb{E}[W|\ \nu].\end{aligned}$$

By substituting the value of $\frac{d}{d\nu}\tilde{Q}_0(\nu)$ into $\frac{d}{d\nu}\Pi_\theta(\nu)$, we get

$$\frac{d}{d\nu}\Pi_\theta(\nu) = (r+\eta-c) - (r+\eta-\zeta)\cdot\tilde{Q}_0(\nu)\cdot(1+\mathbb{E}[W|\ \nu]).$$

By rearranging the first order condition $(r+\eta-c)-(r+\eta-\zeta)\cdot\tilde{Q}_0(\nu_\theta)\cdot(1+\mathbb{E}[W|\ \nu_\theta])=0$, we obtain the desired optimality condition.

The concavity of $\Pi_\theta(\nu)$ implies that the first derivative is monotone decreasing in ν. Hence $\tilde{Q}_0(\nu)\cdot(1+\mathbb{E}[W|\ \nu])$, which is on the right side of the optimality condition, is monotone increasing in ν. As $(r+\eta-c)/(r+\eta-\zeta)$ increases on the left side of the optimality condition, the optimal solution ν_θ also increases. □

Proof. **Proof of Lemma 3**: Since the left side of the optimality condition in Lemma 2 is fixed at $(r-c)/r$, the optimal ν_θ remains the same for all values of κ. Hence, $\mathbb{E}[W|\ \mu_\theta,\lambda_\theta]$ does not change for any value of κ.

Next, λ_θ increases as a function of κ, and $\mu_\theta = \nu_\theta\cdot\left(\frac{1}{\lambda_\theta}+\frac{1}{\phi}\right)^{-1}$. Hence, μ_θ increases as κ increases. Since ϕ is fixed, $\mathbb{E}[S|\ \mu_\theta,\lambda_\theta] = \frac{\mu_\theta}{\phi}\cdot(1-\tilde{Q}_0(\nu_\theta))$ increases as κ increases. This implies that the expected number of citizens in the remaining group, $\mathbb{E}[D|\ \mu_\theta,\lambda_\theta]$, must decrease as κ increases. □

Proof. **Proof of Corollary 4**: The proof of Lemma 3 tells us that ν_θ remains unchanged for all values of κ, and that μ_θ increases as κ increases. Hence, the cost of implementing the consumer-voucher policy can be rewritten as $C(\theta)=\kappa\mu_\theta\cdot(1-\tilde{Q}_0(\nu_\theta)) = \kappa\phi\cdot\mathbb{E}[S|\ \mu_\theta,\lambda_\theta]$, which increases as κ increases. □

Proof. **Proof of Proposition 5:** First, we compare the value of $\lambda_{\theta'}$ against the value of λ_θ by using the change in the expected number of satisfied citizens:

$$\frac{\mu_{\theta'}}{\phi}\cdot\left(1-\tilde{Q}_0(\nu_{\theta'})\right) = \mathbb{E}[S|\ \mu_{\theta'},\lambda_{\theta'}] = \psi\cdot\mathbb{E}[S|\ \mu_\theta,\lambda_\theta] = \psi\cdot\frac{\mu_\theta}{\phi}\cdot\left(1-\tilde{Q}_0(\nu_\theta)\right).$$

As the consumer-voucher policy only affects the service provider indirectly, we have $\nu_0=\nu_\theta=\nu_{\theta'}$. The above equality implies that $\mu_{\theta'}=\psi\mu_\theta$. Furthermore, $\nu_\theta=\nu_{\theta'}$ also implies that:

$$\mu_{\theta'}\cdot\left(\frac{1}{\phi}+\frac{1}{\lambda_{\theta'}}\right) = \mu_\theta\cdot\left(\frac{1}{\phi}+\frac{1}{\lambda_\theta}\right)\ \Rightarrow\ \lambda_{\theta'} = \frac{1}{\frac{1}{\psi}\cdot\left(\frac{1}{\phi}+\frac{1}{\lambda_\theta}\right)-\frac{1}{\phi}} = \frac{\psi\phi}{\phi-(\psi-1)\cdot\lambda_\theta}\cdot\lambda_\theta = \gamma\lambda_\theta.$$

Notice that if $\psi > 1 + \frac{\phi}{\lambda_\theta}$, then the denominator in the second last term is negative and $\lambda_{\theta'}$ does not exist. Hence, ψ cannot be too large.

Since $f(\kappa)$ is an increasing function, its inverse is also an increasing function, which implies that $\kappa' = f^{-1}(\lambda_{\theta'}) = f^{-1}(\gamma\lambda_\theta)$. Using the cost function, we have:

$$\frac{C(\theta')}{C(\theta)} = \frac{\kappa'\mu_{\theta'} \cdot \left(1 - \tilde{Q}_0(\nu_{\theta'})\right)}{\kappa\mu_\theta \cdot \left(1 - \tilde{Q}_0(\nu_\theta)\right)} = \frac{f^{-1}(\gamma\lambda_\theta) \cdot \phi \cdot \mathbb{E}[S|\ \mu_{\theta'}, \lambda_{\theta'}]}{f^{-1}(\lambda_\theta) \cdot \phi \cdot \mathbb{E}[S|\ \mu_\theta, \lambda_\theta]} = \frac{f^{-1}(\gamma\lambda_\theta)}{f^{-1}(\lambda_\theta)} \cdot \psi,$$

where the last inequality is due to $\mathbb{E}[S|\ \mu_{\theta'}, \lambda_{\theta'}] = \psi \cdot \mathbb{E}[S|\ \mu_{\theta'}, \lambda_{\theta'}]$. □

Proof. **Proof of Lemma 6**: As η or ζ increase, the left side of the optimality condition in Lemma 2 increases, which causes ν_θ to increase. Hence, it suffices to focus on the expected number of citizens in each state as ν_θ increases for a fixed demand rate.

We claim that $h(\nu) := \nu \cdot \left(1 - \tilde{Q}_0(\nu)\right)$ is a monotone increasing function. The first derivative of $h(\nu)$ follows from the proof of Lemma 2:

$$\frac{d}{d\nu}h(\nu) = 1 - \tilde{Q}_0(\nu) \cdot (1 + \mathbb{E}[W|\ \nu])\,.$$

We want to show that $\frac{d}{d\nu}h(\nu) \geq 0$ for all ν.

From the proof of Lemma 2, $\tilde{Q}_0(\nu) \cdot (1 + \mathbb{E}[W|\ \nu])$ is monotone increasing in ν. Hence $\frac{d}{d\nu}h(\nu)$ is a monotone decreasing function, and it suffices to prove that $\lim_{\nu\to\infty} \frac{d}{d\nu}h(\nu) \geq 0$. In particular, we show that $\lim_{\nu\to\infty} \tilde{Q}_0(\nu) = 1$ and $\lim_{\nu\to\infty} \mathbb{E}[W|\ \nu] = 0$. For the first limit, we use the expanded form of $\tilde{Q}_0(\nu)$ as presented in the proof of Proposition 1:

$$\lim_{\nu\to\infty} \tilde{Q}_0(\nu) = \lim_{\nu\to\infty} \frac{\nu^N/N!}{\sum_{w=0}^{N} \nu^{N-w}/(N-w)!} = \lim_{\nu\to\infty} \frac{\nu^N}{\nu^N + \sum_{w=0}^{N-1} \frac{N!}{w!} \cdot \nu^w} = 1.$$

The last equality recognizes that the leading term in the numerator and denominator is ν^N for a fixed N, and hence the limit is 1. For the second limit, we have:

$$\begin{aligned}
\lim_{\nu\to\infty} \mathbb{E}[W|\ \nu] &= \lim_{\nu\to\infty} N - \nu \cdot (1 - \tilde{Q}_0(\nu)) \\
&= N - \lim_{\nu\to\infty} \nu \cdot \frac{\sum_{w=1}^{N} \nu^{N-w}/(N-w)!}{\sum_{w=0}^{N} \nu^{N-w}/(N-w)!} \\
&= N - \lim_{\nu\to\infty} \frac{\sum_{w=1}^{N} \nu^{w}/(w-1)!}{\sum_{w=0}^{N} \nu^{w}/w!} \\
&= N - \lim_{\nu\to\infty} \frac{\frac{1}{(N-1)!} \cdot \sum_{w=1}^{N} \frac{(N-1)!}{(w-1)!} \cdot \nu^w}{\frac{1}{N!} \cdot \sum_{w=0}^{N} \frac{N!}{w!} \cdot \nu^w} \\
&= N - N \lim_{\nu\to\infty} \frac{\nu^N + \sum_{w=1}^{N-1} \frac{(N-1)!}{(w-1)!} \cdot \nu^w}{\nu^N + \sum_{w=0}^{N-1} \frac{N!}{w!} \cdot \nu^w} = N - N = 0
\end{aligned}$$

The last line again recognizes that the leading term in the numerator and denominator is ν^N for a fixed N. Hence, $\lim_{\nu\to\infty} \frac{d}{d\nu} h(\nu) = 1 - 1 = 0$ as required.

Observe that $E[S|\ \mu_\theta, \lambda_\theta] = \frac{\lambda_\theta}{\lambda_\theta+\phi} \cdot h(\nu_\theta)$, which is increasing in ν_θ for a fixed λ_θ. The same argument applies to $E[D|\ \mu_\theta, \lambda_\theta] = \frac{\phi}{\lambda_\theta+\phi} \cdot h(\nu_\theta)$. Finally, $\mathbb{E}[W|\ \mu_\theta, \lambda_\theta] = N - \mathbb{E}[S|\ \mu_\theta, \lambda_\theta] - \mathbb{E}[D|\ \mu_\theta, \lambda_\theta]$, which must then be monotone decreasing in ν_θ. □

Proof. **Proof of Corollary 7**: Lemma 2 states that the provider's best response ν_θ increases when η or ζ increases, which implies that μ_θ also increases. Under the fee-for-service policy, the government's cost can be rewritten as $C(\theta) = \eta\mu_\theta \cdot (1 - \tilde{Q}_0(\nu_\theta)) = \eta\phi \cdot \mathbb{E}[S|\ \mu_\theta, \lambda_\theta]$, which increases as η increases by Lemma 6.

Under the downtime-rebate policy, the proof of Lemma 2 showed that $\frac{d}{d\nu}\tilde{Q}_0(\nu) \geq 0$, which implies that $\tilde{Q}_0(\nu_\theta)$ increases as ζ increases. Since μ_θ also increases with ζ, the government's cost of $C(\theta) = \zeta\mu_\theta \cdot \tilde{Q}_0(\nu_\theta)$ increases as the rebate ζ increases. □

Proof. **Proof of Proposition 8:** We divide the proof into three parts.

Part 1, bounding the service rate: The proof is the same for either policy because $\lambda_\theta = \lambda_{\theta'}$ and we are only concerned with the change in $\mu_{\theta'}$. Observe that $\alpha > \psi > 1$ and $\alpha\mu_\theta$ is increasing in $\tilde{Q}_0(\nu_\theta)$ if $1 < \psi < (\tilde{Q}_0(\nu_\theta))^{-1}$. Outside of this range, α becomes negative, which indicates that it is not possible to achieve $\psi \cdot \mathbb{E}[S|\ \mu_\theta, \lambda_\theta]$.

Suppose by contradiction that the government can achieve $\mathbb{E}[S|\ \mu_{\theta'}, \lambda_{\theta'}] = \psi \cdot \mathbb{E}[S|\ \mu_\theta, \lambda_\theta]$ by offering enough subsidy to incentivize the service provider to operate at $\mu_{\theta'} \leq \alpha\mu_\theta$. The new capacity corresponds to $\nu_{\theta'} \leq \alpha\nu_\theta$. Lemma 6 implies that the expected number of satisfied citizens increases with the service rate if the demand rate does not change, which leads to $\mathbb{E}[S|\ \mu_{\theta'}, \lambda_{\theta'}] \leq \mathbb{E}[S|\ \alpha\mu_\theta, \lambda_\theta]$. We construct an upper-bound on the latter term to derive our contradiction. First, we construct an lower-bound on the resulting probability of no waiting citizens:

$$\begin{aligned}
\tilde{Q}_0(\alpha\nu_\theta) = \left(\sum_{j=0}^{N} \frac{N!}{(N-j)!} \cdot (\alpha\nu_\theta)^{-j}\right)^{-1} &> \left(1 + \sum_{j=1}^{N} \frac{N!}{(N-j)!} \cdot \nu_\theta^{-j}\alpha^{-1}\right)^{-1} \\
&= \left(1 + \left(\frac{1}{\tilde{Q}_0(\nu_\theta)} - 1\right) \cdot \alpha^{-1}\right)^{-1} \\
&= \left(1 + \frac{1 - \tilde{Q}_0(\nu_\theta)}{\tilde{Q}_0(\nu)} \cdot \frac{1 - \psi \cdot \tilde{Q}_0(\nu_\theta)}{\psi \cdot \left(1 - \tilde{Q}_0(\nu_\theta)\right)}\right)^{-1} = \psi \cdot \tilde{Q}_0(\nu_\theta)
\end{aligned}$$

The second equality recognizes that $\sum_{j=1}^{N} \frac{N!}{(N-j)!} \cdot \nu_0^{-j} = \frac{1}{\tilde{Q}_0(\nu_0)} - 1$. Next, we can upper-bound the expected number of satisfied citizens:

$$\begin{aligned}
\mathbb{E}[S|\ \mu_{\theta'}, \lambda_{\theta'}] \leq \mathbb{E}[S|\ \alpha\mu_\theta, \lambda_\theta] = \frac{\alpha\mu_\theta}{\phi} \cdot \left(1 - \tilde{Q}_0(\alpha\nu_\theta)\right) &< \frac{\mu_\theta}{\phi} \cdot \frac{\psi \cdot \left(1 - \tilde{Q}_0(\nu_\theta)\right)}{1 - \psi \cdot \tilde{Q}_0(\nu_\theta)} \cdot \left(1 - \psi \cdot \tilde{Q}_0(\nu_\theta)\right) \\
&= \psi \cdot \mathbb{E}[S|\ \mu_\theta, \lambda_\theta],
\end{aligned}$$

which contradicts the assumption that we could achieve the government's target. Hence, $\mu_{\theta'} > \alpha\mu_\theta$.

Part 2, government's cost of fee-for-service policy: Observe that $g_\eta(\nu)$ is an increasing function because ν increases with η, which implies that $g_\eta(\nu_{\theta'}) > g_\eta(\alpha\nu_\theta)$. Using the cost function with $\kappa = \zeta = 0$, we have:

$$\frac{C(\theta')}{C(\theta)} = \frac{\eta'\mu_{\theta'} \cdot \left(1 - \tilde{Q}_0(\nu_{\theta'})\right)}{\eta\mu_\theta \cdot \left(1 - \tilde{Q}_0(\nu_\theta)\right)} > \frac{g_\eta(\alpha\nu_\theta) \cdot \phi \cdot \mathbb{E}[S|\ \mu_{\theta'}, \lambda_{\theta'}]}{g_\eta(\nu_\theta) \cdot \phi \cdot \mathbb{E}[S|\ \mu_\theta, \lambda_\theta]} = \frac{g_\eta(\alpha\nu_\theta)}{g_\eta(\nu_\theta)} \cdot \psi.$$

Part 3, government's cost of downtime-rebate policy: Similar to part 2, $g_\zeta(\nu)$ is an increasing function because ν increases with ζ, which implies that $g_\zeta(\nu_{\theta'}) > g_\zeta(\alpha\nu_\theta)$. Using the cost function with $\kappa = \eta = 0$, we have:

$$\frac{C(\theta')}{C(\theta)} = \frac{\zeta'\mu_{\theta'} \cdot \tilde{Q}_0(\nu_{\theta'})}{\zeta\mu_\theta \cdot \tilde{Q}_0(\nu_\theta)} > \frac{g_\zeta(\alpha\nu_\theta) \cdot \alpha\mu_\theta \cdot \psi \cdot \tilde{Q}_0(\nu_\theta)}{g_\zeta(\nu_\theta) \cdot \mu_\theta \cdot \tilde{Q}_0(\nu_\theta)} = \frac{g_\zeta(\alpha\nu_\theta)}{g_\zeta(\nu_\theta)} \cdot \alpha\psi.$$

The inequality uses $\mu_{\theta'} > \alpha\mu_\theta$ and the proof of part 1, which tells us that $\tilde{Q}_0(\nu_{\theta'}) > \tilde{Q}_0(\alpha\nu_\theta) > \psi \cdot \tilde{Q}_0(\nu_\theta)$. □

Proof. **Proof of Lemma 9**: Let $\theta_\eta = (0, \eta, 0)$ and $\theta_\zeta = (0, 0, \zeta)$ be two policies which result in the same service rate. Let $\mu_\theta = \mu_{\theta_\eta} = \mu_{\theta_\zeta}$ and $\nu_\theta = \nu_{\theta_\eta} = \nu_{\theta_\zeta}$. This ensures that both policy achieves the same number of satisfied citizens in expectation. The optimality condition from Lemma 2 tells us that:

$$\frac{r - c}{r - \zeta} = \tilde{Q}_0(\nu_\theta) \cdot (1 + \mathbb{E}[W|\ \nu_\theta]) = \frac{r + \eta - c}{r + \eta}.$$

This implies that $\eta = \frac{(r-c)\cdot\zeta}{c-\zeta}$. We can compare the cost of the two policies as follows:

$$\begin{aligned}
\frac{C(\theta_\eta)}{C(\theta_\zeta)} = \frac{\eta \cdot \mu_\theta \cdot \left(1 - \tilde{Q}_0(\nu_\theta)\right)}{\zeta \cdot \mu_\theta \cdot \tilde{Q}_0(\nu_\theta)} = \frac{r - c}{c - \zeta} \cdot \frac{1 - \tilde{Q}_0(\nu_\theta)}{\tilde{Q}_0(\nu_\theta)} &= \frac{r - c}{c - \zeta} \cdot \left(\frac{1}{\tilde{Q}_0(\nu_\theta)} - 1\right) \\
&= \frac{r - c}{c - \zeta} \cdot \left((1 + \mathbb{E}[W|\ \nu_\theta]) \cdot \frac{r - \zeta}{r - c} - 1\right) \\
&= \frac{r - \zeta}{c - \zeta} \cdot \mathbb{E}[W|\ \nu_\theta] + 1 \geq 1.
\end{aligned}$$

The second line is true by rearranging the optimality condition for ν_θ. Hence, the fee-for-service policy has a higher cost than the downtime-rebate policy for a given target on the number of satisfied citizens.

Finally, if we rearrange the first line of the proof and solve for ζ instead of η, then we will see that $\zeta = c \cdot \frac{\eta}{(r-c)+\eta}$, which implies that ζ grows sublinearly in η because $r > c$. □

Preamble to the proof of Propositions 10 and 11: We begin by proving two

lemmas which investigate the change in the expected number of satisfied citizens when we increases the demand rate and service rate by a factor of $\gamma > 1$. Recall that λ_0, μ_0 are the citizen's demand rate and provider's service rate respectively when there are no subsidies.

Lemma A1. *For some $\gamma > 1$, suppose the government implements the consumer-voucher policy θ such that $\lambda_\theta = \gamma\lambda_0$. Then:*

$$\mu_\theta = \frac{\gamma \cdot (\lambda + \phi)}{\gamma\lambda + \phi} \cdot \mu_0, \quad \textit{and} \quad \mathbb{E}[S|\ \mu_\theta, \lambda_\theta] = \frac{\gamma \cdot (\lambda_0 + \phi)}{\gamma\lambda_0 + \phi} \cdot \mathbb{E}[S|\ \mu_0, \lambda_0].$$

Proof. Proof. Recall that ν_θ depends only on the gross profit margin, and does not change with λ_θ. Hence, $\nu_\theta = \nu_0$ even when we have a faster demand rate. This leads to a corresponding increase in the service provider's service rate:

$$\mu_\theta \cdot \left(\frac{1}{\gamma\lambda_0} + \frac{1}{\phi}\right) = \mu_0 \cdot \left(\frac{1}{\lambda_0} + \frac{1}{\phi}\right) \quad \Rightarrow \quad \mu_\theta = \frac{\gamma \cdot (\lambda_0 + \phi)}{\gamma\lambda_0 + \phi} \cdot \mu_0.$$

As a result, the expected number of served citizens becomes:

$$E[S|\ \mu_\theta, \lambda_\theta] = \frac{\mu_\theta}{\phi} \cdot (1 - \tilde{Q}_0(\nu_\theta)) = \frac{\gamma \cdot (\lambda_0 + \phi)}{\gamma\lambda_0 + \phi} \cdot \frac{\mu_0}{\phi} \cdot (1 - \tilde{Q}_0(\nu_0)) = \frac{\gamma \cdot (\lambda_0 + \phi)}{\gamma\lambda_0 + \phi} \cdot E[S|\ \mu_0, \lambda_0].$$

Before ending the proof, we also point out that:

$$\frac{\mu_\theta}{\lambda_\theta} = \frac{\frac{\gamma\cdot(\lambda_0+\phi)}{\gamma\lambda_0+\phi} \cdot \mu_0}{\gamma\lambda_0} = \frac{\lambda_0 + \phi}{\gamma\lambda_0 + \phi} \cdot \frac{\mu_0}{\lambda_0} < \frac{\mu_0}{\lambda_0}$$

because $\gamma \geq 1$. □

Lemma A2. *For some $\gamma > 1$, suppose the government implements a provider-based policy θ such that $\mu_\theta = \gamma\mu_0$. Then $\nu_\theta = \gamma\nu_0$ and:*

$$\frac{\gamma}{(\gamma^N - 1) \cdot \tilde{Q}_0(\nu_0) + 1} \cdot \mathbb{E}[S|\ \mu_0, \lambda_0] \leq \mathbb{E}[S|\ \mu_\theta, \lambda_\theta] \leq \frac{\gamma}{(\gamma - 1) \cdot \tilde{Q}_0(\nu_0) + 1} \cdot \mathbb{E}[S|\ \mu_0, \lambda_0].$$

Proof. Proof. Since there is no change in the demand rate, ν_0 increases by the same factor of γ. We obtain the upper bound of $E[S|\ \mu_\theta, \lambda_\theta]$ by the proof of Proposition 8, where we can replace α by γ and ψ by $\frac{\gamma}{(\gamma-1)\cdot\tilde{Q}_{\nu_0}+1}$. Next, we upper bound $\tilde{Q}_0(\nu_\theta)$:

$$\begin{aligned}
\tilde{Q}_0(\nu_\theta) &= \left(\sum_{j=0}^{N} \frac{N!}{(N-j)!} \cdot (\gamma\nu_0)^{-j}\right)^{-1} \\
&\leq \left(1 + \sum_{j=1}^{N} \frac{N!}{(N-j)!} \cdot \nu_0^{-j}\gamma^{-N}\right)^{-1} = \frac{\gamma^N}{(\gamma^N - 1) \cdot \tilde{Q}_0(\nu_0) + 1} \cdot \tilde{Q}_0(\nu_0).
\end{aligned}$$

Using the bound on $\tilde{Q}_0(\nu_0)$, we can lower-bound $E[S|\ \mu_\theta, \lambda_\theta]$ as

$$\begin{aligned} E[S|\ \mu_\theta, \lambda_\theta] = \frac{\gamma\mu_0}{\phi} \cdot (1 - \tilde{Q}_0(\gamma\nu_0)) &\geq \frac{\gamma\mu_0}{\phi} \cdot \left(1 - \frac{\gamma^N}{(\gamma^N - 1) \cdot \tilde{Q}_0(\nu_0) + 1} \cdot \tilde{Q}_0(\nu_0)\right) \\ &= \frac{\gamma}{(\gamma^N - 1) \cdot \tilde{Q}_0(\nu_0) + 1} \cdot \frac{\mu_0}{\phi} \cdot (1 - \tilde{Q}_0(\nu_0)) \\ &= \frac{\gamma}{(\gamma^N - 1) \cdot \tilde{Q}_0(\nu_0) + 1} \cdot E[S|\ \mu_0, \lambda_0]. \end{aligned}$$

Putting the two inequalities together gives us our desired bound. □

Proof. **Proof of Proposition 10**: We prove the proposition in two parts.

Part 1: Suppose by contradiction that there exists an alternative fee-for-service policy $\theta_{\eta'} = (0, \eta', 0)$ which dominates θ_κ. We will show that θ_κ dominates θ_η, as defined in the statement of the proposition, and that $\theta_{\eta'}$ cannot exist.

First, we claim that if $\tilde{Q}_0(\nu_0) \in \left[\frac{\lambda_0}{\lambda_0+\phi}, 1\right]$, then $\mathbb{E}[S|\ \mu_{\theta_\kappa}, \lambda_{\theta_\kappa}] \geq \mathbb{E}[S|\ \mu_{\theta_\eta}, \lambda_{\theta_\eta}]$. By rearranging the bound on $\tilde{Q}_0(\nu_0)$, we can obtain:

$$(\lambda_0 + \phi) \cdot (\gamma - 1) \cdot \tilde{Q}_0(\nu_0) \geq \lambda_0 \cdot (\gamma - 1) \quad \Leftrightarrow \quad \frac{\lambda_0 + \phi}{\gamma\lambda_0 + \phi} \geq \frac{1}{(\gamma - 1) \cdot \tilde{Q}_0(\nu_0) + 1}.$$

The second inequality can be obtained by adding $(\lambda_0 + \phi)$ to both sides of the first inequality, and then rearranging the result. We can multiply both sides of the second inequality by $\gamma \cdot \mathbb{E}[S|\ \mu_0, \lambda_0]$ to obtain the desired result:

$$\mathbb{E}[S|\ \mu_{\theta_\kappa}, \lambda_{\theta_\kappa}] = \frac{\gamma \cdot (\lambda_0 + \phi)}{\gamma\lambda_0 + \phi} \cdot \mathbb{E}[S|\ \mu_0, \lambda_0] \geq \frac{\gamma}{(\gamma - 1) \cdot \tilde{Q}_0(\nu_0) + 1} \cdot \mathbb{E}[S|\ \mu_0, \lambda_0] > \mathbb{E}[S|\ \mu_{\theta_\eta}, \lambda_{\theta_\eta}],$$

where the first equality is true due to Lemma A1 and the last inequality is due to Lemma A2.

Next, we need to consider the cost of the two policies. Given the second condition in part 1, we can apply Lemma A1 and Lemma A2 to get:

$$\kappa \leq \frac{\gamma\lambda_0 + \phi}{\lambda_0 + \phi} \cdot \frac{1}{(\gamma^N - 1) \cdot \tilde{Q}_0(\nu_0) + 1} \cdot \eta \leq \frac{\gamma \cdot \mathbb{E}[S|\ \mu_0, \lambda_0]}{\mathbb{E}[S|\ \mu_{\theta_\kappa}, \lambda_{\theta_\kappa}]} \cdot \frac{\mathbb{E}[S|\ \mu_{\theta_\eta}, \lambda_{\theta_\eta}]}{\gamma \cdot \mathbb{E}[S|\ \mu_0, \lambda_0]} \cdot \eta = \frac{\mathbb{E}[S|\ \mu_{\theta_\eta}, \lambda_{\theta_\eta}]}{\mathbb{E}[S|\ \mu_{\theta_\kappa}, \lambda_{\theta_\kappa}]} \cdot \eta,$$

which rearranges to gives us

$$C(\theta_\kappa) = \kappa \cdot \mathbb{E}[S|\ \mu_{\theta_\kappa}, \lambda_{\theta_\kappa}] \leq \eta \cdot \mathbb{E}[S|\ \mu_{\theta_\eta}, \lambda_{\theta_\eta}] = C(\theta_\eta).$$

Hence, θ_κ dominates θ_η. Since we assumed that $\theta_{\eta'}$ dominates θ_κ, we must have: $\mathbb{E}[S|\ \mu_{\theta_{\eta'}}, \lambda_{\theta_{\eta'}}] \geq \mathbb{E}[S|\ \mu_{\theta_\kappa}, \lambda_{\theta_\kappa}] > \mathbb{E}[S|\ \mu_{\theta_\eta}, \lambda_{\theta_\eta}]$ and $C(\theta_{\eta'}) \leq C(\theta_\kappa) \leq C(\theta_\eta)$. However, Corollary 7 implies $C(\theta_\eta) < C(\theta_{\eta'})$ because $\theta_{\eta'}$ results in more satisfied citizens and we have a contradiction.

Part 2: We will show that θ_η dominates θ_κ. We follow the structure from part 1. If

$\tilde{Q}_0(\nu_0) \in \left[0, \frac{\lambda_0}{\lambda_0+\phi} \cdot \frac{\gamma-1}{\gamma^N-1}\right]$, then we can obtain:

$$(\lambda_0+\phi)\cdot(\gamma^N-1)\cdot\tilde{Q}_0(\nu_0) \le \lambda_0\cdot(\gamma-1) \quad \Leftrightarrow \quad \frac{\lambda_0+\phi}{\gamma\lambda_0+\phi} \le \frac{1}{(\gamma^N-1)\cdot\tilde{Q}_0(\nu_0)+1}.$$

To arrive at the second inequality, we add $(\lambda_0+\phi)$ to both sides of the first inequality. We can multiply both sides of the second inequality by $\gamma\cdot\mathbb{E}[S|\ \mu_0,\lambda_0]$ to obtain the desired result:

$$\mathbb{E}[S|\ \mu_{\theta_\kappa},\lambda_{\theta_\kappa}] = \frac{\gamma\cdot(\lambda_0+\phi)}{\gamma\lambda_0+\phi}\cdot\mathbb{E}[S|\ \mu_0,\lambda_0] \le \frac{\gamma}{(\gamma^N-1)\cdot\tilde{Q}_0(\nu_0)+1}\cdot\mathbb{E}[S|\ \mu_0,\lambda_0] \le \mathbb{E}[S|\ \mu_{\theta_\eta},\lambda_{\theta_\eta}],$$

where the first equality is true due to Lemma A1 and the last inequality is due to Lemma A2.

Next, we consider the cost. Given the second condition in part 2, we can apply Lemma A1 and Lemma A2 to get:

$$\kappa \ge \frac{\gamma\lambda_0+\phi}{\lambda_0+\phi}\cdot\frac{1}{(\gamma-1)\cdot\tilde{Q}_0(\nu_0)+1}\cdot\eta \ge \frac{\gamma\cdot\mathbb{E}[S|\ \mu_0,\lambda_0]}{\mathbb{E}[S|\ \mu_{\theta_\kappa},\lambda_{\theta_\kappa}]}\cdot\frac{\mathbb{E}[S|\ \mu_{\theta_\eta},\lambda_{\theta_\eta}]}{\gamma\cdot\mathbb{E}[S|\ \mu_0,\lambda_0]}\cdot\eta = \frac{\mathbb{E}[S|\ \mu_{\theta_\eta},\lambda_{\theta_\eta}]}{\mathbb{E}[S|\ \mu_{\theta_\kappa},\lambda_{\theta_\kappa}]}\cdot\eta,$$

which rearranges to gives us

$$C(\theta_\kappa) = \kappa\cdot\mathbb{E}[S|\ \mu_{\theta_\kappa},\lambda_{\theta_\kappa}] \ge \eta\cdot\mathbb{E}[S|\ \mu_{\theta_\eta},\lambda_{\theta_\eta}] = C(\theta_\eta).$$

Hence θ_η dominates θ_κ and at least one such policy exists.

□

Proof. **Proof of Proposition 11**: We prove the proposition in two parts.

Part 1: Suppose by contradiction that there exists an alternative downtime-rebate policy $\theta_{\zeta'} = (0,0,\zeta')$ which dominates θ_κ, similar to Proposition 10. We will show that θ_κ dominates θ_ζ, and that $\theta_{\zeta'}$ cannot exist.

From the proof of Proposition 10, we know that $\mathbb{E}[S|\ \mu_{\theta_\kappa},\lambda_{\theta_\kappa}] \ge \mathbb{E}[S|\ \mu_{\theta_\zeta},\lambda_{\theta_\zeta}]$. It suffices to focus on the cost. From the second condition in part 1, we can show:

$$\begin{aligned}
\kappa &\le \frac{\gamma\lambda_0+\phi}{\lambda_0+\phi}\cdot\frac{1}{(\gamma-1)\cdot\tilde{Q}_0(\nu_0)+1}\cdot\frac{\gamma\lambda_0}{\phi}\cdot\zeta \\
&\le \frac{\gamma\lambda_0+\phi}{\lambda_0+\phi}\cdot\frac{\gamma}{(\gamma-1)\cdot\tilde{Q}_0(\nu_0)+1}\cdot\frac{\tilde{Q}_0(\nu_0)}{1-\tilde{Q}_0(\nu_0)}\cdot\zeta \\
&\le \frac{\gamma\lambda_0+\phi}{\lambda_0+\phi}\cdot\frac{\tilde{Q}_0(\nu_{\theta_\zeta})}{1-\tilde{Q}_0(\nu_{\theta_\kappa})}\cdot\zeta \\
&= \frac{\mu_{\theta_\zeta}\cdot\tilde{Q}_0(\nu_{\theta_\zeta})}{\mu_{\theta_\kappa}\cdot(1-\tilde{Q}_0(\nu_{\theta_\kappa}))}\cdot\zeta,
\end{aligned}$$

where the second inequality uses $\frac{\tilde{Q}_0(\nu_0)}{1-\tilde{Q}_0(\nu_0)} \ge \frac{\lambda}{\phi}$ for $\tilde{Q}_0(\nu_0) \in \left[\frac{\lambda}{\lambda+\phi}, 1\right]$. The last equality

applies $\mu_{\theta_\zeta} = \gamma\mu_0$ and $\mu_{\theta_\kappa} = \frac{\gamma\cdot(\lambda_0+\phi)}{\gamma\lambda_0+\phi}$ from Lemma A1. Rearranging the above inequality gives us:

$$C(\theta_\kappa) = \kappa\mu_{\theta_\kappa} \cdot \left(1 - \tilde{Q}_0(\nu_{\theta_\kappa})\right) \leq \zeta\mu_{\theta_\zeta} \cdot \tilde{Q}_0(\nu_{\theta_\zeta}) = C(\theta_\zeta).$$

Hence, θ_κ dominates θ_ζ. Since we assumed that $\theta_{\zeta'}$ dominates θ_κ, we must have: $\mathbb{E}[S|\ \mu_{\theta_{\zeta'}}, \lambda_{\theta_{\zeta'}}] \geq \mathbb{E}[S|\ \mu_{\theta_\kappa}, \lambda_{\theta_\kappa}] > \mathbb{E}[S|\ \mu_{\theta_\zeta}, \lambda_{\theta_\zeta}]$ and $C(\theta_{\zeta'}) \leq C(\theta_\kappa) \leq C(\theta_\zeta)$. However, Corollary 7 implies $C(\theta_\zeta) < C(\theta_{\zeta'})$ because $\theta_{\zeta'}$ results in more satisfied citizens and we have a contradiction.

Part 2: We will show that θ_ζ dominates θ_κ. Again, the proof of Proposition 10 tells us that $\mathbb{E}[S|\ \mu_{\theta_\kappa}, \lambda_{\theta_\kappa}] \leq \mathbb{E}[S|\ \mu_{\theta_\zeta}, \lambda_{\theta_\zeta}]$ and it suffices to focus on the costs. Before proceeding, we present a slightly complicated upper-bound on $\frac{\tilde{Q}_0}{1-\tilde{Q}_0}$. For $\tilde{Q}_0 \in \left[0, \frac{\lambda}{\lambda+\phi} \cdot \frac{\gamma-1}{\gamma^N-1}\right]$, observe that:

$$\begin{aligned}\frac{\tilde{Q}_0}{1-\tilde{Q}_0} \leq \frac{\frac{\lambda}{\lambda+\phi} \cdot \frac{\gamma-1}{\gamma^N-1}}{1 - \frac{\lambda}{\lambda+\phi} \cdot \frac{\gamma-1}{\gamma^N-1}} &= \frac{\lambda \cdot (\gamma-1)}{(\lambda+\phi)\cdot(\gamma^N-1) - \lambda\cdot(\gamma-1)} \\ &\leq \frac{\lambda}{\phi \cdot \frac{\gamma^N-1}{\gamma-1} + \lambda \cdot \left(\frac{\gamma^N-1}{\gamma-1} - (\gamma-1)\right)} \leq \frac{\lambda}{\phi} \cdot \frac{\gamma-1}{\gamma^N-1}.\end{aligned}$$

Using the second condition of part 2 of the proposition, we can show that:

$$\begin{aligned}\kappa &\geq \frac{\gamma\lambda_0+\phi}{\lambda_0+\phi} \cdot \frac{1}{(\gamma^N-1)\cdot\tilde{Q}_0(\nu_0)+1} \cdot \frac{\gamma\lambda_0}{\phi} \cdot \zeta \\ &\geq \frac{\gamma\lambda_0+\phi}{\lambda_0+\phi} \cdot \frac{\gamma^N}{(\gamma^N-1)\cdot\tilde{Q}_0(\nu_0)+1} \cdot \frac{\gamma-1}{\gamma^N-1} \cdot \frac{\lambda_0}{\phi} \cdot \zeta \\ &\geq \frac{\gamma\lambda_0+\phi}{\lambda_0+\phi} \cdot \frac{\gamma^N}{(\gamma^N-1)\cdot\tilde{Q}_0(\nu_0)+1} \cdot \frac{\tilde{Q}_0(\nu_0)}{1-\tilde{Q}_0(\nu_0)} \cdot \zeta \\ &\geq \frac{\gamma\lambda_0+\phi}{\lambda_0+\phi} \cdot \frac{\tilde{Q}_0(\nu_{\theta_\zeta})}{1-\tilde{Q}_0(\nu_{\theta_\kappa})} \cdot \zeta \\ &= \frac{\mu_{\theta_\zeta}\cdot\tilde{Q}_0(\nu_{\theta_\zeta})}{\mu_{\theta_\kappa}\cdot(1-\tilde{Q}_0(\nu_{\theta_\kappa}))} \cdot \zeta,\end{aligned}$$

where the last inequality follows from Lemmas A1 and A2, $\mu_{\theta_\zeta} = \gamma\mu_0$, and $\mu_{\theta_\kappa} = \frac{\gamma\cdot(\lambda_0+\phi)}{\gamma\lambda_0+\phi}$. Rearranging the above inequality gives us:

$$C(\theta_\kappa) = \kappa\mu_{\theta_\kappa} \cdot \left(1 - \tilde{Q}_0(\nu_{\theta_\kappa})\right) \geq \zeta\mu_{\theta_\zeta} \cdot \tilde{Q}_0(\nu_{\theta_\zeta}) = C(\theta_\zeta).$$

Hence θ_ζ dominates θ_κ and at least one such policy exists. □

Preamble to the proof of Propositions 12: We begin by proving a lemma on the stationary probabilities in the multi-server model.

Lemma A3. *Under the multi-server model with M service stations, the stationary probabilities are of the form $n_{s,w}(M,\lambda)/\Gamma(M,\lambda)$, where*

$$n_{s,w}(M,\lambda) = \begin{cases} \frac{\lambda^{s+w}}{\phi^s \cdot (M\mu)^w} \cdot \frac{N!}{s!(N-s-w)!} \cdot \frac{M!}{w!\ M^{M-w}} & \text{if } w \le M-1 \\ \frac{\lambda^{s+w}}{\phi^s \cdot (M\mu)^w} \cdot \frac{N!}{s!(N-s-w)!} & \text{if } w \ge M \end{cases}$$

and $\Gamma(M,\lambda) = \sum_{s=0}^{N} \sum_{w=0}^{N-s} n_{s,w}(M,\lambda)$.

Proof. Proof.

For cleanliness, we fix λ and μ and denote $Q_{s,w} = Q_{s,w}(M,\lambda)$. Figure 1 would be modified so that state (s,w) transitions to $(s+1,w-1)$ with rate $\min\{w,M\}\cdot\mu$. Hiding the dependence on λ, M in $n_{s,w}$ and Γ, the balance equations of the M-server model are expressed below:

$$\begin{aligned}
&(N-s-w+1)\cdot\lambda\cdot n_{s,w-1} + \min\{w+1,M\}\cdot\mu\cdot n_{s-1,w+1} + (s+1)\cdot\phi\cdot n_{s+1,w} \\
&= (\min\{w+1,M\}\cdot\mu + s\phi + (N-s-w)\cdot\lambda)\cdot n_{s,w} \quad \text{if } s,w>0; s+w<N \\
&\lambda\cdot n_{s,w-1} + \min\{w+1,M\}\cdot\mu\cdot n_{s-1,w+1} = \big(\min\{w+1,M\}\cdot\mu + s\phi\big)\cdot n_{s,w} \quad \text{if } s,w>0; s+w=N \\
&(N-w+1)\cdot\lambda\cdot n_{0,w-1} + \phi\cdot n_{1,w} = \big(\min\{w+1,M\}\cdot\mu + (N-w)\cdot\lambda\big)\cdot n_{0,w} \quad \text{if } s=0; w\neq 0,N \\
&\lambda\cdot n_{0,N-1} = M\mu\cdot n_{0,N} \quad \text{if } s=0; w=N \\
&\mu\cdot n_{s-1,1} + (s+1)\cdot\phi\cdot n_{s+1,0} = \big(s\phi + (N-s)\cdot\lambda\big)\cdot n_{s,0} \quad \text{if } w=0; s\neq 0,N \\
&\mu\cdot n_{N-1,1} = N\phi\cdot n_{N,0} \quad \text{if } w=0; s=N \\
&\phi\cdot n_{1,0} = N\lambda\cdot n_{0,0} \quad \text{if } w=0; s=0.
\end{aligned}$$

We leave the verification that $n_{s,w}$ satisfies the above balance equation as an exercise, as the proof is similar to Proposition 1. □

Proof. **Proof of Proposition 12**:

For cleanliness of notation, we assume that M and λ are fixed and omit them from the notation. First, we compute $\tilde{Q}_w$ for a general $w = 0,\dots,N$. Based on Lemma A3, observe that $n_{s,w} = \left(\frac{\lambda}{\phi}\right)^s \cdot \binom{N-w}{s} \cdot n_{0,w}$ for any w. Hence, if $w \ge M$:

$$\begin{aligned}
\sum_{s=0}^{N-w} n_{s,w} = n_{0,w}\cdot\sum_{s=0}^{N-w}\binom{N-w}{s}\cdot\left(\frac{\lambda}{\phi}\right)^s &= \frac{N!}{(N-w)!}\cdot\left(\frac{\lambda}{M\mu}\right)^w\cdot\left(\frac{\lambda}{\phi}+1\right)^{N-w} \\
&= \frac{N!}{(N-w)!}\cdot\left(\frac{\lambda}{M\mu}\right)^N\cdot(M\nu)^{N-w}
\end{aligned}$$

Similarly, if $w \le M-1$, the numerator simplifies to:

$$\sum_{s=0}^{N-w} n_{s,w} = n_{0,w}\cdot\sum_{s=0}^{N-w}\binom{N-w}{s}\cdot\left(\frac{\lambda}{\phi}\right)^s = \frac{N!}{(N-w)!}\cdot\left(\frac{\lambda}{M\mu}\right)^N\cdot(M\nu)^{N-w}\cdot\frac{M!}{w!\ M^{M-w}}$$

Specifically, if $w = 0$, then:

$$\begin{aligned}\tilde{Q}_0 &= \frac{\frac{(M\nu)^N}{N!} \cdot \frac{M!}{M^M}}{\sum_{w=0}^{M-1} \frac{(M\nu)^{N-w}}{(N-w)!} \cdot \frac{M!}{w!\, M^{M-w}} + \sum_{w=M}^{N} \frac{(M\nu)^{N-w}}{(N-w)!}} \\ &= \left[\sum_{w=0}^{M-1} \binom{N}{w} \cdot \nu^{-w} + \sum_{w=M}^{N} \binom{N}{w} \cdot \nu^{-w} \cdot \frac{w!}{M!\, M^{w-M}} \right]^{-1};\end{aligned}$$

and if $w > 0$, we can obtain the desired form by observing that $\tilde{Q}_w = \frac{n_{s,w}}{n_{s,0}} \cdot \tilde{Q}_0$ for any $s = 0, \ldots, N$.

Next, we compute the expected number of waiting citizens, $\mathbb{E}[W]$. As in Proposition 1, we can write $\mathbb{E}[W] = N - \mathbb{E}[N - W]$ and proceed as follows:

$$\begin{aligned}\mathbb{E}[W] &= N - \sum_{w=0}^{M-1} (N-w) \cdot \tilde{Q}_0 \cdot \binom{N}{w} \cdot \nu^{-w} - \sum_{w=M}^{N} (N-w) \cdot \tilde{Q}_0 \cdot \binom{N}{w} \cdot \nu^{-w} \cdot \frac{w!}{M!\, M^{w-M}} \\ &= N - \sum_{w=0}^{M-1} \tilde{Q}_0 \cdot \frac{N!}{(N-w-1)!\,(w+1)!} \cdot \frac{\nu^{-(w+1)}}{\nu^{-1}} \cdot (w+1) \\ &\quad - \sum_{w=M}^{N-1} \tilde{Q}_0 \cdot \frac{N!}{(N-w-1)!\,(w+1)!} \cdot \frac{\nu^{-(w+1)}}{\nu^{-1}} \cdot \frac{w!}{M!\, M^{w+1-M}} \cdot (w+1) \cdot M \\ &= N - \nu \cdot \left(\sum_{w=1}^{M} w \cdot \tilde{Q}_0 \cdot \binom{N}{w} \cdot \nu^{-w} + \sum_{w=M+1}^{N} M \cdot \tilde{Q}_0 \cdot \binom{N}{w} \cdot \nu^{-w} \cdot \frac{w!}{M!\, M^{w-M}} \right) \\ &= N - \nu \cdot \mathbb{E}[\min\{W, M\}].\end{aligned}$$

The expected number of satisfied and needy citizens, $\mathbb{E}[S]$ and $\mathbb{E}[D]$, are found in the same manner as in the proof of Proposition 1 because $Q_{s,w}/Q_{0,w} = n_{s,w}/n_{0,w} = \left(\frac{\lambda}{\phi}\right)^s \cdot \binom{N-w}{s}$ for all w. □

Proof. **Proof of Lemma 13:** Fix λ. We abuse notation slightly and let $\tilde{Q}_w(M) = \tilde{Q}_w(M, \lambda)$ be the probability that there are w citizens waiting in steady state with M service stations. We first prove that $\tilde{Q}_w(M) \leq \tilde{Q}_w(M+1)$ for $w \leq M-1$ by using the relationship in Proposition 12:

$$\begin{aligned}\frac{\tilde{Q}_w(M+1)}{\tilde{Q}_w(M)} = \frac{\tilde{Q}_0(M+1)}{\tilde{Q}_0(M)} &= \frac{\left[\sum_{w'=0}^{M} \binom{N}{w'} \cdot \nu^{-w'} + \sum_{w'=M+1}^{N} \binom{N}{w'} \cdot \nu^{-w'} \cdot \frac{w'!}{(M+1)!\,(M+1)^{w'-(M+1)}} \right]^{-1}}{\left[\sum_{w'=0}^{M-1} \binom{N}{w'} \cdot \nu^{-w'} + \sum_{w'=M}^{N} \binom{N}{w'} \cdot \nu^{-w'} \cdot \frac{w'!}{M!\, M^{w'-M}} \right]^{-1}} \\ &= \frac{\sum_{w'=0}^{M} \binom{N}{w'} \cdot \nu^{-w'} + \sum_{w'=M+1}^{N} \binom{N}{w'} \cdot \nu^{-w'} \cdot \frac{w'!}{M!\, M^{w'-M}}}{\sum_{w'=0}^{M} \binom{N}{w'} \cdot \nu^{-w'} + \sum_{w'=M+1}^{N} \binom{N}{w'} \cdot \nu^{-w'} \cdot \frac{w'!}{M!\,(M+1)^{w'-M}}} > 1,\end{aligned}$$

where the last equality comes from cleaning up the M-th term in the denominator. At

$w' = M$, we also have:

$$\frac{\tilde{Q}_M(M+1)}{\tilde{Q}_M(M)} = \frac{\tilde{Q}_0(M+1)}{\tilde{Q}_0(M) \cdot \frac{M!}{M!M^0}} = \frac{\tilde{Q}_0(M+1)}{\tilde{Q}_0(M)} > 1,$$

where the analysis is the same as the case with $w' \leq M-1$.

In contrast, if $w \geq M+1$, then we can show that there exist at least one or more w such that $\tilde{Q}_w(M) \geq \tilde{Q}_w(M+1)$:

$$\frac{\tilde{Q}_w(M+1)}{\tilde{Q}_w(M)} = \frac{\tilde{Q}_0(M+1)}{\tilde{Q}_0(M)} \cdot \left(\frac{M}{M+1}\right)^{w-M},$$

which is decreasing in w.

Since $\sum_{w=0}^{N} \tilde{Q}_w(M) = \sum_{w=0}^{N} \tilde{Q}_w(M+1) = 1$ and $\tilde{Q}_w(M) < \tilde{Q}_w(M+1)$ for all $w \leq M$, there exist $\bar{w} \geq M+1$ such that $\tilde{Q}_w(M) > \tilde{Q}_w(M+1)$ for all $w \geq \bar{w}$ and $\tilde{Q}_w(M) < \tilde{Q}_w(M+1)$ for all $w < \bar{w}$. Turning to $\mathbb{E}[W|\ M, \lambda]$, we have:

$$\begin{aligned}
\mathbb{E}[W|\ M+1, \lambda] - \mathbb{E}[W|\ M, \lambda] =& \sum_{w=0}^{\bar{w}-1} w \cdot (\tilde{Q}_w(M+1) - \tilde{Q}_w(M)) + \sum_{w=\bar{w}}^{N} w \cdot (\tilde{Q}_w(M+1) - \tilde{Q}_w(M)) \\
<& \sum_{w=0}^{\bar{w}-1} \bar{w} \cdot (\tilde{Q}_w(M+1) - \tilde{Q}_w(M)) + \sum_{w=\bar{w}}^{N} \bar{w} \cdot (\tilde{Q}_w(M+1) - \tilde{Q}_w(M)) \\
=& \bar{w} \cdot \left(\sum_{w=0}^{N} \tilde{Q}_w(M+1) - \sum_{w=0}^{N} \tilde{Q}_w(M) \right) = 0
\end{aligned}$$

The strict inequality comes from replacing w with the larger $\bar{w} > w$ whenever $\tilde{Q}_w(M+1) - \tilde{Q}_w(M) > 0$, and replacing w with the smaller $\bar{w} \leq w$ whenever $\tilde{Q}_w(M+1) - \tilde{Q}_w(M) < 0$. By Proposition 12, $\mathbb{E}[\min\{W, M\}|\ M, \lambda]$ increases with M, and $\mathbb{E}[S|\ M, \lambda]$, $\mathbb{E}[D|\ M, \lambda]$ must also increase with M.

□

Proof. **Proof of Lemma 14: Step 1: Constraint reduction.** Since the objective is increasing in U_i, the government lowers U_i until constraints bind. First, IR_H must bind, i.e., $U_H = 0$. If $U_H > 0$, we can reduce both U_H and U_L by a small $\epsilon > 0$ while maintaining IR, satisfying all constraints and reducing cost. Second, IC_L must bind. If $U_L > (c_H - c_L)\mu_H$, we can reduce U_L by a small $\epsilon > 0$ to reduce costs without violating IR. Finally, the target constraint must bind, otherwise, the government could reduce μ_L, μ_H to lower costs. Thus, we proceed with $U_H = 0$ and $U_L = (c_H - c_L)\mu_H$.

Step 2: Solving the relaxed problem. We first consider the government problem without the constraints IC_H and IR_L. Substituting the binding utilities into the objective function, the government minimizes:

$$\min_{\mu_L, \mu_H} \quad \beta(c_L\mu_L + (c_H - c_L)\mu_H) + (1-\beta)c_H\mu_H, \tag{A1}$$

subject to the target constraint $\beta R(\mu_L) + (1-\beta)R(\mu_H) = \mathcal{T}$. Let $\lambda > 0$ be the Lagrange multiplier for the target constraint. The first order conditions (FOCs) with respect to μ_L and μ_H are:

$$\frac{\partial \mathcal{L}}{\partial \mu_L} = \beta c_L - \lambda \beta R'(\mu_L) = 0 \implies R'(\mu_L^*) = \frac{c_L}{\lambda}, \tag{A2}$$

$$\frac{\partial \mathcal{L}}{\partial \mu_H} = [(1-\beta)c_H + \beta \Delta c] - \lambda(1-\beta)R'(\mu_H) = 0 \implies R'(\mu_H^*) = \frac{1}{\lambda}\left(c_H + \frac{\beta}{1-\beta}(c_H - c_L)\right). \tag{A3}$$

Step 3: Proving $\mu_L^* > \mu_H^*$. We compare the marginal production requirements derived in (A2) and (A3). Note that $c_H + \frac{\beta}{1-\beta}\Delta c > c_H > c_L$. Since $\lambda > 0$, it follows directly that $R'(\mu_H^*) > R'(\mu_L^*)$, which suggests that $\mu_L^* > \mu_H^*$ by the concavity of $R(\cdot)$. This confirms that the efficient provider operates at a higher rate, while the inefficient provider's rate is distorted downward to reduce the information rent (U_i) paid to the efficient type.

Step 4: Verification of ignored constraints. We now turn back and verify that the solution satisfies IR_L and IC_H. First, we have $U_L^* = (c_H - c_L)\mu_H^*$. Since $\mu_H^* > 0$ and $(c_H - c_L) > 0$, $U_L^* > 0$. Thus IR_L is satisfied. Next, by substituting $U_H^* = 0$ and $U_L^* = (c_H - c_L)\mu_H^*$, the IC_H constraint requires $0 \geq (c_H - c_L)\mu_H^* - (c_H - c_L)\mu_L^*$, which is redundant since we proved $\mu_L^* > \mu_H^*$ in Step 3.

□

Proof. **Proof of Proposition 15:** Let $m_i = \eta_i - \zeta_i$ be the marginal subsidy in contract i. We first prove that the $m_L > m_H$. We prove it by matching the provider's incentives with the optimal allocation. The utility that provider i can obtain from choosing contract i is $m_i R(\mu_i) - c_i \mu_i$. To implement the target rate μ_i^*, the marginal subsidy must satisfy the provider's first-order condition at the target: $m_i R'(\mu_i^*) = c_i$.

Substituting equation (A2) into the expression for m_L gives us $m_L = \frac{c_L}{c_L/\lambda} = \lambda$. Similarly, substituting equation (A3) into the expression for m_H gives us $m_H = \lambda \cdot \frac{c_H}{c_H + \frac{\beta}{1-\beta}(c_H - c_L)} < \lambda$. Consequently, we have $m_L > m_H$.

Next we prove that $\zeta_H > \zeta_L$. We use the binding incentive compatibility constraint for the efficient type: $U_L^* = U_H^* + (c_H - c_L)\mu_H^*$. Recall that $U_H^* = 0$, and the utility under truth-telling is $U_L^* = P_L^* - c_L \mu_L^*$. The deviation profit for type-L provider mimicking H is $\Pi(H|L) = P_H^* - c_L \mu_H^*$. Substituting these into the binding IC_L gives us $P_L^* - c_L \mu_L^* = P_H^* - c_L \mu_H^*$. By substituting the payment definition $P_i^* = m_i R_i + \zeta_i K$ where $R_i \equiv R(\mu_i^*)$, we have $(m_L R_L + \zeta_L K) - c_L \mu_L^* = (m_H R_H + \zeta_H K) - c_L \mu_H^*$. Rearranging terms gives us that $(\zeta_H - \zeta_L)K = (m_L R_L - m_H R_H) - c_L(\mu_L^* - \mu_H^*)$. We claim that the right-hand side is positive.

Define $g(\mu) = m_L R(\mu) - c_L \mu$. Since μ_L^* is determined such that $m_L R'(\mu_L^*) = c_L$, it is the unique maximizer of $g(\mu)$ by the concavity of $R(\mu)$. Then we have

$$m_L R_L - c_L \mu_L^* > m_L R_H - c_L \mu_H^* > m_H R_H - c_L \mu_H^*,$$

since $\mu_H^* < \mu_L^*$ and $m_H < m_L$. Thus, $(\zeta_H - \zeta_L)K > 0$ and hence $\zeta_H > \zeta_L$.

Then we prove $\eta_L > \eta_H$. By definition, $\eta_i = m_i + \zeta_i$. We analyze the difference $\eta_L - \eta_H =$

$(m_L - m_H) - (\zeta_H - \zeta_L)$. Multiplying by K and substituting $(\zeta_H - \zeta_L)K$ in the proof of $\zeta_H > \zeta_L$ gives us

$$K(\eta_L - \eta_H) = K(m_L - m_H) - [(m_L R_L - m_H R_H) - c_L(\mu_L^* - \mu_H^*)] \tag{A4}$$

$$= m_L(K - R_L) - m_H(K - R_H) + c_L(\mu_L^* - \mu_H^*) + (m_H - m_H)(K - R_L), \tag{A5}$$

$$= (m_L - m_H)(K - R_L) + m_H(R_H - R_L) + c_L(\mu_L^* - \mu_H^*), \tag{A6}$$

which is strictly positive by the assumption that $K > R(\mu_L^*) + \frac{(\eta_H - \zeta_H)[R(\mu_L^*) - R(\mu_H^*)] - c_L(\mu_L^* - \mu_H^*)}{(\eta_L - \zeta_L) - (\eta_H - \zeta_H)}$. Thus, we have $\eta_L > \eta_H$.

□

B Approximation Algorithm for Mixed-Subsidy Policies

In this section, we introduce the algorithm that generates suboptimal mixed policies with a provable performance guarantee. Before diving into the design of the algorithm, we first present a counterexample for a mixed-subsidy policy of the form $\theta = (\kappa, \eta, 0)$. The service provider charges a price of $r = 10$ and incurs operating cost $c = 9$. Let $\phi = 1$. Let the demand rate be $\lambda_\theta = f(\kappa) = \frac{1}{1+e^{-\frac{\kappa-5}{0.5}}}$. Consider two policies $\theta_1 = (4, 2, 0)$ and $\theta_2 = (8, 8, 0)$. Then their costs to the government are $C(\theta_1) = 54.51$ and $C(\theta_2) = 720.13$ respectively. If we take the average policy with $\bar{\theta} = (6, 5, 0)$, then we incur a cost of $C(\bar{\theta}) = 454.78 > 387.32 = \frac{1}{2} \cdot (C(\theta_1) + C(\theta_2))$. Hence, the government's cost is not jointly convex in κ and η. Similarly, we can find a counterexample for mixed-subsidy policies of the form $\theta = (\kappa, 0, \zeta)$ under the same setup. Consider two policies $\theta_1 = (4, 0, 2)$ and $\theta_2 = (8, 0, 8)$. In this case, $C(\theta_1) = 34.84$ and $C(\theta_2) = 377.64$, but the average policy $\bar{\theta} = (6, 0, 5)$ incurs a cost of $C(\bar{\theta}) = 237.43 > 206.24 = \frac{1}{2} \cdot (C(\theta_1) + C(\theta_2))$.

Considering the difficulty of finding the optimal mixed-subsidy policy θ^*, we will construct an algorithm to find a near-optimal policy $\hat{\theta}$, such that $C(\hat{\theta}) \leq (1+\epsilon) \cdot C(\theta^*)$ and $\epsilon \in (0, 1)$ is an accuracy parameter that we can control. The purpose of constructing such an algorithm is to give us a method to investigate properties of the near-optimal mixed-subsidy policies via numerical experiments.

From hereon we will focus on the mixed-subsidy policy $\theta = (\kappa, 0, \zeta)$; the results are analogous for $\theta = (\kappa, \eta, 0)$. Before proceeding, we present a useful property which considers the impact of only increasing the voucher or the rebate. The proof is omitted because it follows directly from Lemma 3, Lemma 6, Corollary 4, and Corollary 7.

Corollary B4. *Consider a mixed-subsidy policy of the form $\theta = (\kappa, 0, \zeta)$. For any fixed κ (resp. ζ), $\mathbb{E}[S|\ \mu_\theta, \lambda_\theta]$ and $C(\theta)$ strictly increase as ζ (resp. κ) increases.*

Corollary B4 suggests that the optimal policy satisfies $\mathbb{E}[S|\ \mu_\theta, \lambda_\theta]/N = B$ if B is a feasible target. This corollary suggests that we can construct an algorithm by discretizing

the possible values of κ and ζ to build a set of candidate policies $\theta = (\kappa, 0, \zeta)$ which satisfy $\mathbb{E}[S|\ \mu_\theta, \lambda_\theta]/N = B$.

A direct discretization of the value of κ and ζ is not ideal for algorithmic efficiency. In order to compute ν_θ for each ζ, we would need to run a bisection search based on the service provider's optimality condition before we can compute $\mathbb{E}[S|\ \mu_\theta, \lambda_\theta]$. However, we can use the one-to-one correspondence between ζ and ν_θ to avoid the bisection search because we can recover the corresponding rebate directly via the formula $g_\zeta(\nu) = r - \frac{r-c}{\tilde{Q}_0(\nu)\cdot(1+\mathbb{E}[W|\ \nu])}$ (see Section 4.2). The algorithm is presented in Algorithm 1.

More specifically, we discretize the possible values of the voucher using a geometric sequence. Let $\bar{\kappa}$ denote a positive maximum voucher, which we can generally set to $\bar{\kappa} = r$. Let $\underline{\kappa}$ denote a positive minimum voucher which we will define later. We consider a sequence $K = \{\kappa \mid \kappa = \bar{\kappa} \cdot (1 + \frac{\epsilon}{2})^{-d},\ \underline{\kappa} \leq \kappa \leq \bar{\kappa},\ d \in \mathbb{N}_{\geq 0}, \} \cup \{0\}$. We will show later that either there exists $\hat{\kappa} \in K$ such that $\hat{\kappa}/\left(1 + \frac{\epsilon}{2}\right) < \kappa^* \leq \hat{\kappa}$, or the cost of the vouchers is negligible.

For each value of $\kappa \in K$, we first determine whether it is possible to achieve the government's target. Define $\nu^{\min}, \nu^{\max}$ as the minimum and maximum values of ν that we need to consider. The value of $\nu^{\max}$ can reflect the maximum service rate that can be handled by the service provider. The value of $\nu^{\min}$ could be set to ν_0, or the service provider's decision when no subsidies are given, but the value of ν_0 still needs to be computed via a bisection search. We claim that we can focus on $1 \leq \nu \leq N^2$ if the service provider operates with realistic gross profit margins. Hence, after computing ν_0 with a bisection search on $[1, N^2]$, we can set $\nu^{\min} = \nu_0$ and $\nu^{\max} = N^2$, where the latter term should be adjusted to reflect the maximum feasible service rate, if known.

Claim B5. *If $\nu \leq 1$, then $\frac{r+\eta-c}{r+\eta-\zeta} < \frac{1}{(N-1)!}$. If $\nu \geq N^2$, then $\frac{r+\eta-c}{r+\eta-\zeta} > \frac{N-1}{N}$.*

Proof. Proof: Since $\tilde{Q}_0(\nu)$ is increasing in ν, we can simply consider $\nu = 1$ and $\nu = N^2$. If $\nu = 1$, then:

$$\frac{r+\eta-c}{r+\eta-\zeta} = \tilde{Q}_0(1) \cdot (1 + \mathbb{E}[W|\ 1]) \leq \frac{1}{\sum_{w=0}^{N} \frac{N!}{(N-w)!}} \cdot (1+N) \leq \frac{1}{2 \cdot N!}(1+N) \leq \frac{1}{(N-1)!}$$

If $\nu = N^2$, then:

$$\frac{r+\eta-c}{r+\eta-\zeta} = \tilde{Q}_0(N^2) \cdot \left(1 + \mathbb{E}[W|\ N^2]\right) \geq \frac{1}{\sum_{w=0}^{N} \frac{N^2}{(N-w)!(N^2)^w}} \cdot (1+0) \geq \frac{1}{\sum_{w=0}^{\infty} \left(\frac{1}{N}\right)^w} = \frac{N-1}{N}$$

The second inequality comes from lower-bounding each term in the denominator: $\frac{N}{N^2} \cdot \frac{N-1}{N^2} \cdots \frac{N-w+1}{N^2} \leq \left(\frac{1}{N}\right)^w$. $\square$

Let $\bar{\zeta} = g_\zeta(\nu^{\max})$ and let $\hat{\kappa} \in K$ be the current voucher value. If the initial $\overline{\theta} = (\hat{\kappa}, 0, \bar{\zeta}) = (\hat{\kappa}, 0, \bar{\zeta})$ results in $\mathbb{E}[S|\ \mu_\theta, \lambda_\theta]/N < B$, then the government's policy is infeasible for all $\kappa \leq \hat{\kappa}$ and we can ignore the smaller voucher values in K (lines 11-12). On the other hand, if the initial $\underline{\theta} = (\hat{\kappa}, 0, g_\zeta(\underline{\nu})) = (\hat{\kappa}, 0, 0)$ already achieves the government's target, then a bisection search over ν to find the corresponding rebate is not necessary. Hence, we only need to proceed into the bisection search for ν if $\mathbb{E}[S|\ \mu_{\underline{\theta}}, \lambda_{\underline{\theta}}]/N \leq B \leq \mathbb{E}[S|\ \mu_{\overline{\theta}}, \lambda_{\overline{\theta}}]/N$. Define ν_κ

and ζ_κ to be the policy values which would achieve equality. In each step of the bisection search (lines 16-23), we maintain a lower estimate $\underline{\nu} \geq \nu^{\min}$ and an upper estimate $\overline{\nu} \leq \nu^{\max}$ such that $\underline{\nu} \leq \nu_\kappa \leq \overline{\nu}$. For $\hat{\lambda} = f(\kappa)$ and $\underline{\mu}$, $\overline{\mu}$ constructed from $\underline{\nu}$, $\overline{\nu}$, the bisection search ensures that $\mathbb{E}[S|\ \underline{\mu}, \hat{\lambda}]/N \leq B \leq \mathbb{E}[S|\ \overline{\mu}, \hat{\lambda}]/N$.

To end the bisection search on ν for the current κ, we depart from the standard literature and use the cost of the rebates instead of the values of $\underline{\nu}$, $\overline{\nu}$ to define our termination condition. Define $c_\zeta(\theta) := \zeta \cdot \mu_\zeta \cdot \tilde{Q}_0(\nu_\theta)$ as the cost due to rebates. The purpose of using $c_\zeta(\theta)$ to build the termination condition is to ensure that we can measure the performance loss when we end our bisection search. The termination condition is presented on line 16 and ensures that the cost of rebates is off by a factor of at most $1 + \epsilon/4$ for the two policies being considered.

So far, we have not defined the smallest positive voucher, $\underline{\kappa}$. To determine whether the current κ is small enough, we need to ensure that vouchers make up a negligible portion of the total cost to the government. In line 25 of Algorithm 1, we construct an auxiliary policy by decreasing the rebate slightly and removing the subsidy. If the cost of the auxiliary policy is almost the same as the candidate policy, then we can terminate the search over grid K. Otherwise, we reduce the value of κ. Note that if we enter line 13 of the algorithm because the voucher is sufficiently large and the initial $\underline{\theta}$ satisfy $\mathbb{E}[S|\ \mu_{\underline{\theta}}, \lambda_{\underline{\theta}}]/N \geq B$, then we always proceed to reduce the value of κ on line 25 because θ^- would be the no-subsidy policy in this case.

Algorithm 1 ends by considering the single-subsidy policies as special cases in line 32. It returns the policy with the lowest cost among all policies which can achieve the government's target. Proposition B6 proves that Algorithm 1 identifies a mixed-subsidy policy $\hat{\theta}$ such that the total cost is within a $(1+\epsilon)$-factor of optimality.

Proposition B6. *Let $\hat{\theta} = (\hat{\kappa}, 0, \hat{\zeta})$ be the near-optimal policy returned by Algorithm 1, and let $\theta^* = (\kappa^*, 0, \zeta^*)$ be the optimal policy. Then $C(\hat{\theta}) \leq (1+\epsilon) \cdot C(\theta^*)$.*

Proof. Proof:

If the policy $\theta^{\max} = (\bar{\kappa}, 0, \bar{\zeta})$ cannot achieve the government's target so that $\mathbb{E}[S|\ \mu_{\theta^{\max}}, \lambda_{\theta^{\max}}]/N < B$, then no mixed-subsidy policy can achieve B and the government's target is infeasible. If single-subsidy policies are optimal, then they are captured on line 32. Hence, we may focus on $\kappa^*, \zeta^* > 0$. Let $\underline{\kappa}$ be the smallest positive voucher considered in lines 8-29 of the algorithm.

Case 1: $\kappa^* \geq \underline{\kappa}$

There exists $\tilde{\kappa} \in K$ such that $\frac{\tilde{\kappa}}{(1+\frac{\epsilon}{2})} < \kappa^* \leq \tilde{\kappa}$. We would have entered the bisection search on lines 16-23, so define $\overline{\zeta} = g_\zeta(\overline{\nu})$ and $\underline{\zeta} = g_\zeta(\underline{\nu})$ based on termination of the bisection search for $\tilde{\kappa}$. By construction, $\tilde{\theta} = (\tilde{\kappa}, 0, \overline{\zeta})$ ensures that $\mathbb{E}[S|\ \mu_{\tilde{\theta}}, \lambda_{\tilde{\theta}}]/N > B$.

Define ζ_κ to be the rebate such that $\theta_\kappa = (\tilde{\kappa}, 0, \zeta_\kappa)$ would result in $\mathbb{E}[S|\ \mu_{\theta_\kappa}, \lambda_{\theta_\kappa}]/N = B$. That is, θ_κ ensures that there is no approximation error from the bisection search on rebates at $\tilde{\kappa}$. Corollary B4 implies that $\zeta_\kappa \leq \overline{\zeta}$. We bound the cost of the rebates and vouchers separately.

Total cost of rebates: To bound the cost of rebates to the government, observe that $c_\zeta(\tilde{\theta}) \leq \left(1 + \frac{\epsilon}{4}\right) \cdot c_\zeta(\theta_\kappa)$ by the termination condition on the bisection search. We will focus on bounding the cost of rebates from θ_κ against the optimal θ^*. To do so, we first need to show that $\nu_{\theta_\kappa} \leq \nu_{\theta^*}$.

To prove $\nu_{\theta_\kappa} \leq \nu_{\theta^*}$, we can equivalently prove that $\zeta_\kappa \leq \zeta^*$. Consider an alternative policy $\theta' = (\kappa^*, 0, \zeta_\kappa)$. Then we must have:

$$\mathbb{E}[S|\ \mu_{\theta^*}, \lambda_{\theta^*}] = \mathbb{E}[S|\ \mu_{\theta_\kappa}, \lambda_{\theta_\kappa}] \geq \mathbb{E}[S|\ \mu_{\theta'}, \lambda_{\theta'}],$$

where the equality is true because θ_κ and θ^* ensure that a fraction B of the population is satisfied. The inequality applies Corollary B4 to θ_κ and θ' by observing that both policies offer the same rebate ζ_κ, but $\tilde{\kappa} \geq \kappa^*$. By comparing θ^* and θ', which both offer the same voucher κ^*, we can conclude that $\zeta^* \geq \zeta_\kappa$, again by applying Corollary B4.

Next, we prove that $\mu_{\theta_\kappa} \leq \mu_{\theta^*}$. Observe that both policies achieve the government's target exactly, so that:

$$\mu_{\theta^*} \cdot \left(1 - \tilde{Q}_0(\nu_{\theta^*})\right) = \phi \cdot \mathbb{E}[S|\ \mu_{\theta^*}, \lambda_{\theta^*}] = \phi \cdot \mathbb{E}[S|\ \mu_{\theta_\kappa}, \lambda_{\theta_\kappa}] = \mu_{\theta_\kappa} \cdot \left(1 - \tilde{Q}_0(\nu_{\theta_\kappa})\right).$$

Since we showed that $\nu_{\theta_\kappa} \leq \nu_{\theta^*}$, we must have $1 - \tilde{Q}_0(\nu_{\theta_\kappa}) \geq 1 - \tilde{Q}_0(\nu_{\theta^*})$, which implies that $\mu_{\theta_\kappa} \leq \mu_{\theta^*}$. Finally, we can bound the cost of rebates as follows:

$$c_\zeta(\tilde{\theta}) \leq \left(1 + \frac{\epsilon}{4}\right) \cdot c_\zeta(\theta_\kappa) = \left(1 + \frac{\epsilon}{4}\right) \cdot \zeta_\kappa \cdot \mu_{\theta_\kappa} \cdot \tilde{Q}_0(\nu_{\theta_\kappa}) \leq \left(1 + \frac{\epsilon}{4}\right) \cdot \zeta^* \cdot \mu_{\theta^*} \cdot \tilde{Q}_0(\nu_{\theta^*}) = \left(1 + \frac{\epsilon}{4}\right) \cdot c_\zeta(\theta^*),$$

where the second inequality follows from $\zeta_\kappa \leq \zeta^*$, $\nu_{\theta_\kappa} \leq \nu_{\theta^*}$, and $\mu_{\theta_\kappa} \leq \mu_{\theta^*}$ proven earlier.

Total cost of vouchers: To bound the cost of vouchers to the government, we need to compare the service rates under $\tilde{\theta}$ and θ_κ. Again, we use the termination condition based on the cost of rebates:

$$\overline{\zeta} \cdot \mu_{\tilde{\theta}} \cdot \tilde{Q}_0(\nu_{\tilde{\theta}}) = c_\zeta(\tilde{\theta}) \leq \left(1 + \frac{\epsilon}{4}\right) \cdot c_\zeta(\theta_\kappa) = \left(1 + \frac{\epsilon}{4}\right) \cdot \zeta_\kappa \cdot \mu_{\theta_\kappa} \cdot \tilde{Q}_0(\nu_{\theta_\kappa}),$$

which rearranges to

$$\frac{\mu_{\tilde{\theta}}}{\mu_{\theta_\kappa}} \leq \left(1 + \frac{\epsilon}{4}\right) \cdot \frac{\zeta_\kappa \cdot \tilde{Q}_0(\nu_{\theta_\kappa})}{\overline{\zeta} \cdot \tilde{Q}_0(\nu_{\tilde{\theta}})} \leq \left(1 + \frac{\epsilon}{4}\right).$$

The last inequality is true because $\zeta_\kappa \leq \overline{\zeta}$ and $\theta_\kappa, \tilde{\theta}$ offer the same voucher. Hence, we have $\mu_{\theta_\kappa} \leq \mu_{\tilde{\theta}} \leq \left(1 + \frac{\epsilon}{4}\right) \cdot \mu_{\theta_\kappa}$. The total cost of vouchers is bounded as follows:

$$\begin{aligned}
\tilde{\kappa} \cdot \mu_{\tilde{\theta}} \cdot \left(1 - \tilde{Q}_0(\nu_{\tilde{\theta}})\right) \leq \tilde{\kappa} \cdot \left(1 + \frac{\epsilon}{4}\right) \cdot \mu_{\theta_\kappa} \cdot \left(1 - \tilde{Q}_0(\nu_{\theta_\kappa})\right) &= \left(1 + \frac{\epsilon}{4}\right) \cdot \tilde{\kappa} \cdot \phi \cdot \mathbb{E}[S|\ \mu_{\theta_\kappa}, \lambda_{\theta_\kappa}] \\
&\leq \left(1 + \frac{\epsilon}{4}\right) \cdot \left(1 + \frac{\epsilon}{2}\right) \cdot \kappa^* \cdot \phi \cdot \mathbb{E}[S|\ \mu_{\theta^*}, \lambda_{\theta^*}] \\
&\leq (1 + \epsilon) \cdot \kappa^* \cdot \mu_{\theta^*} \cdot \left(1 - \tilde{Q}_0(\nu_{\theta^*})\right).
\end{aligned}$$

The first inequality uses $\mu_{\tilde{\theta}} \leq \left(1 + \frac{\epsilon}{4}\right) \cdot \mu_{\theta_\kappa}$ and $\nu_{\theta_\kappa} \leq \nu_{\tilde{\theta}}$. The second inequality uses the definition of $\tilde{\kappa}$, as well as the fact that both θ_κ and θ^* satisfy the government's target exactly. By combining the two sets of costs above, we prove that $C(\hat{\theta}) \leq C(\tilde{\theta}) \leq (1 + \epsilon) \cdot C(\theta^*)$.

Case 2: $\kappa^* < \underline{\kappa}$

Define $\tilde{\kappa} = \underline{\kappa}$. If we terminate at $\underline{\kappa}$ because it could not satisfy the government's target

in lines 4-5, then the target would not have been achievable with a voucher of $\kappa^* < \underline{\kappa}$ and we have a contradiction. Hence, we must have started the bisection search for the corresponding rebate. We use the definitions of $\underline{\zeta}, \overline{\zeta}, \zeta_\kappa$, and $\tilde{\theta} = (\tilde{\kappa}, 0, \overline{\zeta})$ from Case 1.

We abuse notation slightly and directly use a tuple to represent the policies for ease of comparing the policies when we modify one parameter at a time. Hence, we have:

$$\begin{aligned} C(\tilde{\theta}) = C((\tilde{\kappa}, 0, \overline{\zeta})) \leq (1+\epsilon) \cdot C((0,0,\underline{\zeta})) &\leq (1+\epsilon) \cdot C((0,0,\zeta_\kappa)) \\ &\leq (1+\epsilon) \cdot C((\kappa^*, 0, \zeta_\kappa)) \leq (1+\epsilon) \cdot C((\kappa^*, 0, \zeta^*)). \end{aligned}$$

The first inequality is due to the termination condition in lines 25-26. In the fourth inequality, the proof that $\zeta_\kappa \leq \zeta^*$ follows $\tilde{\kappa} \geq \kappa^*$ and the discussion in Case 1. As such, $C(\hat{\theta}) \leq C(\tilde{\theta}) \leq (1+\epsilon) \cdot C(\theta^*)$. $\square$

C Robustness Checks

In this appendix, we run several sets of numerical experiments to check the robustness of our policy insights.

C1 Convex Operating Costs

In the main text, we assumed a linear operating cost structure $c\mu$ for tractability. However, in many service settings, particularly healthcare, the marginal cost of increasing capacity may rise due to overtime pay, staff fatigue, or the nonlinear costs of adding new resources. In this section, we examine the results under a convex cost function.

We consider a convex cost function $C(\mu)$ satisfying $C'(\mu) > 0$ and $C''(\mu) > 0$. The service provider's profit maximization problem under policy θ becomes:

$$\max_{\mu} \Pi_\theta(\mu) = \max_{\mu} (r+\eta) \cdot \mu \cdot (1 - \tilde{Q}_0(\mu, \lambda_\theta)) + \zeta\mu \cdot \tilde{Q}_0(\mu, \lambda_\theta) - C(\mu). \tag{C1}$$

Note that it is more difficult to establish the structural property of $\Pi_\theta(\mu)$ in eq. (C1) since $C(\mu)$ may not be well expressed in terms of the utilization ν.

We conduct numerical experiments to examine the robustness of our policy insights. We adopt a power function form of $C(\mu)$ where $C(\mu) = c_0\mu^k$, with $k = 1.15$ to represent accelerating marginal costs. We calibrate c_0 such that the baseline service rate remains consistent with the linear model in the main text. The results are summarized in Figures C1-C4.

It is worth noting that the introduction of convex costs reveals a critical distinction in the **feasibility boundaries** of the subsidy policies. Specifically, when the profit margin is extremely low, the single-subsidy policies are infeasible even for the lowest target, as shown in the blank spaces in Figure C1 and Figure C3.

It is intuitive that the simple consumer voucher could be insufficient to induce higher provider capacity when her profit margin is too low, especially under convex costs. Regarding the two provider-based policies, at an extremely low profit margin, the downtime-rebate

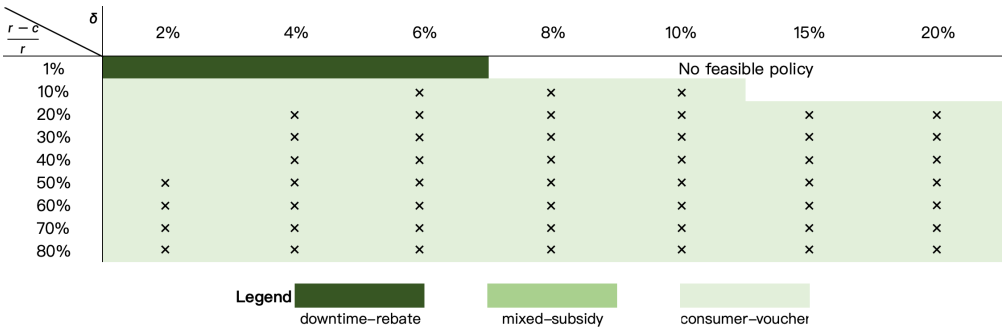


Figure C1: Comparison of the consumer-voucher policy and the fee-for-service policy. The darker cells indicate that the fee-for-service policy has a lower cost than the consumer-voucher policy for the corresponding δ and $\frac{r-c}{r}$. An $\times$ indicates that the fee-for-service policy is infeasible.

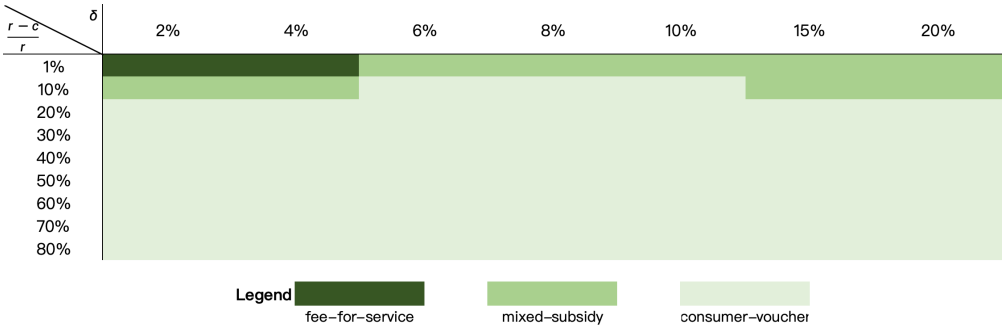


Figure C2: Comparison of the mixed-subsidy policy $\hat{\theta} = (\kappa, \eta, 0)$ against the optimal single-subsidy policies distributing either κ or η.

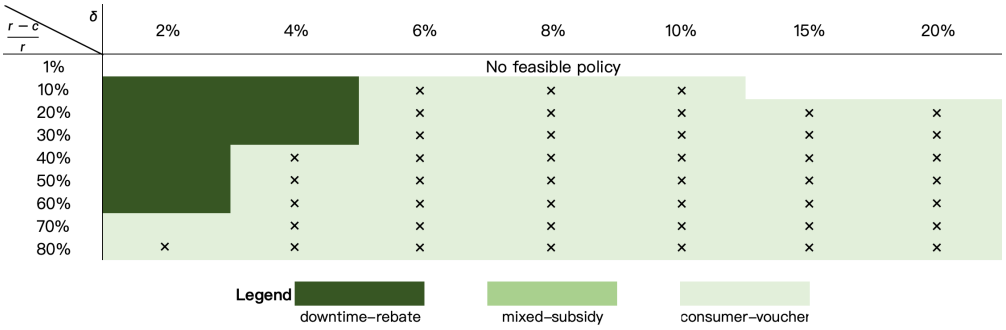


Figure C3: Comparison of the consumer-voucher policy and the downtime-rebate policy. The darker cells indicate that the downtime-rebate policy has a lower cost than the consumer-voucher policy for the corresponding δ and $\frac{r-c}{r}$. An $\times$ indicates that the downtime-rebate policy is infeasible.

Algorithm 1 Mixed-subsidy Policy (Consumer-voucher, Downtime-rebate)

Use a bisection search on $[1, N^2]$ to find ν_0 such that $\frac{r-c}{r} = \tilde{Q}_0(\nu_0) \cdot (1 + \mathbb{E}[W|\ \nu_0])$
Initialize $\nu^{\min} \leftarrow \nu_0$ and $\nu^{\max} \leftarrow N^2$ (or set $\nu^{\max}$ based on the maximum feasible service rate and $\lambda = f(r)$)
Initialize $\kappa = r$ and $\theta^{\text{best}} \leftarrow (\kappa, 0, g_\zeta(\nu^{\max}))$
if $\mathbb{E}[S|\ \mu_{\theta^{\text{best}}}, \lambda_{\theta^{\text{best}}}]/N < B$ **then**
    Return "Target is infeasible"
else
    Initialize TERMINATE = false
    while TERMINATE $\leftarrow$ false **do**
        Set $\underline{\nu} \leftarrow \nu^{\min}$ and $\overline{\nu} \leftarrow \nu^{\max}$
        Set $\underline{\theta} \leftarrow (\kappa, 0, g_\zeta(\underline{\nu}))$ and $\overline{\theta} \leftarrow (\kappa, 0, g_\zeta(\overline{\nu}))$
        if $\mathbb{E}[S|\ \mu_{\overline{\theta}}, \lambda_{\overline{\theta}}]/N < B$ **then**
            TERMINATE $\leftarrow$ true
        else if $\mathbb{E}[S|\ \mu_{\underline{\theta}}, \lambda_{\underline{\theta}}]/N \geq B$ **then**
            Set $\overline{\theta} = \underline{\theta}$
        else
            while $\left(1 + \frac{\epsilon}{4}\right) \cdot c_\zeta(\underline{\theta}) < c_\zeta(\overline{\theta})$ **do**
                Set $\nu^{\text{avg}} \leftarrow \frac{1}{2}(\overline{\nu} + \underline{\nu})$ and set $\theta^{\text{avg}} \leftarrow (\kappa, 0, g_\zeta(\nu^{\text{avg}}))$
                if $\mathbb{E}[S|\ \mu_{\theta^{\text{avg}}}, \lambda_{\theta^{\text{avg}}}]/N > B$ **then**
                    Set $\overline{\nu} \leftarrow \nu^{\text{avg}}$ and $\overline{\theta} \leftarrow \theta^{\text{avg}}$
                else
                    Set $\underline{\nu} \leftarrow \nu^{\text{avg}}$ and $\underline{\theta} \leftarrow \theta^{\text{avg}}$
                end if
            end while
            Update $\theta^{\text{best}} \leftarrow \arg\min\{C(\theta)|\ \theta \in \{\theta^{\text{best}}, \overline{\theta}\}\}$
            if $C(\overline{\theta}) > (1 + \epsilon) \cdot C(\theta^-)$ for $\theta^- = (0, 0, g_\zeta(\underline{\nu}))$ **then**
                Set $\kappa \leftarrow \kappa \cdot (1 + \epsilon)^{-1}$
            else
                Set TERMINATE $\leftarrow$ true
            end if
        end if
    end while
    Run bisection search for optimal single-subsidy $\theta_\kappa = (\kappa', 0, 0)$, $\theta_\zeta = (0, 0, \zeta')$, if feasible
    Return "Target is feasible" and policy $\theta^{\text{best}} \leftarrow \arg\min\{C(\theta)|\ \theta \in \{\theta^{\text{best}}, \theta_\kappa, \theta_\zeta\}\}$
end if

policy seems to have 'lost' to the fee-for-service policy, in the sense that it becomes totally infeasible. This reveals the weakness of the downtime-rebate policy and is magnified by convex costs. When the profit margin is low, the provider simply limits capacity to control costs, forcing the system into a high-congestion state where utilization is near 100% and the probability of idleness is small. In this regime, the downtime-rebate policy loses its leverage. Under convex costs that explode as the capacity increases, the provider further reduces her capacity. The policy becomes infeasible because it attempts to subsidize idleness that effectively does not exist. Conversely, the fee-for-service policy thrives here. It subsidizes throughput, which is high in a congested system, providing a direct incentive to overcome

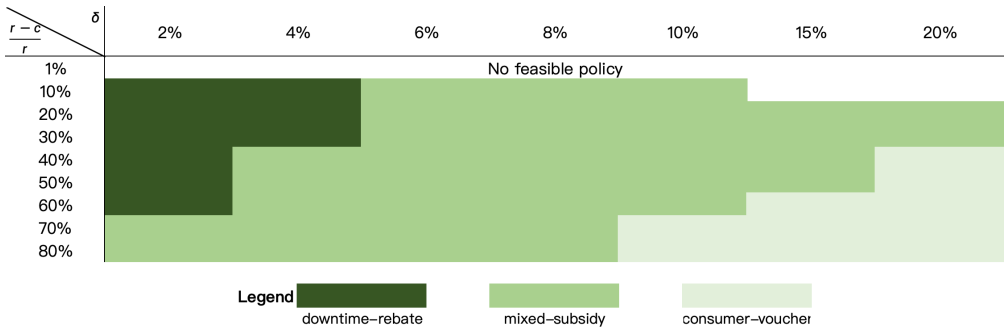


Figure C4: Comparison of the mixed-subsidy policy $\hat{\theta} = (\kappa, 0, \zeta)$ against the optimal single-subsidy policies distributing either κ or ζ.

the convex costs of capacity expansion.

However, despite this feasibility distinction, our numerical results confirm that the main insight regarding cost-effectiveness of the downtime-rebate policy over the fee-for-service policy holds. Wherever the downtime-rebate policy is feasible, that is, when the system is not in a state of extreme congestion, it strictly dominates the fee-for-service policy in terms of government cost. Furthermore, the mixed-subsidy policy continues to offer superior performance, particularly in scenarios with intermediate profit margins and targets, by balancing the strengths of both instruments.

C2 Welfare maximization

In our main analysis, we modeled the government's objective as minimizing fiscal costs subject to a target service level. In this appendix, we consider the government to maximize citizen welfare subject to a provider participation constraint.

C2.1 Model Setup for Welfare Maximization

We define the citizen welfare function as the total utility derived from the service minus the costs incurred by the citizens including waiting costs. The formulation of citizen welfare under policy is as follows:

$$CW = \beta \cdot \mathbb{E}[S] - h \cdot \mathbb{E}[W] - (r - \kappa) \cdot \mu \cdot (1 - \tilde{Q}_0).$$

Here, β represents the social benefit or utility parameter for each satisfied citizen. h denotes the unit cost of waiting time for citizens. $(r - \kappa) \cdot \mu \cdot (1 - \tilde{Q}_0)$ is the out-of-pocket payment of citizens.

The government solves the following optimization problem:

$$\begin{aligned} \max_{\theta} \quad & CW(\theta) && \text{(C2)} \\ \text{s.t.} \quad & (\kappa + \eta) \cdot \mu(1 - \tilde{Q}_0) + \zeta \cdot \mu \tilde{Q}_0 \leq \mathcal{B}, && \text{(Budget Constraint)} \\ & (r + \eta) \cdot \mu(1 - \tilde{Q}_0) + \zeta \cdot \mu \tilde{Q}_0 - c\mu \geq 0. && \text{(Provider Participation)} \end{aligned}$$

The provider participation constraint ensures that the policy does not force the service provider to operate at a loss.

C2.2 Numerical Experiments for Maximizing Social Welfare

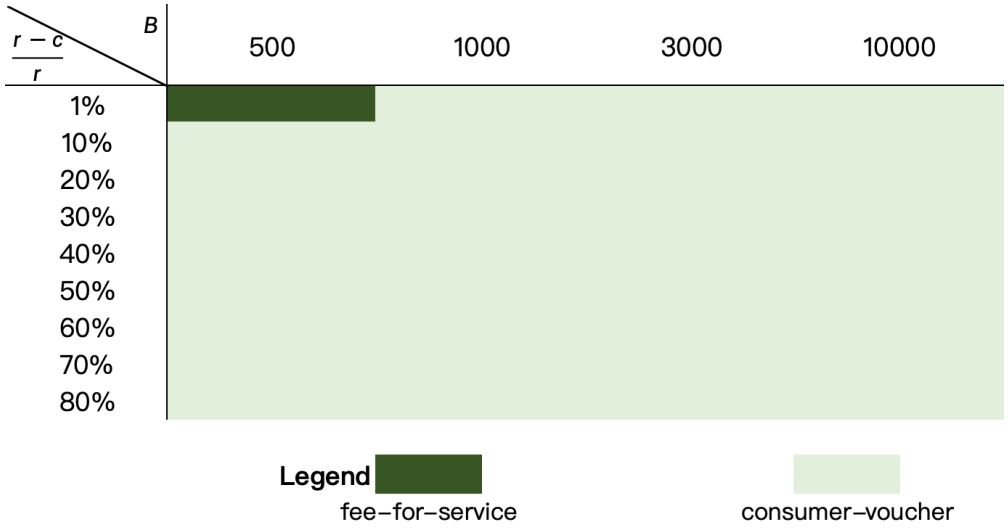


Figure C5: Comparison of the consumer-voucher policy and the fee-for-service policy when maximizing citizen welfare. The darker cells indicate that the fee-for-service policy has lower costs than the consumer-voucher policy for the corresponding δ and $\frac{r-c}{r}$. An $\times$ indicates that the fee-for-service policy is infeasible.

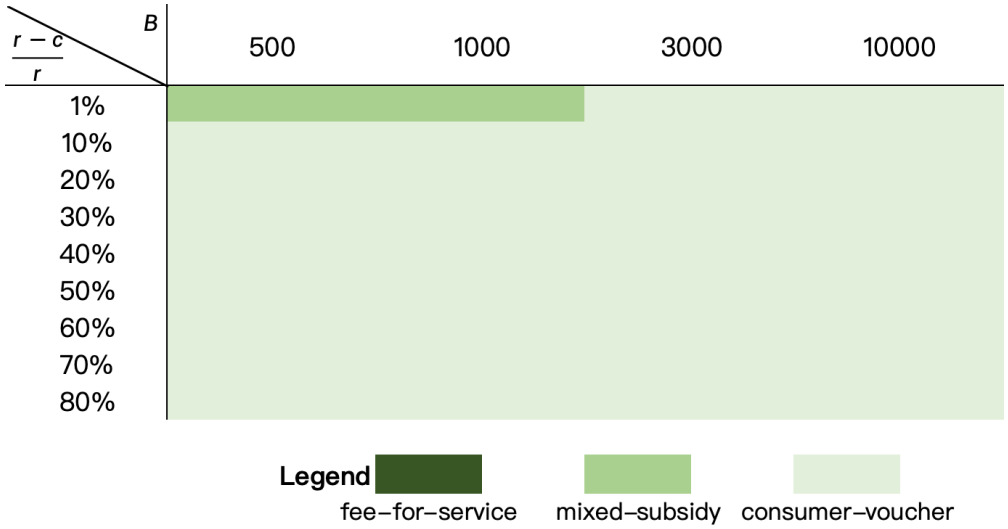


Figure C6: Comparison of the mixed policy $(\kappa, \eta, 0)$ against the optimal single-incentive policy distributing either κ or η, when maximizing citizen welfare.

The results in Figures C5-C8 of the welfare maximization model align closely with our cost-minimization findings. The optimal policy choice remains dependent on the system's

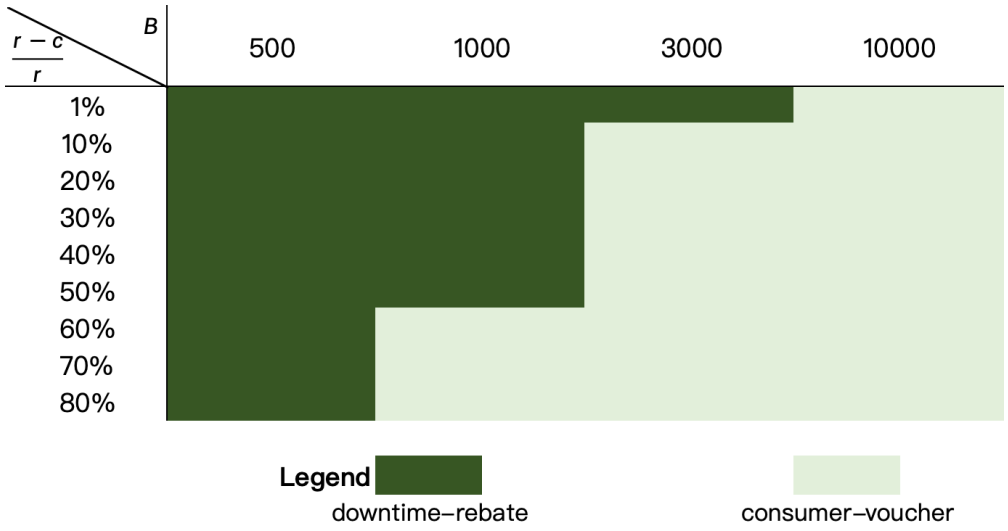


Figure C7: Comparison of the consumer-voucher policy and the downtime-rebate policy when maximizing citizen welfare. The darker cells indicate that the downtime-rebate policy has lower costs than the consumer-voucher policy for the corresponding δ and $\frac{r-c}{r}$. An $\times$ indicates that the downtime-rebate policy is infeasible.

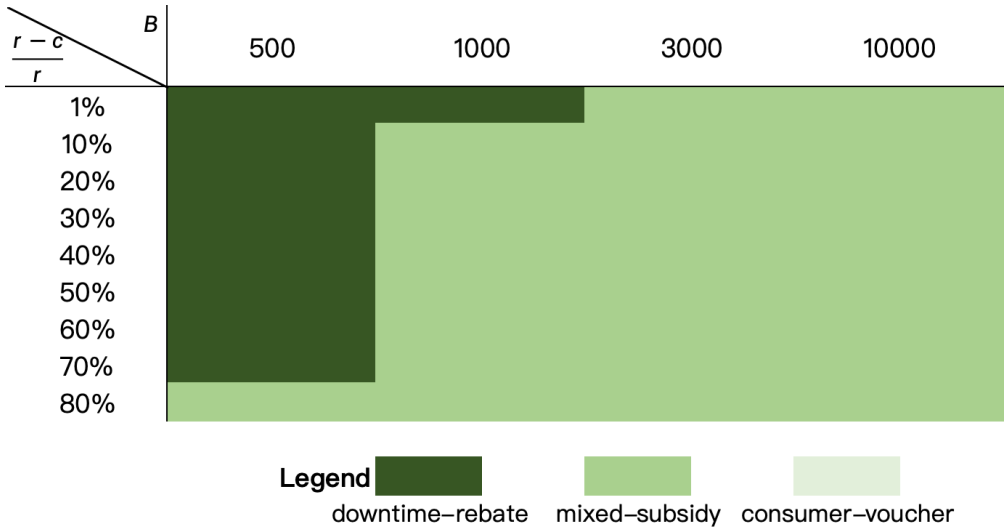


Figure C8: Comparison of the mixed policy $(\kappa, 0, \zeta)$ against the optimal single-incentive policy distributing either κ or ζ, when maximizing citizen welfare.

primary bottleneck. When the provider operates efficiently with high margins, waiting costs are naturally low. The welfare gain comes primarily from reducing the financial burden on citizens. Thus, the consumer-voucher policy yields the highest welfare. While the provider has low margins, her capacity is constrained, and waiting costs become the dominant negative factor in welfare. Here, provider-based subsidies are superior because they effectively incentivize capacity expansion and reduce waiting costs, which outweighs the benefit of cheaper prices. Under tight budget constraints or intermediate profit margins, the mixed policy frequently outperforms single policies. By allocating a portion of the budget to rebates to control waiting costs and ensure provider viability, and the remainder to vouchers to improve affordability, the government can achieve a higher citizen welfare than by focusing on a single subsidy.

C3 Distributional Assumption of System Parameters

We build the system based on exponential distributions for tractability, which effectively allow us to tackle the problem with Erlang-loss systems. To verify the robustness of our insights, we develop a comprehensive Discrete Event Simulation (DES) model that relaxes the exponential assumptions. We conduct extensive experiments comparing our baseline model against three alternative distributional scenarios, ensuring that the mean rates (λ, ϕ) remain consistent across all cases to allow for a fair comparison.

We tested the following distributions to capture different realistic features of service systems:

- Deterioration Process: Weibull distribution. Instead of the memoryless Exponential distribution for the deterioration process, we implemented a Weibull distribution with a shape parameter $k = 1.5$. This models an increasing hazard rate, reflecting the reality that the probability of needing care increases as time passes since the last treatment, e.g., the aging or wear-out effect.

- Service Process: Log-Normal distribution (high variability). To model human-delivered services which often have a long tail, e.g., occasional complex cases taking much longer, we used a Log-Normal distribution with a coefficient of variation of 0.5.

- Service Process: Deterministic distribution (zero variability). As a benchmark for maximum efficiency, we tested fixed service times, representing highly standardized, machine-driven services.

For each subsidy policy, we simulated the system dynamics over a range of subsidy values. To ensure the results represent the optimal response of the system, we implemented a simulation-based optimization routine. For every subsidy level and distribution type, the provider's capacity μ was optimized to maximize her profit, and the resulting system metrics, $\mathbb{E}[S], \mathbb{E}[W]$ and $\mathbb{E}[D]$, were recorded after a long-run steady state was reached. The results of our robustness check are visualized in Figures C9-C11. We observe a consistency of trends across all distributions. The impact of the three policies on the key metrics follows a similar qualitative trajectory.

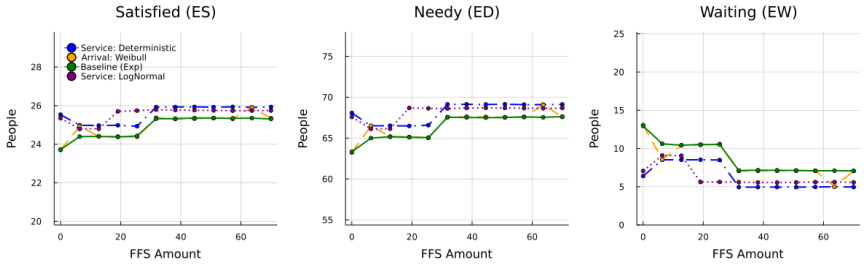


Figure C9: Performance of the fee-for-service policy under different distributions.

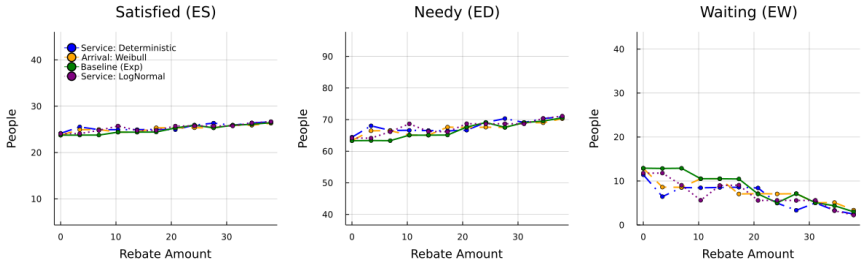


Figure C10: Performance of the downtime-rebate policy under different distributions.

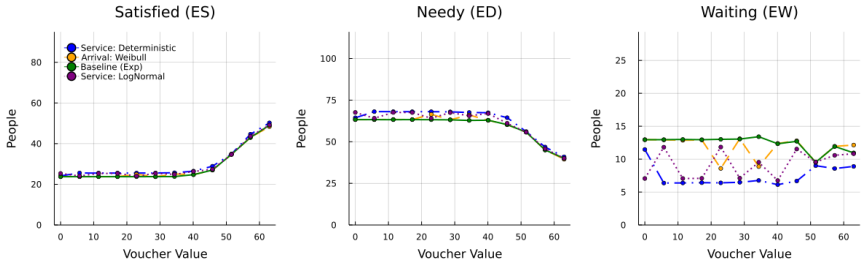


Figure C11: Performance of the consumer-voucher policy under different distributions.