# Joint Design of Treatment Allocation and Recommendation*

Li Guo† Penghuan Yan‡

May 29, 2026

**Abstract**

We study a model in which a sender allocates limited treatment that improves agents' quality and later recommends selected agents to a receiver, seeking to maximize the number of agents accepted by the receiver. All agents value treatment, but treatment must be allocated before the sender observes agents' quality; recommendation occurs after quality is learned. A natural idea is to design the two instruments separately: allocate treatment randomly first, and then recommend agents from the top down afterward. We show that this can be strictly improved by jointly designing treatment allocation and recommendation. In the optimal joint mechanism, treatment is non-monotone in quality: an intermediate group has a lower treatment probability than both higher- and lower-quality agents, but is compensated with a guaranteed recommendation when treatment is realized. We provide an implementation through contracts that induce self-selection and discuss applications to education, industrial policy, startup incubation, and bank runs. The takeaway is simple: jointly design treatment allocation and recommendation.

*This paper has previously been circulated under the title "Multi-Section Grading Design". We thank Cuimin Ba, J. Aislinn Bohren, Kevin He, Navin Kartik, Xiao Lin, George J. Mailath, Margaret Meyer, Stephen Morris, Alessandro Pavan, Andrew Postlewaite, Daniel Rappoport, Doron Ravid, Collin Raymond, Juuso Toikka, Maren Vairo, and Rakesh V. Vohra, as well as participants at the Penn Micro Theory Lunch, the Midwest Trade & Theory Conference, the Pennsylvania Economics Theory Conference (PETCO), the ACM Conference on Economics and Computation (EC'26), and the North American Summer Meeting of the Econometric Society for valuable comments and suggestions.

†University of Pennsylvania. gary16@sas.upenn.edu

‡University of Pennsylvania. yanph@sas.upenn.edu

# 1 Introduction

Many institutions both expend resources to improve agents' quality and recommend selected agents to outside evaluators, with the objective of maximizing acceptance by those evaluators. Schools provide students with educational resources and later recommend them to employers for hiring. Incubators support startups and later recommend selected startups to investors who decide whether to invest in them. We refer to such an institution as the sender. In these settings, the sender must make two choices: how to initially allocate limited treatment that improves the treated agents' quality, and how to later recommend agents to an outside receiver.

Treatment is limited and must be allocated before the sender knows agents' quality. All agents value treatment, but the sender cannot yet tell who should receive it. A school may allocate educational resources before knowing which students will become strong candidates. An incubator may allocate support before knowing which startups have real potential. Only after treatment is allocated does the sender learn more: schools observe course performance, and incubators observe demand data. The sender can then recommend selected agents to the outside receiver.

In this paper, we focus on how a sender should allocate limited treatment and design recommendations to maximize the number of agents accepted by the receiver.

To answer this question, we study a joint mechanism and information design problem that involves a sender, a receiver, and a continuum of agents. The sender wants to maximize the mass of agents accepted by the receiver. The sender has two instruments. First, before observing agents' true types, the sender can allocate treatment to a limited mass of agents. Treatment improves treated agents' quality. Second, after treatment is realized and agents' types are observed, the sender can decide whether to recommend each agent to the receiver. The receiver observes only the sender's recommendation and accepts recommended agents only if their expected quality is sufficiently high. Agents value both treatment and acceptance, but value acceptance more.

A natural idea is to design the two instruments separately. The sender first allocates treatment, and later, after observing agents' quality, decides whom to recommend. Since the sender cannot yet identify quality and all agents want treatment, treatment is assigned at random and each agent receives treatment with the same probability. After treatment is realized and quality is observed, the sender recommends agents in descending order of post-treatment quality. The recommendation cutoff is chosen so that the average quality of recommended agents is just high enough for the receiver to accept them.

Our main result shows that the sender can do strictly better by designing treatment

allocation and recommendation jointly. The optimal joint mechanism targets a group of intermediate-quality agents. These agents receive a lower probability of treatment than others, but are compensated with a recommendation as long as they are treated. In this way, the sender reduces the amount of treatment spent on this group while preserving their incentives to report truthfully. As a result, high-quality agents, who are more likely to be recommended, receive treatment with higher probability, thereby improving the overall quality of recommended agents.

The treatment allocation in the optimal joint mechanism is non-monotone. Treatment is useful to the sender only when it reaches agents who are eventually recommended. The sender therefore wants to reserve treatment for high-quality agents, who are more likely to enter the recommended pool and make it easier to satisfy the receiver's acceptance constraint. The mechanism instead offers a special arrangement to an intermediate group: these agents receive a lower treatment probability and are compensated with treatment-contingent recommendation. This saves capacity and redirects treatment toward higher-quality agents. At the same time, the sender does not want the lowest-quality agents to choose this arrangement, because recommending them would lower the average quality of the recommended pool too much. To separate them from the intermediate group, the mechanism gives the lowest-quality agents a higher treatment probability than the intermediate group, while giving them no chance of recommendation.

In standard information design, a recommendation is mainly a signal sent to the receiver. Here, recommendation also serves as an incentive device for agents because the sender can commit to the recommendation rule before treatment is allocated. The same instrument therefore performs two tasks: it persuades the receiver after quality is learned, and it screens agents before treatment is assigned.

We provide a simple implementation of the optimal joint mechanism using a menu with two contracts that induces self-selection. Accepting the non-default contract means receiving a lower probability of treatment, but facing a more lenient standard for recommendation when treated. In equilibrium, only the targeted group accepts the non-default contract; all other agents choose the default option and face the regular treatment and recommendation rule. We also discuss applications to education, industrial policy, startup incubation, and bank runs.

### Related Literature

Our paper builds on Bayesian persuasion and information design. In this literature, a sender designs signals to influence a receiver's beliefs and actions (Rayo and Segal, 2010; Kamenica and Gentzkow, 2011; Bergemann and Morris, 2016; Kolotilin, 2018; Bergemann and Morris,

2019; Kamenica, 2019). Our paper has the same persuasion motive: the sender uses information to induce acceptance by a receiver. The difference is that the payoff-relevant state is endogenous: the quality distribution entering the recommendation problem depends on the sender's treatment-allocation rule.

The treatment-allocation part of the model relates to mechanism design and screening. In this literature, a designer allocates goods or contracts among privately informed agents subject to incentive constraints (Mussa and Rosen, 1978; Myerson, 1981; Baron and Myerson, 1982; Maskin and Riley, 1984; Laffont and Martimort, 2002). Our model is especially related to mechanism design without monetary transfers (Martimort and Semenov, 2006; Schummer and Vohra, 2007; Miralles, 2012; Ben-Porath, Dekel and Lipman, 2014). In our model, the only instrument through which the sender can transfer utility to agents is the later recommendation. At the same time, the sender's objective is to use this recommendation to induce receiver acceptance. This dual role of recommendation is the key difference: the same recommendation rule must provide incentives to agents before treatment is allocated and remain credible to the receiver afterward.

The closest strand of literature studies the joint design of information and mechanisms (Bergemann and Pesendorfer, 2007; Kolotilin, Mylovanov, Zapechelnyuk and Li, 2017; Bergemann, Heumann and Morris, 2026). These papers share with ours the idea that information and mechanisms should be designed jointly. The difference is that their information design and mechanism design are directed at the same side of the market. Our environment is three-sided. The sender designs a treatment-allocation mechanism for agents, but designs information for a separate receiver. The recommendation connects the two sides: it is the sender's only instrument for transferring utility to agents, and it is also the information device used to induce receiver acceptance.

Within this literature, the closest paper is Dworczak (2020). He studies a setting in which a designer first runs a mechanism and then chooses whether and how to disclose information elicited by that mechanism to a third party. In our model, treatment allocation is not itself the sender's payoff-relevant objective and does not serve to elicit information; its role is to change agents' quality so that later recommendation can satisfy the receiver's requirement.

Our paper is also related to work on certification, ratings, grading, testing, and information as an incentive device. Classic work on certification and quality disclosure studies how intermediaries or firms disclose quality information to markets (Lizzeri, 1999; Dranove and Jin, 2010). More recent work studies how certification, ratings, grades, or tests can induce agents to take costly actions that improve quality, such as effort, investment, or input choices (Zubrickas, 2015; Boleslavsky and Kim, 2018; Zapechelnyuk, 2020; Xiao, 2024; Camboni and Carnehl, 2025; Saeedi and Shourideh, 2026). In these papers, the analogue

of treatment is typically costly for agents: it lowers their utility but raises the quality later revealed or rewarded by certification, ratings, grades, or tests. We study a complementary case in which treatment also raises quality but is valued by agents. The sender therefore does not use recommendation to induce agents to take a costly action; it uses recommendation to screen agents and improve the allocation of a scarce treatment.

The rest of the paper is organized as follows. Section 2 sets up the model. Section 3 characterizes the optimal joint mechanism, develops the main intuition, and provides an implementation of the mechanism via contracts that induce self-selection. Section 4 discusses four applications: education, industrial policy, startup incubation, and bank runs. Section 5 concludes.

# 2 Model

## 2.1 Environment

We study a joint mechanism and information design problem in which a sender seeks to maximize the mass of agents accepted by a receiver. The sender has two instruments. First, before agents' types are observed by the sender, the sender can allocate treatment to a limited mass of agents, which improves treated agents' quality. Second, after treatment is realized and agents' types are observed, the sender can decide whether to recommend each agent to the receiver.

Formally, there is a unit mass of agents. Each agent has a type $\theta \in \Theta = [0, 1]$, drawn from a distribution $F$ with positive density everywhere. The type $\theta$ represents the agent's initial quality.

Treatment raises an agent's quality. If a type-$\theta$ agent receives treatment, his quality becomes $\theta + v(\theta)$; otherwise, his quality remains $\theta$. Let $d \in \{0, 1\}$ denote realized treatment status, where $d = 1$ means that the agent is treated. Treatment is limited: the sender can treat at most a mass $\kappa \in (0, 1)$ of agents.

After treatment decisions are realized, the sender observes each agent's resulting quality. She then sends messages about the agents' quality to the receiver. For each agent the sender chooses a message $m \in \mathcal{M}$, where $\mathcal{M}$ denotes the message space. The sender's messaging strategy may depend on the agent's realized quality and treatment status.

The receiver decides whether to accept an agent. He wants to accept only agents whose expected quality is sufficiently high. Specifically, after observing the sender's message, the receiver accepts an agent only if the posterior expected quality of that agent is at least $\underline{\theta}$. The sender's payoff is the mass of agents accepted by the receiver.

Agents value both treatment and acceptance. The payoff from being accepted is normalized to 1. A type-$\theta$ agent obtains utility $u(\theta)$ from receiving treatment, where $0 < u(\theta) < 1$. Thus, acceptance is more valuable than treatment for every type, but all agents strictly prefer receiving treatment to not receiving it.

The information structure is as follows. Agents privately know their types when treatment is allocated. The sender observes agents' true types only after treatment has been assigned and realized. The receiver observes neither agents' types nor their treatment statuses; he observes only the sender's message.

The timing is as follows. First, the sender publicly commits to a mechanism that specifies how treatment allocation and recommendation depend on subsequent information. Second, each agent observes his type and communicates with the sender. Third, treatment is assigned according to the mechanism, and treatment status is realized. Fourth, the sender observes the agent's true type and treatment status, and sends a recommendation message to the receiver according to the mechanism. Finally, the receiver observes the sender's message and decides whether to accept the agent.

### 2.2 Direct Mechanisms

By the revelation principle, we restrict attention to direct mechanisms. In a direct mechanism, each agent reports a type $\hat{\theta} \in \Theta$, and truth-telling requires $\hat{\theta} = \theta$ on the equilibrium path.

A direct mechanism is denoted by $\pi = (p, q)$. The first component, $p : \Theta \to [0, 1]$, is a treatment allocation rule. If an agent reports $\hat{\theta}$, he receives treatment with probability $p(\hat{\theta})$. The second component, $q$, is a recommendation rule. A recommendation rule is a mapping

$$q : \Theta \times \Theta \times \{0, 1\} \to \Delta(\mathcal{M}),$$

where $q(m \mid \hat{\theta}, \theta, d)$ is the probability that the sender sends message $m \in \mathcal{M}$ after observing report $\hat{\theta}$, true type $\theta$, and realized treatment status $d$.

Since the receiver's action is binary, we restrict attention without loss to a binary message space, $\mathcal{M} = \{0, 1\}$. We interpret $m = 1$ as a recommendation and $m = 0$ as no recommendation. With binary messages, we write

$$q(\hat{\theta}, \theta, d) := q(1 \mid \hat{\theta}, \theta, d)$$

for the probability of recommendation. A recommendation is intended to induce acceptance by the receiver, subject to the obedience constraint below.

### 2.3 Receiver Obedience and Agent Incentives

Under truth-telling $\hat{\theta} = \theta$, the posterior mean quality of a recommended agent is

$$\mu_\pi := \frac{\displaystyle\int_\Theta \overbrace{[p(\theta)q(\theta,\theta,1)\big(\theta + v(\theta)\big)}^{\text{treated quality}} + \overbrace{\big(1-p(\theta)\big)q(\theta,\theta,0)\ \theta}^{\text{untreated quality}}]\mathrm{d}F(\theta)}{\underbrace{\displaystyle\int_\Theta \big[p(\theta)q(\theta,\theta,1) + \big(1-p(\theta)\big)q(\theta,\theta,0)\big]\,\mathrm{d}F(\theta)}_{\text{mass of recommended agents}}},$$

whenever the denominator is positive. This expression integrates out treatment status: the receiver does not observe whether a recommended agent was treated, but infers the composition of the recommended pool from the sender's committed rules.

Receiver obedience requires

$$\mu_\pi \geq \underline{\theta}.$$

Given the agents' preferences, if a type-$\theta$ agent reports $\hat{\theta}$, his expected payoff is

$$U_\pi(\theta,\hat{\theta}) = \underbrace{p(\hat{\theta})\Big[u(\theta) + q(\hat{\theta},\theta,1)\Big]}_{\text{payoff when treated}} + \underbrace{\big(1-p(\hat{\theta})\big)q(\hat{\theta},\theta,0)}_{\text{payoff when untreated}}.$$

With probability $p(\hat{\theta})$, the agent receives treatment and obtains treatment utility $u(\theta)$; conditional on being treated, he is recommended with probability $q(\hat{\theta},\theta,1)$. With probability $1-p(\hat{\theta})$, the agent is untreated and obtains only the acceptance payoff, which occurs with probability $q(\hat{\theta},\theta,0)$.

Truthful reporting requires

$$U_\pi(\theta,\theta) \geq U_\pi(\theta,\hat{\theta}), \quad \forall \theta,\hat{\theta} \in \Theta.$$

### 2.4 Sender's problem

The sender chooses a feasible direct mechanism to maximize the mass of recommended agents. Under truth-telling, the sender solves

$$\begin{aligned}
\max_{p,q} \quad & \int_\Theta \big[p(\theta)q(\theta,\theta,1) + \big(1-p(\theta)\big)q(\theta,\theta,0)\big]\,\mathrm{d}F(\theta), & \\
\text{s.t.} \quad & \int_\Theta p(\theta)\,\mathrm{d}F(\theta) \leq \kappa, & \text{(Capacity)} \\
& \mu_\pi \geq \underline{\theta}, & \text{(Obedience)} \\
& U_\pi(\theta,\theta) \geq U_\pi(\theta,\hat{\theta}), \quad \forall \theta,\hat{\theta} \in \Theta. & \text{(IC)}
\end{aligned} \tag{2.1}$$

The model combines a mechanism-design problem on the agent side with an information-design problem on the receiver side. Treatment is allocated before the sender observes agents' types, so the allocation rule must respect agents' reporting incentives. Recommendation occurs later, after the sender observes true types and realized treatment statuses, so the recommendation rule determines the receiver's posterior belief.

## 3 Optimal Direct Mechanism

In this section, we characterize the optimal direct mechanism. To keep the analysis transparent and to highlight the main intuition, we impose throughout this section the simplifying assumption that

$$u(\theta) \equiv u \quad \text{and} \quad v(\theta) \equiv v,$$

where $0 < u < 1$ and $v > 0$. The same economic forces remain present with more general functions $u(\theta)$ and $v(\theta)$. We focus on the nondegenerate interior case in which the receiver obedience constraint binds, i.e.

$$\underline{\theta} > \int_{\Theta} \theta \, \mathrm{d}F(\theta) + v\kappa.$$

Our main result is that the optimal direct mechanism takes a four-segment form.

**Theorem 1.** Suppose $u(\theta) \equiv u$ and $v(\theta) \equiv v$, with $0 < u < 1$ and $v > 0$. In the nondegenerate interior case, an optimal direct mechanism either satisfies $q(\hat{\theta}, \theta, d) = 0$ almost everywhere (recommends no one), or takes the following form:

$$\big(p(\theta), q(\theta,\theta,1), q(\theta,\theta,0)\big) = \begin{cases} (P, 1, 1), & \theta \in (\theta_3, 1], \\ (P, 1, 0), & \theta \in (\theta_2, \theta_3], \\ \left(\dfrac{u}{1+u}P, 1, 0\right), & \theta \in (\theta_1, \theta_2], \\ (P, 0, 0), & \theta \in [0, \theta_1], \end{cases}$$

where $\kappa < P \leq 1$ and $0 \leq \theta_1 < \theta_2 \leq \theta_3 \leq 1$ are cutoffs, and

$$q(\hat{\theta}, \theta, d) = 0, \quad \text{for all } \hat{\theta} \neq \theta \text{ and } d \in \{0, 1\}.$$

The mechanism divides agents into four groups. The top group receives the high treatment probability $P$ and is recommended regardless of treatment status. The upper-middle

group also receives the high treatment probability $P$ and is recommended only if treated. The lower-middle group is the target group: these agents receive a lower treatment probability, $\frac{u}{1+u}P$, and are recommended conditional on being treated. The bottom group receives the high treatment probability $P$ but is never recommended.

To understand the intuition behind Theorem 1, it is useful to compare the optimal mechanism with a separated-design benchmark.

### 3.1 Separated-Design Benchmark

In this benchmark, treatment allocation and recommendation are designed separately. At the treatment-allocation stage, the sender only chooses which agents receive treatment. After treatment is realized and types are observed, the sender publicly commits to a recommendation rule. The solution to this recommendation problem is characterized by a cutoff $\theta^S$, and the benchmark mechanism is

$$\big(p^S(\theta), q^S(\theta,\theta,1), q^S(\theta,\theta,0)\big) = \begin{cases} \big(P^S,1,1\big), & \theta \in (\theta^S,1], \\ \big(P^S,1,0\big), & \theta \in (\theta^S - v, \theta^S], \\ \big(P^S,0,0\big), & \theta \in [0,\theta^S - v], \end{cases} \qquad P^S = \kappa. \tag{S}$$

The cutoff $\theta^S$ is chosen so that the receiver's obedience constraint binds.

The cutoff form has a simple interpretation. Since treatment allocation is designed separately from recommendation and all agents value treatment, before types are observed, the sender has no way to direct treatment toward particular types. Treatment is therefore allocated uniformly: $P^S = \kappa$.

Once treatment is realized and types are observed, the sender faces a standard recommendation problem. Since treatment raises quality by $v$, a treated agent of type $\theta$ has quality $\theta + v$, while an untreated agent has quality $\theta$. The sender therefore recommends agents from the top down in realized quality. Untreated agents are recommended only if $\theta > \theta^S$, while treated agents are recommended if $\theta + v > \theta^S$, or equivalently $\theta > \theta^S - v$. Thus high types are recommended regardless of treatment status, intermediate types are recommended only if treated, and low types are never recommended.

The limitation of the separated benchmark is that treatment cannot be targeted. Treatment helps the sender only when it is assigned to agents who are eventually recommended. If a treated agent is not recommended, the treatment uses capacity but does not improve the quality of the recommended pool. However, in the separated benchmark, treatment is forced to be assigned randomly, so some treatment is spent on agents who are too low-quality to

be recommended even after treatment.

The sender would instead like treatment to reach agents with high initial quality, since these agents are more likely to enter the recommended pool. Treating such agents raises the quality of the recommended pool and relaxes the receiver's obedience constraint, which allows the sender to recommend more agents. The problem is that, at the treatment-allocation stage, the sender cannot distinguish these agents from lower-quality agents. The optimal joint mechanism in Theorem 1 improves on the separated benchmark by using the future recommendation rule to create incentives that help direct treatment toward the relevant types.

### 3.2 Intuition for Theorem 1

The optimal joint mechanism differs from the separated benchmark by introducing a target group. In the separated benchmark, low-type agents are never recommended, regardless of treatment status, but they still receive the same treatment probability as all other agents. The optimal joint mechanism instead selects some relatively low types and gives them a lower treatment probability, while compensating them with a recommendation when treatment is realized. This is the lower-middle group,

$$\theta \in (\theta_1, \theta_2], \qquad \big(p(\theta), q(\theta,\theta,1), q(\theta,\theta,0)\big) = \left(\frac{u}{1+u}P, 1, 0\right).$$

These agents receive a lower treatment probability than the other groups. They are not recommended when untreated, and recommended when treated.

The purpose of this design is to save treatment capacity. The target group consists of relatively low types, so the sender would rather reserve treatment for higher types. However, if the target group simply received a lower treatment probability, these agents would prefer to mimic types who receive treatment with probability $P$.

The mechanism prevents this deviation by attaching recommendation to treatment. Conditional on receiving treatment, a target agent obtains both the treatment payoff $u$ and the acceptance payoff 1. Hence his truthful payoff is

$$\frac{u}{1+u}P(1+u) = Pu,$$

which is exactly the payoff from mimicking a type who receives treatment with probability $P$ but is not recommended. Thus the target group is willing to report truthfully, while the sender reduces the treatment probability spent on them from $P$ to $\frac{u}{1+u}P$.

This compensation is not free. Recommending lower-quality agents lowers the average

quality of the recommended pool and tightens the receiver's obedience constraint. Hence the sender faces a trade-off. Reducing treatment probability saves scarce capacity, but compensating agents through recommendation makes the obedience constraint harder to satisfy.

This explains the non-monotone structure of the mechanism. The target group is neither the highest group nor the lowest group. It is not optimal to reduce treatment probability for the highest types because the sender wants these agents to receive treatment: they are most valuable for the receiver's obedience constraint. It is also not optimal to target the lowest types because compensating them through recommendation is too costly. Their quality is too low, so recommending them would sharply reduce the posterior mean quality of recommended agents. For the lowest types, the mechanism instead gives the high treatment probability $P$ but no recommendation. The target group is therefore an intermediate group.

The preceding discussion shows why treatment allocation and recommendation must be designed jointly. The key limitation of the separated design is that it uses the sender's commitment power only toward the receiver, not toward agents. Joint design also uses the sender's commitment power toward agents. By committing ex-ante to treatment-contingent recommendation rules, the sender can use future recommendation to screen agents before treatment is allocated. This allows the sender to redirect treatment capacity toward more valuable types. Corollary 1.1 shows that this joint design strictly improves on the separated-design benchmark.

**Corollary 1.1.** Suppose $u(\theta) \equiv u$ and $v(\theta) \equiv v$, with $0 < u < 1$ and $v > 0$. In the nondegenerate interior case, the optimal joint mechanism characterized in Theorem 1 achieves a strictly larger mass of accepted agents than the separated-design benchmark in (S).

Corollary 1.1 follows directly from the restriction in Theorem 1 that $P$ be strictly greater than $\kappa$.

### 3.3 Implementation

The optimal direct mechanism asks agents to report their types. In practice, such direct reports may be unrealistic. We now describe an indirect implementation of the same allocation. Agents do not report their types directly. Instead, they choose between two contracts offered by the sender: a default contract and a non-default contract. The mechanism is described below.

**Indirect implementation**

The sender publicly commits to the following mechanism. The sender offers each agent a menu with two contracts: a default contract and a non-default contract. Under the non-default contract, the agent receives treatment with fixed probability $\frac{u}{1+u}P$. The recommendation rule for agents who choose the non-default contract is

$$\big(q^N(\theta,1), q^N(\theta,0)\big) = \begin{cases} (1,1), & \theta \in (\theta_3, 1], \\ (1,0), & \theta \in (\theta_1, \theta_3], \\ (0,0), & \theta \in [0, \theta_1]. \end{cases}$$

Thus, for agents who choose the non-default contract, the recommendation threshold is $\theta_1$ when treated and $\theta_3$ when untreated.

Under the default contract, the agent receives treatment with fixed probability $P$. Agents who choose the default contract face the regular recommendation rule:

$$\big(q^D(\theta,1), q^D(\theta,0)\big) = \begin{cases} (1,1), & \theta \in (\theta_3, 1], \\ (1,0), & \theta \in (\theta_2, \theta_3], \\ (0,0), & \theta \in [0, \theta_2]. \end{cases}$$

Thus, agents who choose the default contract face a higher recommendation threshold when treated: $\theta_2 > \theta_1$. The threshold when untreated remains $\theta_3$.

We show that this menu implements the optimal direct mechanism under tie-breaking in favor of the non-default contract. In the candidate equilibrium, agents in the target group, $\theta \in (\theta_1, \theta_2]$, choose the non-default contract, while all other agents choose the default contract. We verify that no type has a profitable deviation:

Top-group agents, $\theta \in (\theta_3, 1]$, choose the default contract. Under the default contract, they receive treatment with probability $P$ and are recommended regardless of treatment status. Under the non-default contract, they are still recommended regardless of treatment status, but receive treatment with the lower probability $\frac{u}{1+u}P$. Hence they strictly prefer the default contract.

Upper-middle agents, $\theta \in (\theta_2, \theta_3]$, also choose the default contract. Under the default contract, they receive treatment with probability $P$ and are recommended whenever treated. Under the non-default contract, they are also recommended whenever treated, but receive treatment with the lower probability $\frac{u}{1+u}P$. Hence they strictly prefer the default contract.

Bottom agents, $\theta \in [0, \theta_1]$, choose the default contract as well. They are never recommended under either contract. Choosing the non-default contract only lowers their treatment probability, so they strictly prefer the default contract.

Finally, consider target agents, $\theta \in (\theta_1, \theta_2]$. If a target agent chooses the default contract, he receives treatment with probability $P$ but is never recommended, so his payoff is $Pu$. If he chooses the non-default contract, he receives treatment with probability $\frac{u}{1+u}P$, and treatment is bundled with recommendation. His payoff is

$$\frac{u}{1+u}P(1+u) = Pu.$$

Thus target agents are indifferent between the two contracts. With tie-breaking in favor of the non-default contract, all target agents choose it.

The induced allocation coincides with the optimal direct mechanism. Target agents choose the non-default contract, receive treatment with probability $\frac{u}{1+u}P$, and are recommended when treated. All other agents choose the default contract, receive treatment with probability $P$, and face the regular recommendation rule. Hence top agents are recommended regardless of treatment status, upper-middle agents are recommended only when treated, and bottom agents are never recommended.

This menu induces self-selection. Choosing the non-default contract means giving up treatment probability in exchange for a lower recommendation threshold when treated. Only the target group finds this trade-off weakly attractive, so the sender identifies them through their contract choice.

### 3.4 A Proof Sketch of Theorem 1

In this subsection, we present a proof sketch for our main result, Theorem 1. A complete proof is provided in Appendix A.

The proof starts with the following two observations. First, the probability of recommendation is the only instrument that the sender can use to punish agents who are found to have misreported their types. Since the sender has full commitment power, it is optimal to commit to the harshest punishment for misreports by never recommending agents whose reported type is inconsistent with their true type, i.e., $q(\hat{\theta}, \theta, d) = 0$ for all $\hat{\theta} \neq \theta$, $d \in \{0, 1\}$. As a result, the only potential benefit for an agent from deviating from truthful reporting is a higher probability of treatment. The treatment allocation rule $p$ depends only on the reported type. Therefore, if an agent is to deviate, it is optimal for him to report a type that receives the highest probability of treatment. We denote this highest probability by $P$, where $P = \sup_{\theta \in \Theta} p(\theta)$.

The second observation is that, apart from treatment, agents care only about the ex-ante probability of being recommended. That is, conditional on this probability, it does not matter whether recommendation occurs when they are treated or untreated. We denote the ex-ante recommendation rule by $Q$, where $Q(\theta)$ is the ex-ante probability of a type-$\theta$ agent being recommended given truthful reporting, i.e., $Q(\theta) = p(\theta)q(\theta,\theta,1) + [1-p(\theta)]q(\theta,\theta,0)$. It is sufficient for the sender to design $(p,Q)$ as the incentive scheme for the agents. Moreover, the same agent has higher quality when treated than when untreated. Therefore, for any fixed ex-ante recommendation probability $Q(\theta)$, the sender prefers to assign recommendation probability to the treated state first. Only after the agent is recommended with probability one when treated does the sender assign positive recommendation probability to the untreated state.

With these two observations, it can be shown that the sender's problem (2.1) is equivalent to the following problem:

$$\begin{aligned}
\max_{p,Q,P} \quad & \int_\Theta Q(\theta)\mathrm{d}F(\theta), \\
\text{s.t.} \quad & \int_\Theta p(\theta)\,\mathrm{d}F(\theta) \le \kappa, && \text{(Capacity)} \\
& \int_\Theta [(\theta-\underline{\theta})Q(\theta) + v\min\{p(\theta),Q(\theta)\}]\mathrm{d}F(\theta) \ge 0, && \text{(Obedience)} \\
& Q(\theta) + p(\theta)u \ge Pu, \quad \forall\theta\in\Theta, && (\mathrm{IC}_1) \\
& p(\theta) \le P, \quad \forall\theta\in\Theta, && (\mathrm{IC}_2)
\end{aligned}$$

The optimal recommendation rule $q$ can be recovered from $p$ and $Q$.

To solve this problem, we first fix the value of $P$ and optimize over $p$ and $Q$. When $P$ is fixed, the problem is a convex optimization problem satisfying Slater's condition. We therefore proceed with Lagrangian methods and show that given $P$, the solution takes either of the following two forms:

$$(p(\theta),Q(\theta)) = \begin{cases} (P,1), & \theta_3<\theta\le 1 \\ (P,P), & \theta_2<\theta\le\theta_3 \\ \left(\frac{u}{1+u}P,\frac{u}{1+u}P\right), & \theta_1<\theta\le\theta_2 \\ (P,0), & 0\le\theta\le\theta_1 \end{cases} \text{ or } \begin{cases} (0,1), & \theta_3<\theta\le 1 \\ (0,uP), & \tilde{\theta}_2<\theta\le\theta_3 \\ \left(\frac{u}{1+u}P,\frac{u}{1+u}P\right), & \theta_1<\theta\le\tilde{\theta}_2 \\ (P,0), & 0\le\theta\le\theta_1 \end{cases},$$

where $\theta_1$, $\theta_2$, $\tilde{\theta}_2$, and $\theta_3$ are cutoffs, whose values are pinned down by the Lagrangian multipliers.

We then optimize over $P\in[\kappa,1]$. We assume that at the optimum, the sender can

recommend a strictly positive measure of agents, otherwise the sender's maximal payoff is zero. It can be shown that for a given $P$, if the solution takes the second form, it is strictly dominated by the solution when $P = \kappa$. Furthermore, it can be shown that $P = \kappa$ is also not optimal. Therefore, the optimal direct mechanism takes the first form with $P > \kappa$, which represents a mechanism that takes exactly the four-segment form presented in Theorem 1.

# 4 Applications

The model has a rich set of applications. This section discusses four examples: education and placement, industrial policy, startup incubation, and bank runs.

## 4.1 Education and Placement

A natural application is education. The sender is a school or program; the agents are students; and the receiver is the labor market, employers, or another institution that evaluates students after they leave the school. The treatment is a scarce educational resource, such as access to strong instructors, small honors sections, research positions, mentoring, internships, or placement support.

When these resources are allocated, the school does not yet know students' true ability. Admission files may not reveal which students will become strong candidates later, while students may have private information about their preparation, motivation, or fit. All students value scarce educational resources and successful placement.

After students take courses, complete projects, and interact with faculty, the school learns much more about their ability. It can then recommend students through grades, honors, or recommendation letters. Employers do not observe all of the school's internal information, so they decide whether to hire recommended students based on the expected quality of the recommended pool.

This maps directly into the model. The school wants to maximize the number of students accepted by employers. It has a limited treatment that improves student quality, but must allocate it before fully learning students' ability. Later, once ability is better observed, the school can recommend students to employers. The receiver's obedience constraint captures the credibility of these recommendations: if the school recommends too many weak students, employers discount the signal.

The model suggests that educational resources and placement support should be designed jointly. A school can implement this logic through honors programs or academic tracks. These tracks may differ both in access to scarce educational resources and in the probability of later recommendation. Students in a less resource-intensive track may receive

fewer educational inputs, but may also face stronger placement support or a lower threshold for endorsement if they perform well.

Such tracks induce self-selection. Students choose based not only on current resources, but also on the future recommendation attached to each track. The school can therefore use track choice to identify a target group before it fully observes students' ability, implementing the same logic as the optimal joint mechanism.

### 4.2 Industrial Policy and Firm Certification

A second application is local industrial policy. The sender is a local government; the agents are firms; and the receiver is an upper-level government. The treatment is limited policy support, such as access to an industrial park, subsidized land, tax relief, talent-policy support, loan guarantees, infrastructure support, or procurement opportunities.

The local government wants to cultivate firms that can later be recognized by the upper-level government as important firms or strategic industries. Such recognition may serve as evidence of local policy success. When local support is allocated, however, the government cannot perfectly distinguish high-potential firms from low-potential firms mainly seeking subsidies. Firms know more about their own productivity, technology, commitment, and growth prospects. All firms value policy support and recognition by the upper-level government.

Over time, the local government learns much more about firm quality. It observes production, employment, tax payments, project completion, orders, financing, patents, exports, and other performance measures. It can then recommend firms to the upper-level government. The upper-level government does not observe all local information, so it relies on the local government's recommendation. The recommendation must remain credible: if too many weak firms are recommended, the upper-level government will discount the signal.

This setting captures a central difficulty in industrial policy. Early policy support is valuable, but it may attract weak firms mainly interested in obtaining subsidies. The local government would like scarce support to reach firms that can later become credible candidates for upper-level recognition, but these firms are difficult to identify at the initial allocation stage.

The model highlights that industrial policy is not only about allocating subsidies. It is also about designing later certification. A local government may use future recommendation to induce firms to reveal information before scarce support is allocated. Some firms may accept less policy support in exchange for a better chance of later certification if they meet performance standards. This allows the government to save scarce resources while still identifying firms that can become valuable candidates for upper-level recognition.

### 4.3 Incubators and Accelerators

A third application is startup incubation. The sender is an incubator, accelerator, university entrepreneurship center, or local startup platform; the agents are startups; and the receiver is a group of later-stage investors, strategic partners, acquirers, or other outside evaluators. The treatment is a scarce developmental input, such as seed funding, office space, technical assistance, mentorship, legal support, cloud credits, network access, or investor introductions.

Early in the process, the incubator does not know which startups are truly high quality. Many startups are still at the idea or prototype stage, and the available information may consist mainly of founders' claims, pitch decks, and limited early traction. Founders may know more about their own technology, effort, or market fit than the incubator does. All startups value support from the incubator, and all startups value being recommended to investors or partners later.

After the incubation period, the incubator learns much more. It observes whether the team builds the product, whether customers respond, whether founders execute, and whether the business model improves. At that point, the incubator can recommend startups through demo days, curated introductions, endorsement letters, or access to a preferred investor network. Investors do not observe all of the incubator's private information, but they may trust the recommendation if the recommended pool is sufficiently strong.

This application captures the joint role of early support and later certification. If early support is allocated without using the later recommendation rule, some resources may go to startups that never become investor-ready. The incubator would instead like support to reach startups that are more likely to become credible candidates for later financing, but these startups are difficult to identify at the beginning.

The model suggests that incubators should design early support and later investor access jointly. Some startups may accept less early support in exchange for a better chance of later investor introduction if they reach certain milestones. This contract is attractive only to startups that expect to benefit from the later recommendation. The incubator can therefore use the promise of future certification to identify a target group and reserve more intensive resources for startups that are more valuable to the later recommendation pool.

### 4.4 Bank Runs and Liquidity Support

A fourth application is a bank-run setting. The sender is a central bank or financial regulator; the agents are financial institutions; and the receiver is depositors. The sender's objective is to minimize the number of institutions that fail because depositors run. The treatment is

limited liquidity support, such as emergency lending, access to a standing facility, discount-window lending, or other stabilization tools. The recommendation is a public assurance or supervisory certification that an institution is sufficiently sound.

This setting is relevant when systemic risk begins to appear. Early in a stress episode, the central bank may first use liquidity support to stabilize institutions and markets. At that stage, it may need to act before fully assessing each institution's condition. A full audit of all institutions may be too costly and too slow. At the same time, every institution wants support: during financial stress, each institution would like to protect itself from a possible run. The central bank, however, would prefer to support institutions that can survive with liquidity assistance, rather than institutions for which support is unlikely to make a difference.

Public assurance or certification becomes more important when financial stress becomes more severe. At that point, liquidity support alone may not be enough to prevent a run. As supervisory information improves, the central bank can communicate with depositors through public statements, supervisory assessments, or certification. Depositors do not observe all supervisory information, so they rely on the central bank's signal. The signal must remain credible: if the central bank reassures too many weak institutions, depositors will discount the reassurance and may still run.

This maps into the model. Liquidity support is scarce and can improve an institution's chance of surviving a period of stress, but it must be allocated before the central bank fully knows institutional quality. Later, once more information is available, the central bank can decide which institutions to reassure. Depositors accept an institution by keeping their funds in it; they reject it by running. The receiver's obedience constraint captures the credibility of public reassurance: depositors stay only if reassured institutions are expected to be sufficiently sound.

The model suggests that liquidity support and public reassurance should be designed jointly. A central bank may use future reassurance to induce institutions to reveal information before support is allocated. Some institutions may accept lower access to liquidity support in exchange for a more favorable reassurance rule if support is realized and later information is sufficiently good. This allows the central bank to save scarce support while still identifying institutions that can become credible candidates for public reassurance.

Across these examples, the same logic appears. The sender controls an early scarce resource and a later recommendation. The early resource is valuable, but the sender does not yet know which agents are best to treat. The main implication of the model is that these two instruments should be designed jointly: future recommendation can be used not only to

persuade the receiver, but also to improve the allocation of scarce resources before quality is fully observed.

## 5 Conclusion

This paper studies a joint mechanism and information design problem. A sender wants to maximize the mass of agents accepted by a receiver. The sender has two instruments: a scarce treatment that can improve agents' quality, and a later recommendation that affects the receiver's belief. The key friction is informational timing. Treatment must be allocated before the sender fully observes agents' types, while recommendation occurs later, after more information is revealed.

The main result is that treatment allocation and recommendation should be designed jointly. The optimal joint mechanism targets an intermediate group: the sender reduces treatment probability not for the lowest or highest types, but for agents in the middle. These agents receive a lower probability of treatment, but are compensated by a recommendation when treatment is realized. This non-monotone treatment allocation allows the sender to save treatment capacity and redirect it toward higher types, whose treated quality is more valuable for satisfying the receiver's acceptance constraint. The mechanism therefore uses recommendation in two ways. It communicates information to the receiver, and it also helps elicit information from agents before treatment is allocated.

The model applies to many settings. In education, schools allocate scarce educational resources before fully learning students' ability and later recommend students to employers. In industrial policy, local governments allocate limited support before fully learning firms' quality and later promote selected firms to upper-level governments as evidence of local policy success. In startup incubation, accelerators allocate scarce support before fully learning startups' potential and later recommend startups to investors. In financial regulation, regulators allocate limited liquidity support before fully learning banks' quality and later reassure depositors about selected banks. Across these applications, the central takeaway is the same: later recommendation should not be treated only as an information device, but also as an incentive device that improves the early allocation of scarce resources.

## References


**Baron, David P. and Roger B. Myerson**, "Regulating a Monopolist with Unknown Costs," *Econometrica*, 1982, *50* (4), 911–930.

**Ben-Porath, Elchanan, Eddie Dekel, and Barton L Lipman**, "Optimal allocation with costly verification," *American Economic Review*, 2014, *104* (12), 3779–3813.

**Bergemann, Dirk and Martin Pesendorfer**, "Information Structures in Optimal Auctions," *Journal of Economic Theory*, 2007, *137* (1), 580–609.

_ **and Stephen Morris**, "Information Design, Bayesian Persuasion, and Bayes Correlated Equilibrium," *American Economic Review*, 2016, *106* (5), 586–591.

_ **and** _ , "Information Design: A Unified Perspective," *Journal of Economic Literature*, 2019, *57* (1), 44–95.

_ **, Tibor Heumann, and Stephen Morris**, "Screening with Persuasion," *Journal of Political Economy*, 2026. Forthcoming.

**Boleslavsky, Raphael and Kyungmin Kim**, "Bayesian Persuasion and Moral Hazard," 2018. SSRN working paper.

**Camboni, Matteo and Christoph Carnehl**, "Inputs or Outputs: What to Test and How to Test," in "Proceedings of the 26th ACM Conference on Economics and Computation" 2025, p. 577.

**Dranove, David and Ginger Zhe Jin**, "Quality Disclosure and Certification: Theory and Practice," *Journal of Economic Literature*, 2010, *48* (4), 935–963.

**Dworczak, Piotr**, "Mechanism Design with Aftermarkets: Cutoff Mechanisms," *Econometrica*, 2020, *88* (6), 2629–2661.

**Kamenica, Emir**, "Bayesian Persuasion and Information Design," *Annual Review of Economics*, 2019, *11*, 249–272.

_ **and Matthew Gentzkow**, "Bayesian Persuasion," *American Economic Review*, 2011, *101* (6), 2590–2615.

**Kolotilin, Anton**, "Optimal Information Disclosure: A Linear Programming Approach," *Theoretical Economics*, 2018, *13* (2), 607–635.

_ **, Tymofiy Mylovanov, Andriy Zapechelnyuk, and Ming Li**, "Persuasion of a privately informed receiver," *Econometrica*, 2017, *85* (6), 1949–1964.

**Laffont, Jean-Jacques and David Martimort**, *The Theory of Incentives: The Principal-Agent Model*, Princeton, NJ: Princeton University Press, 2002.

**Lizzeri, Alessandro**, "Information Revelation and Certification Intermediaries," *The RAND Journal of Economics*, 1999, *30* (2), 214–231.

**Martimort, David and Aggey Semenov**, "Continuity in Mechanism Design without Transfers," *Economics Letters*, 2006, *93* (2), 182–189.

**Maskin, Eric and John Riley**, "Monopoly with Incomplete Information," *The RAND Journal of Economics*, 1984, *15* (2), 171–196.

**Miralles, Antonio**, "Cardinal Bayesian Allocation Mechanisms without Transfers," *Journal of Economic Theory*, 2012, *147* (1), 179–206.

**Mussa, Michael and Sherwin Rosen**, "Monopoly and Product Quality," *Journal of Economic Theory*, 1978, *18* (2), 301–317.

**Myerson, Roger B.**, "Optimal Auction Design," *Mathematics of Operations Research*, 1981, *6* (1), 58–73.

**Rayo, Luis and Ilya Segal**, "Optimal Information Disclosure," *Journal of Political Economy*, 2010, *118* (5), 949–987.

**Saeedi, Maryam and Ali Shourideh**, "Optimal Rating Design under Moral Hazard," 2026. Working paper, arXiv:2008.09529.

**Schummer, James and Rakesh V. Vohra**, "Mechanism Design without Money," in Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani, eds., *Algorithmic Game Theory*, Cambridge: Cambridge University Press, 2007, chapter 10, pp. 243–265.

**Xiao, Peiran**, "Incentivizing Agents through Ratings," 2024. Working paper, arXiv:2407.10525.

**Zapechelnyuk, Andriy**, "Optimal Quality Certification," *American Economic Review: Insights*, 2020, *2* (2), 161–176.

**Zubrickas, Robertas**, "Optimal grading," *International Economic Review*, 2015, *56* (3), 751–776.

## Appendix A Proof of Theorem 1

In this section, we provide a characterization for the solution to problem (2.1), which is restated below.

$$\begin{aligned} \max_{p,q} \quad & \int_\Theta [p(\theta)q(\theta,\theta,1) + (1-p(\theta))q(\theta,\theta,0)]\,\mathrm{d}F(\theta), \\ \text{s.t.} \quad & \int_\Theta p(\theta)\,\mathrm{d}F(\theta) \le \kappa, && \text{(Capacity)} \\ & \mu_\pi \ge \underline{\theta}, && \text{(Obedience)} \\ & U_\pi(\theta,\theta) \ge U_\pi(\theta,\hat{\theta}), \quad \forall \theta,\hat{\theta} \in \Theta, && \text{(IC)} \end{aligned} \tag{2.1}$$

with the assumption that $u(\theta) \equiv u$, $v(\theta) \equiv v$.

**Lemma A1.** There exists an optimal mechanism $(p,q)$ that satisfies $q(\hat{\theta},\theta,d) = 0$ for all $\hat{\theta} \neq \theta$ and $d \in \{0,1\}$.

*Proof.* In problem (2.1), when $\hat{\theta} \neq \theta$, $q(\hat{\theta},\theta,d)$ shows up only in the IC constraint, which expands to

$$p(\theta)[u + q(\theta,\theta,1)] + (1-p(\theta))q(\theta,\theta,0) \ge p(\hat{\theta})[u + q(\hat{\theta},\theta,1)] + (1-p(\hat{\theta}))q(\hat{\theta},\theta,0).$$

Clearly, the right-hand side is increasing in $q(\hat{\theta},\theta,0)$ and $q(\hat{\theta},\theta,1)$. Therefore, it is optimal to set them such that the IC constraint is as slack as possible, i.e., $q(\hat{\theta},\theta,0) = q(\hat{\theta},\theta,1) = 0$ whenever $\hat{\theta} \neq \theta$. This is essentially assigning the harshest punishment available whenever a lie is detected. □

Lemma A1 implies that we can restrict our attention to the mechanisms that impose the harshest punishment to agents that misreport.

We define the "'ex-ante recommendation rule" $Q : [0,1] \to [0,1]$ as

$$Q(\theta) := p(\theta)q(\theta,\theta,1) + [1-p(\theta)]q(\theta,\theta,0).$$

"Ex-ante" here means before assigning the treatment.

With Lemma A1, problem (2.1) can be rewritten using the definition of $Q$ as

$$\begin{aligned} \max_{p,q} \quad & \int_\Theta Q(\theta)\mathrm{d}F(\theta), & \\ \text{s.t.} \quad & \int_\Theta p(\theta)\,\mathrm{d}F(\theta) \le \kappa, & \text{(Capacity)} \\ & \int_\Theta [(\theta-\underline{\theta})Q(\theta) + vp(\theta)q(\theta,\theta,1)]\mathrm{d}F(\theta) \ge 0, & \text{(Obedience)} \\ & Q(\theta) + p(\theta)u \ge p(\hat{\theta})u, \quad \forall \theta,\hat{\theta}\in\Theta. & \text{(IC)} \end{aligned} \tag{A.1}$$

The obedience constraint is expanded and simplified according to the definition of $\mu_\pi$. The IC constraint can be further rewritten as

$$Q(\theta) + p(\theta)u \ge \sup_{\hat{\theta}\in\Theta} p(\hat{\theta})u, \quad \forall \theta\in\Theta.$$

**Lemma A2.** The optimal mechanism $(p,q)$ satisfies

$$\begin{aligned} q(\theta,\theta,1) &= \min\left\{1, \frac{Q(\theta)}{p(\theta)}\right\}, \\ q(\theta,\theta,0) &= \max\left\{\frac{Q(\theta)-p(\theta)}{1-p(\theta)}, 0\right\}. \end{aligned}$$

When $p(\theta)\in\{0,1\}$, the value of $q(\theta,\theta,1-p(\theta))$ is irrelevant.

*Proof.* Given fixed $p(\theta)$ and $Q(\theta)$, the obedience constraint becomes slacker when $q(\theta,\theta,1)$ increases. Therefore, it is optimal to set $q(\theta,\theta,1)$ to be its maximal feasible value. The only two constraints on the upper bound of $q(\theta,\theta,1)$ are $q(\theta,\theta,1)\le 1$, and

$$p(\theta)q(\theta,\theta,1) + [1-p(\theta)]q(\theta,\theta,0) = Q(\theta) \le 1 \implies q(\theta,\theta,1) \le \frac{Q(\theta)}{p(\theta)},$$

because $p(\theta)$, $q(\theta,\theta,0)$ and $1-p(\theta)$ are non-negative. Hence, the optimal $q(\theta,\theta,1)$ is $\min\{1, \frac{Q(\theta)}{p(\theta)}\}$, and the corresponding optimal $q(\theta,\theta,0)$ takes the value such that $p(\theta)q(\theta,\theta,1)+[1-p(\theta)]q(\theta,\theta,0) = Q(\theta)$. □

Lemma A2 implies that it is sufficient for the sender to design $p$ and ex-ante recommendation rule $Q$ as the mechanism. The optimal treatment-specific recommendation rule $q(\theta,\theta,0)$ and $q(\theta,\theta,1)$ can be recovered from the optimal $p$ and $Q$. This further simplifies

the sender's problem into

$$\begin{aligned}
\max_{p,Q} \quad & \int_\Theta Q(\theta)\mathrm{d}F(\theta), \\
\text{s.t.} \quad & \int_\Theta p(\theta)\,\mathrm{d}F(\theta) \le \kappa, && \text{(Capacity)} \\
& \int_\Theta [(\theta-\underline{\theta})Q(\theta) + v\min\{p(\theta),Q(\theta)\}]\mathrm{d}F(\theta) \ge 0, && \text{(Obedience)} \\
& Q(\theta) + p(\theta)u \ge \sup_{\hat\theta\in\Theta} p(\hat\theta)u, \quad \forall\theta\in\Theta. && \text{(IC)}
\end{aligned} \tag{A.2}$$

We hereby refer to $(p, Q)$ also as a mechanism.

Define $P = \sup_{\hat\theta\in\Theta} p(\hat\theta)$. Since the capacity constraint requires exactly $\kappa$ mass of agents to be treated, we know that $P \in [\kappa, 1]$. To solve (A.2), we follow a two-step procedure to linearize the constraints. We first characterize the optimal $(p, Q)$ for a given $P$ by solving

$$\begin{aligned}
\max_{p,Q} \quad & \int_\Theta Q(\theta)\mathrm{d}F(\theta), \\
\text{s.t.} \quad & \int_\Theta p(\theta)\,\mathrm{d}F(\theta) \le \kappa, && \text{(Capacity)} \\
& \int_\Theta [(\theta-\underline{\theta})Q(\theta) + v\min\{p(\theta),Q(\theta)\}]\mathrm{d}F(\theta) \ge 0, && \text{(Obedience)} \\
& Q(\theta) + p(\theta)u \ge Pu, \quad \forall\theta\in\Theta, && (\mathrm{IC}_1) \\
& p(\theta) \le P, \quad \forall\theta\in\Theta, && (\mathrm{IC}_2)
\end{aligned} \tag{A.3}$$

We then substitute the resulting solution $(p_P, Q_P)$ into the objective and choose $P$ to solve

$$\max_{P\in[\kappa,1]} \int_\Theta Q_P(\theta)\mathrm{d}F(\theta), \quad \text{s.t.} \quad (p_P, Q_P) \text{ solves (A.3)}.$$

To solve problem (A.3), we introduce Lagrangian multipliers for the capacity and obedience constraints:

$$\begin{aligned}
\max_{p,Q} \quad & \int_\Theta \big[Q(\theta) - \lambda p(\theta) + \eta((\theta-\underline{\theta})Q(\theta) + v\min\{p(\theta),Q(\theta)\})\big]\mathrm{d}F(\theta) + \lambda\kappa, \\
\text{s.t.} \quad & (\mathrm{IC}_1)\ Q(\theta) + p(\theta)u \ge Pu, \quad \forall\theta\in\Theta, \\
& (\mathrm{IC}_2)\ p(\theta) \le P, \quad \forall\theta\in\Theta.
\end{aligned} \tag{A.4}$$

**Lemma A3.** For a fixed $P \in [\kappa, 1]$, if the constraint set of (A.3) is non-empty, $(p^*, Q^*)$ is the solution to problem (A.3) if and only if there exists $\lambda, \eta \ge 0$ satisfying the KKT conditions such that $(p^*, Q^*)$ is also a solution to problem (A.4).

*Proof.* In problem (A.3), the objective function is linear in $(p, Q)$. All constraints except the obedience constraint are affine in $(p, Q)$. It is easy to verify that $\min\{p(\theta), Q(\theta)\}$ is a concave function in $(p(\theta), Q(\theta))$, and therefore the left-hand side of the obedience constraint is concave in $(p, Q)$. Hence, the constraint set is a convex set, and therefore problem (A.3) is a convex optimization problem.

We can linearize the constraint set by replacing $\min\{p(\theta), Q(\theta)\}$ with $m(\theta)$ and introducing constraint $p(\theta), Q(\theta) \geq m(\theta) \geq 0$. The new constraint set is affine in $(p, Q, m)$. For an affine constraint set, the Slater's condition only requires that a feasible point exists. As long as the original constraint set is non-empty, $(p, Q, \min\{p, Q\})$ is feasible in the linearized constraint set. Therefore, the Slater's condition is satisfied, which finishes the proof. □

Lemma A3 implies that to solve problem (A.3), we can first solve problem (A.4) for any given pair of multipliers, and then find a specific combination of multipliers such that its corresponding solution to (A.4) also satisfies the capacity and obedience constraints.

Notice that the objective function of (A.3) is increasing in $Q$, and that the IC constraint becomes slacker when $Q$ increases. Since the obedience constraint is continuous in $Q$, as long as it is slack, the objective can be improved by increasing $Q$ slightly without violating the constraints. Therefore, at optimum, the obedience constraint binds, and hence $\eta > 0$.

Given $\lambda$ and $\eta$, problem (A.4) is a pointwise maximization problem with linear constraints. Let $(p_{\lambda,\eta}, Q_{\lambda,\eta})$ be the solution to problem (A.4) given the multipliers $\lambda$ and $\eta$. We further define

$$\Lambda = \frac{\lambda}{\eta}, \ \mathrm{H} = \underline{\theta} - \frac{1}{\eta}.$$

**Proposition A1.** Given $\lambda, \eta \geq 0$, the solution to problem (A.4) is given by

$$(p_{\lambda,\eta}(\theta), Q_{\lambda,\eta}(\theta)) = \begin{cases} (P, 1), & \mathrm{H} \leq \theta, \ \Lambda \leq v, \\ (0, 1), & \mathrm{H} \leq \theta, \ \Lambda \geq v, \\ (P, P), & \mathrm{H} \geq \theta, \ \Lambda + \mathrm{H} \leq \theta + v, \\ (P, 0), & \mathrm{H} \geq \theta, \ \mathrm{H} - \dfrac{\Lambda}{u} \geq \theta + v, \\ (0, uP), & \mathrm{H} \geq \theta, \ \mathrm{H} - \dfrac{\Lambda}{u} \leq \theta - \dfrac{v}{u}, \\ \left(\dfrac{u}{1+u}P, \dfrac{u}{1+u}P\right), & \text{otherwise.} \end{cases} \tag{A.5}$$

*Proof.* Problem (A.4) is a pointwise optimization problem. For a given $\theta$, we introduce $m = \min\{p(\theta), Q(\theta)\}$ to linearize the problem. Equivalently, we can rewrite the problem

into the following linear programming problem with $\Lambda$ and H:

$$\begin{aligned} \max_{p(\theta),Q(\theta),m} \quad & -\Lambda \cdot p(\theta) + (\theta - \mathrm{H}) \cdot Q(\theta) + v \cdot m \\ \text{s.t.} \quad & Q(\theta) + p(\theta)u \geq Pu, \\ & p(\theta) \leq P, \\ & 0 \leq m \leq p(\theta), Q(\theta) \leq 1. \end{aligned}$$

Here, we scaled the objective function by $\eta^{-1}$. We know that $\Lambda \geq 0$.

The objective function is optimized at at least one of the extreme points of the constraint set. All extreme points such that at least one of $m = p(\theta)$ and $m = Q(\theta)$ holds and the corresponding values of the objective function are listed in Table A.1. For convenience, we label the values $V_1$–$V_6$. The solution to the problem is the extreme point with the highest corresponding objective function value.

| Extreme point $(p(\theta), Q(\theta), m)$ | Value of objective |
|---|---|
| $(P, 1, P)$ | $(v - \Lambda)P + \theta - \mathrm{H} =: V_1$ |
| $(0, 1, 0)$ | $\theta - \mathrm{H} =: V_2$ |
| $(P, P, P)$ | $(v - \Lambda + \theta - \mathrm{H})P =: V_3$ |
| $(P, 0, 0)$ | $-\Lambda P =: V_4$ |
| $(0, uP, 0)$ | $(\theta - \mathrm{H})uP =: V_5$ |
| $\left(\frac{u}{1+u}P, \frac{u}{1+u}P, \frac{u}{1+u}P\right)$ | $(v - \Lambda + \theta - \mathrm{H})\frac{u}{1+u}P =: V_6$ |

Table A.1. Extreme points and corresponding values of the objective function

**Case 1:** $\mathrm{H} \leq \theta,\ \Lambda \leq v$**.**

In this case, $v - \Lambda,\ \theta - \mathrm{H} \geq 0$. For any $a \in [0, P], b \in [0, 1]$, $(v - \Lambda)a + (\theta - \mathrm{H})b \leq (v - \Lambda)P + \theta - \mathrm{H} = V_1$. Therefore, $V_1 \geq V_2, V_3, V_5, V_6 \geq 0 \geq V_4$, and the solution is $(p(\theta), Q(\theta)) = (P, 1)$.

**Case 2:** $\mathrm{H} \leq \theta,\ \Lambda \geq v$**.**

In this case, $v - \Lambda \leq 0,\ \theta - \mathrm{H} \geq 0$. For any $a \in [0, P], b \in [0, 1]$, $(v - \Lambda)a + (\theta - \mathrm{H})b \leq \theta - \mathrm{H} = V_2$. Therefore, $V_2 \geq V_1, V_3, V_5, V_6$. Also, $V_2 \geq 0 \geq V_4$, and hence the solution is $(p(\theta), Q(\theta)) = (0, 1)$.

**Case 3:** $\mathrm{H} \geq \theta,\ \Lambda + \mathrm{H} \leq \theta + v$**.**

In this case, it is easy to verify that $v - \Lambda \geq 0$, $\theta - \mathrm{H} \leq 0$, and $v - \Lambda + \theta - \mathrm{H} \geq 0$. Therefore, $V_2, V_4, V_5 \leq 0 \leq V_3$. Furthermore,

$$V_1 = (v - \Lambda)P + \theta - \mathrm{H} \leq (v - \Lambda)P + (\theta - \mathrm{H})P = V_3,$$

and

$$V_6 = (v - \Lambda + \theta - \mathrm{H})\frac{u}{1+u}P \leq (v - \Lambda + \theta - \mathrm{H})P = V_3.$$

Hence, $V_3$ is the largest, and the solution is $(p(\theta), Q(\theta)) = (P, P)$.

**Case 4:** $\mathrm{H} \geq \theta,\ \mathrm{H} - \frac{\Lambda}{u} \geq \theta + v$**.**

In this case, it is easy to verify that $\theta - \mathrm{H} \leq 0$ and $v - \Lambda + \theta - \mathrm{H} \leq 0$. Therefore, $V_1 \leq V_3 \leq V_6 \leq 0$, $V_2 \leq V_5 \leq 0$.

Since $\mathrm{H} - \frac{\Lambda}{u} \geq \theta + v$, we know that $\mathrm{H} - \frac{\Lambda}{u} - \theta \geq v$, and that

$$V_4 - V_5 = \left(\mathrm{H} - \theta - \frac{\Lambda}{u}\right) uP \geq vuP \geq 0 \implies V_4 \geq V_5.$$

Furthermore, $\mathrm{H} - \frac{\Lambda}{u} - \theta - v \geq 0$, and

$$\begin{aligned} V_4 - V_6 &= \left(-\frac{1+u}{u}\Lambda - v + \Lambda - \theta + \mathrm{H}\right)\frac{u}{1+u}P \\ &= \left(\mathrm{H} - \frac{\Lambda}{u} - \theta - v\right)\frac{u}{1+u}P \\ &\geq 0 \implies V_4 \geq V_6. \end{aligned}$$

Hence, $V_4$ is the largest, and the solution is $(p(\theta), Q(\theta)) = (P, 0)$.

**Case 5:** $\mathrm{H} \geq \theta,\ \mathrm{H} - \frac{\Lambda}{u} \leq \theta - \frac{v}{u}$**.**

In this case, it is easy to verify that $v - \Lambda \leq 0$ and $\theta - \mathrm{H} \leq 0$. Therefore, $V_3 \leq V_6 \leq 0$, $V_1 \leq V_2 \leq V_5 \leq 0$.

Since $\mathrm{H} - \frac{\Lambda}{u} \leq \theta - \frac{v}{u}$ and therefore $\theta - \mathrm{H} + \frac{\Lambda}{u} \geq \frac{v}{u}$, we have

$$V_5 - V_4 = \left(\theta - \mathrm{H} + \frac{\Lambda}{u}\right) uP \geq \frac{v}{u}uP \geq 0 \implies V_5 \geq V_4,$$

and

$$V_5 - V_6 = [(1+u)(\theta - \mathrm{H}) - (v - \Lambda + \theta - \mathrm{H})]\frac{u}{1+u}P$$

$$
\begin{aligned}
&= \left(\theta - \mathrm{H} + \frac{\Lambda}{u} - \frac{v}{u}\right)\frac{u^2}{1+u}P \\
&\geq \left(\frac{v}{u} - \frac{v}{u}\right)\frac{u^2}{1+u}P \\
&= 0 \implies V_5 \geq V_6.
\end{aligned}
$$

Hence, $V_5$ is the largest, and the solution is $(p(\theta), Q(\theta)) = (0, uP)$.

**Case 6:** $\mathrm{H} \geq \theta,\ \Lambda + \mathrm{H} \geq \theta + v,\ \theta - \frac{v}{u} \leq \mathrm{H} - \frac{\Lambda}{u} \leq \theta + v$**.**

In this case, it is easy to verify that $\theta - \mathrm{H} \leq 0$ and $v - \Lambda + \theta - \mathrm{H} \leq 0$. Therefore, $V_1 \leq V_3 \leq V_6 \leq 0$, $V_2 \leq V_5 \leq 0$.

$\theta - \frac{v}{u} \leq \mathrm{H} - \frac{\Lambda}{u} \leq \theta + v$ implies $\theta - \mathrm{H} + v + \frac{\Lambda}{u} \geq 0$ and $\frac{v}{u} - \frac{\Lambda}{u} - \theta + \mathrm{H} \geq 0$. Therefore,

$$
\begin{aligned}
V_6 - V_4 &= \left(v - \Lambda + \theta - \mathrm{H} + \frac{1+u}{u}\Lambda\right)\frac{u}{1+u}P \\
&= \left(\theta - \mathrm{H} + v + \frac{\Lambda}{u}\right)\frac{u}{1+u}P \\
&\geq 0 \implies V_6 \geq V_4,
\end{aligned}
$$

and

$$
\begin{aligned}
V_6 - V_5 &= [v - \Lambda + \theta - \mathrm{H} - (1+u)(\theta - \mathrm{H})]\frac{u}{1+u}P \\
&= \left(\frac{v}{u} - \frac{\Lambda}{u} - \theta + \mathrm{H}\right)\frac{u^2}{1+u}P \\
&\geq 0 \implies V_6 \geq V_5.
\end{aligned}
$$

Hence, $V_6$ is the largest, and the solution is $(p(\theta), Q(\theta)) = (\frac{u}{1+u}P, \frac{u}{1+u}P)$.

In summary, the solution to problem (A.4) is as listed in (A.5). □

Figure A.1a provides an illustration for the results in Proposition A1 in $(\Lambda, \mathrm{H})$ space.

The second step, finding a combination of $\lambda$ and H such that $(p_{\lambda,\mathrm{H}}, Q_{\lambda,\mathrm{H}})$ satisfies the capacity and obedience constraint, is not useful for characterizing the optimal mechanism.

By Proposition A1, there are two classes of mechanisms that are potentially optimal, illustrated by the two line segments in Figure A.1b. When $\Lambda \leq v$, the solution to problem

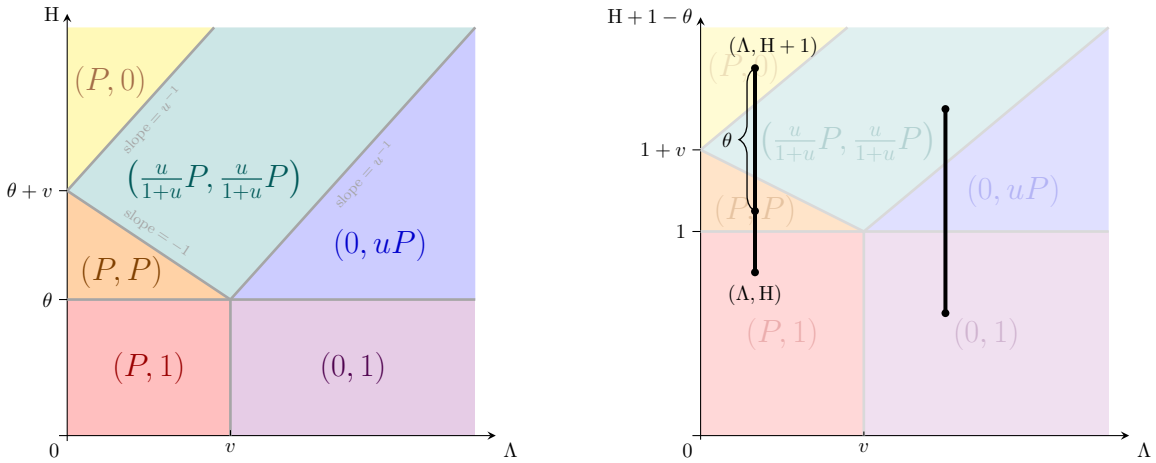


Figure A.1. Illustration of direct mechanisms.

(A.4) is a four-segment mechanism that takes the following form:

$$(p(\theta), Q(\theta)) = \begin{cases} (P,1), & \theta_3 < \theta \leq 1, \\ (P,P), & \theta_2 < \theta \leq \theta_3, \\ \left(\dfrac{u}{1+u}P, \dfrac{u}{1+u}P\right), & \theta_1 < \theta \leq \theta_2, \\ (P,0), & 0 \leq \theta \leq \theta_1, \end{cases} \tag{A.6}$$

where $P \in [\kappa, 1]$, and

$$\begin{aligned} \theta_1 &= \mathrm{H} - v - \frac{\Lambda}{u}, \\ \theta_2 &= \mathrm{H} - v + \Lambda, \\ \theta_3 &= \mathrm{H}. \end{aligned}$$

Each segment could be empty. This is the class of mechanisms that the optimal direct mechanism described in Theorem 1 falls into. When $\Lambda > v$, the solution to problem (A.4) is a mechanism that takes another four-segment form:

$$(p(\theta), Q(\theta)) = \begin{cases} (0,1), & \theta_3 < \theta \leq 1, \\ (0,uP), & \tilde{\theta}_2 < \theta \leq \theta_3, \\ \left(\dfrac{u}{1+u}P, \dfrac{u}{1+u}P\right), & \theta_1 < \theta \leq \tilde{\theta}_2, \\ (P,0), & 0 \leq \theta \leq \theta_1, \end{cases}$$

where $P \in [\kappa, 1]$, and

$$\tilde{\theta}_2 = \mathrm{H} + \frac{v}{u} - \frac{\Lambda}{u}.$$

Each segment could be empty. Lemma A4 shows that it is sufficient to consider the first class of mechanisms.

**Lemma A4.** Let $(p, Q)$ represent the optimal direct mechanism that solves the sender's problem (A.2). There exists $\lambda, \eta$ satisfying $\Lambda = \lambda/\eta \leq v$ such that $(p, Q)$ solves problem (A.4) given $P = \sup_{\hat{\theta} \in \Theta} p(\hat{\theta})$.

*Proof.* To prove this, it is sufficient to prove the following claim.

**Claim.** For a given $P$, let $\pi = (p, Q)$ be the mechanism that solves problem (A.3). If all corresponding pairs of multipliers $(\lambda, \eta)$ satisfy $\Lambda := \lambda/\eta > v$ (i.e., $(p, Q)$ falls only into the second class), then there exists another feasible mechanism $(p', Q')$ that strictly dominates $(p, Q)$ and has corresponding Lagrangian multipliers $\lambda', \eta'$ such that $\Lambda' := \lambda'/\eta' \leq v$.

We move on to prove this claim. When $\Lambda > v$, by Proposition A1, $(p, Q)$ must take the following form:

$$(p(\theta), Q(\theta)) = \begin{cases} (0, 1), & \theta_3 < \theta \leq 1, \\ (0, uP), & \tilde{\theta}_2 < \theta \leq \theta_3, \\ \left(\frac{u}{1+u}P, \frac{u}{1+u}P\right), & \theta_1 < \theta \leq \tilde{\theta}_2, \\ (P, 0), & 0 \leq \theta \leq \theta_1, \end{cases}$$

We assume $0 < \theta_1 < \tilde{\theta}_2 < 1$, i.e., the fourth segment and one of the first two segments is of strictly positive measure. This is without loss of generality because when the first two segments are measure zero, the mechanism also falls into the first class, and when the fourth segment is measure zero, $\operatorname{ess\,sup}_{\theta \in \Theta} p(\theta) = \frac{u}{1+u}P < P$, and is thus weakly dominated by the optimal mechanism of (A.3) when $P$ is replaced by $\frac{u}{1+u}P$. Since $F$ has positive density almost everywhere, this assumption implies $F(\theta_1) > 0$, $F(\tilde{\theta}_2) < 1$, and $\int_\Theta Q(\theta) \mathrm{d}F(\theta) > 0$.

Now, consider the following mechanism $\tilde{\pi}' = (\tilde{p}', \tilde{Q}')$:

$$\tilde{p}'(\theta) \equiv \kappa, \ \tilde{Q}'(\theta) = Q(\theta), \ \forall \theta \in \Theta.$$

It is easy to verify that $(\tilde{p}', \tilde{Q}')$ also satisfies capacity and IC constraints, and yields exactly the same amount of recommendation as $(p, Q)$.

Consider the obedience constraint. When $\kappa > \frac{u}{1+u}P$, it is easy to verify that $\min\{\tilde{p}'(\theta), \tilde{Q}'(\theta)\} \geq \min\{p(\theta), Q(\theta)\}$ for all $\theta \in \Theta$, and the inequality is strict when $\theta \in (\theta_1, \tilde{\theta}_2]$ which is of strictly

positive measure. Therefore, we have

$$\int_\Theta \min\{\tilde{p}'(\theta), \tilde{Q}'(\theta)\}\mathrm{d}F(\theta) > \int_\Theta \min\{p(\theta), Q(\theta)\}\mathrm{d}F(\theta).$$

When $\kappa \leq \frac{u}{1+u}P$, the capacity constraint gives

$$\int_\Theta p(\theta)\mathrm{d}F(\theta) \leq \kappa \implies PF(\theta_1) + \frac{u}{1+u}P(F(\tilde{\theta}_2) - F(\theta_1)) \leq \kappa.$$

This implies

$$\begin{aligned}
&\int_\Theta \min\{\tilde{p}'(\theta), \tilde{Q}'(\theta)\}\mathrm{d}F(\theta) - \int_\Theta \min\{p(\theta), Q(\theta)\}\mathrm{d}F(\theta) \\
&= \kappa(1 - F(\theta_1)) - \frac{u}{1+u}P(F(\tilde{\theta}_2) - F(\theta_1)) \\
&\geq \left[PF(\theta_1) + \frac{u}{1+u}P(F(\tilde{\theta}_2) - F(\theta_1))\right](1 - F(\theta_1)) - \frac{u}{1+u}P(F(\tilde{\theta}_2) - F(\theta_1)) \\
&= PF(\theta_1)\left(1 - \frac{u}{1+u}F(\tilde{\theta}_2) - \frac{1}{1+u}F(\theta_1)\right) \\
&\geq PF(\theta_1)(1 - F(\tilde{\theta}_2)) \\
&> 0.
\end{aligned}$$

Hence,

$$\int_\Theta \min\{\tilde{p}'(\theta), \tilde{Q}'(\theta)\}\mathrm{d}F(\theta) > \int_\Theta \min\{p(\theta), Q(\theta)\}\mathrm{d}F(\theta)$$

holds for both $\kappa > \frac{u}{1+u}P$ and $\kappa \leq \frac{u}{1+u}P$. Since $\tilde{Q}' = Q$, this further indicates that

$$\begin{aligned}
\mu_{\tilde{\pi}'} &= \frac{\displaystyle\int_\Theta \theta\tilde{Q}'(\theta)\mathrm{d}F(\theta) + v\int_\Theta \min\{\tilde{p}'(\theta), \tilde{Q}'(\theta)\}\mathrm{d}F(\theta)}{\displaystyle\int_\Theta \tilde{Q}'(\theta)\mathrm{d}F(\theta)} \\
&> \mu_\pi = \frac{\displaystyle\int_\Theta \theta Q(\theta)\mathrm{d}F(\theta) + v\int_\Theta \min\{p(\theta), Q(\theta)\}\mathrm{d}F(\theta)}{\displaystyle\int_\Theta Q(\theta)\mathrm{d}F(\theta)} \\
&\geq \underline{\theta},
\end{aligned}$$

i.e., the obedience constraint of $(\tilde{p}', \tilde{Q}')$ is strictly slack. Therefore, it is feasible and must be strictly dominated by the optimal solution to (A.3) given $P = \kappa$, which we denote by

$(p', Q')$.

Now, $\sup_{\theta\in\Theta} p'(\theta) = \kappa$ implies that its corresponding multipliers $(\lambda', \eta')$ satisfy $\Lambda' = \lambda'/\eta' = 0 < v$. Also, $(p', Q')$ strictly dominates $(\tilde{p}', \tilde{Q}')$ and therefore $(p, Q)$ since $\tilde{Q}' = Q$. This proves the claim and hence the lemma. □

A remark here is that for a given $P$, the direct mechanism that solves problem (A.3) may still correspond to a $\Lambda > v$. This is a sign that the supremum of $p(\theta)$, $P$, is chosen to be too large.

Lemma A4 implies that the optimal direct mechanism must take the form of (A.6). Applying the result of Lemma A2 to recover the recommendation rule $q$ from the ex-ante probability of recommendation $Q$ in (A.6) yields the exact four-segment form of the optimal direct mechanism as presented in Theorem 1. The last step is to exclude the case where $\sup_{\theta\in\Theta} p(\theta) = \kappa$.

**Proposition A2.** If an optimal mechanism to (A.2), $(p^*, Q^*)$, satisfies $\int_\Theta Q^*(\theta)\mathrm{d}F(\theta) > 0$, then $\sup_{\theta\in\Theta} p^*(\theta) > \kappa$.

*Proof.* When $p(\theta) \equiv \kappa$, $\sup_{\theta\in\Theta} p(\theta) = \kappa$, and when $P = \kappa$, the optimal $p$ in problem (A.3) is $p(\theta) \equiv \kappa$. We go on to prove that any mechanism with $p(\theta) \equiv \kappa$ is always improvable.

According to Proposition A1, the solution to problem (A.3) given $P = \kappa$ takes the following form:

$$p(\theta) \equiv \kappa,\ Q(\theta) = \begin{cases} 1, & \mathrm{H} < \theta \le 1, \\ \kappa, & \mathrm{H} - v < \theta \le \mathrm{H}, \\ 0, & 0 \le \theta \le \mathrm{H} - v, \end{cases}$$

where $\mathrm{H} = \underline{\theta} - \eta^{-1}$ and $(\lambda, \eta)$ are the corresponding multipliers satisfying $\Lambda = \lambda/\eta = 0$.

Denote by $(p^*, Q^*)$ the optimal solution to the sender's problem (A.2). The condition $\int_\Theta Q^*(\theta)\mathrm{d}F(\theta) > 0$ implies that at least someone is recommended in the best case. If $Q(\theta) = 0$ almost everywhere, it is clearly strictly dominated by $(p^*, Q^*)$ which must correspond to $P > \kappa$.

When $\int_\Theta Q(\theta)\mathrm{d}F(\theta) > 0$, given that the obedience constraint binds, $Q$ cannot be equal to 1 almost everywhere, and hence the second segment where $Q(\theta) = \kappa$ must be of strictly positive measure.

Problem (A.4) at $\Lambda = 0$ becomes

$$\begin{aligned} \max_{p,Q}\ & \int_\Theta [(\theta - \mathrm{H})Q(\theta) + v\min\{p(\theta), Q(\theta)\}]\mathrm{d}F(\theta), \\ \text{s.t. }\ & (\mathrm{IC}_1)\ Q(\theta) + p(\theta)u \ge \kappa u,\ \forall\theta\in\Theta, \\ & (\mathrm{IC}_2)\ p(\theta) \le \kappa,\ \forall\theta\in\Theta. \end{aligned}$$

Here, we scaled the objective function by $\eta^{-1}$. We introduce Lagrangian multipliers, $\nu(\theta)$, $\phi(\theta) \geq 0$, for constraint ($\text{IC}_1$) and ($\text{IC}_2$) respectively.

For a given $\theta$, the pointwise optimization problem is

$$\begin{aligned} \max_{p(\theta),Q(\theta)} \ & (\theta - \text{H})Q(\theta) + v \min\{p(\theta), Q(\theta)\}, \\ \text{s.t.} \ & (\text{IC}_1)\ Q(\theta) + p(\theta)u \geq \kappa u, \\ & (\text{IC}_2)\ p(\theta) \leq \kappa. \end{aligned}$$

We dropped the constraints $p(\theta), Q(\theta) \geq 0$ because it is never optimal to decrease $p(\theta)$, and $Q(\theta) \geq 0$ is implied by the two IC constraints.

When $\theta \in (\text{H}, 1]$, the optimal $(p(\theta), Q(\theta)) = (\kappa, 1)$ suggests that ($\text{IC}_1$) is slack, i.e. $\nu(\theta) = 0$, and $\min\{p(\theta), Q(\theta)\} = p(\theta)$. The Lagrangian problem is thus

$$\max_{p(\theta),Q(\theta)} (\theta - \text{H})Q(\theta) + vp(\theta) - \phi(\theta)(p(\theta) - \kappa).$$

First order condition gives $\phi(\theta) = v > 0$.

When $\theta \in (\text{H} - v, \text{H}]$, the optimal $(p(\theta), Q(\theta)) = (\kappa, \kappa)$ suggests that ($\text{IC}_1$) is slack, i.e. $\nu(\theta) = 0$. We write the problem equivalently as

$$\begin{aligned} \max_{p(\theta),Q(\theta),m} \ & (\theta - \text{H})Q(\theta) + v \cdot m, \\ \text{s.t.} \ & p(\theta) \leq \kappa, \\ & p(\theta) \geq m, \\ & Q(\theta) \geq m. \end{aligned}$$

Let $\alpha, \beta \geq 0$ be the Lagrangian multipliers of $p(\theta) \geq m$ and $Q(\theta) \geq m$ respectively. The corresponding Lagrangian problem is

$$\max_{p(\theta),Q(\theta)} (\theta - \text{H})Q(\theta) + v \cdot m - \phi(\theta)(p(\theta) - \kappa) - \alpha(m - p(\theta)) - \beta(m - Q(\theta)).$$

First order conditions give

$$\theta - \text{H} + \beta = 0, \qquad (\text{FOC-}Q(\theta))$$

$$-\phi(\theta) + \alpha = 0, \qquad (\text{FOC-}p(\theta))$$

$$v - \alpha - \beta = 0. \qquad (\text{FOC-}m)$$

These conditions yield

$$\phi(\theta) = \alpha = \theta - \mathrm{H} + v > 0, \ \beta = \mathrm{H} - \theta \geq 0.$$

When $\theta \in [0, \mathrm{H}-v]$, we know from the optimal $(p(\theta), Q(\theta)) = (\kappa, 0)$ that $\min\{p(\theta), Q(\theta)\} = Q(\theta)$. The corresponding Lagrangian problem is thus

$$\max_{p(\theta), Q(\theta)} (\theta - \mathrm{H} + v)Q(\theta) - \nu(\theta)(\kappa u - Q(\theta) - p(\theta)u) - \phi(\theta)(p(\theta) - \kappa).$$

First order condition of $p(\theta)$ gives

$$\nu(\theta)u - \phi(\theta) = 0.$$

In summary, at $P = \kappa$, the Lagrangian multipliers for incentive compatibility constraints in problem (A.4) and therefore in problem (A.3) satisfy $\nu(\theta) = 0$, $\phi(\theta) > 0$ for all $\theta \in (\mathrm{H} - v, 1] \neq \varnothing$, and $\nu(\theta)u - \phi(\theta) = 0$ for all $\theta \in [0, \mathrm{H} - v]$.

Denote the optimal value of problem (A.3) be $V$. By envelope theorem,

$$\begin{aligned}
\left.\frac{\partial V}{\partial P}\right|_{P=\kappa} &= \int_{\Theta} [-\nu(\theta)u + \phi(\theta)] \mathrm{d}F(\theta) \\
&= \int_{\theta \in (\mathrm{H}-v, 1]} [-\nu(\theta)u + \phi(\theta)] \mathrm{d}F(\theta) \ + \ \int_{\theta \in [0, \mathrm{H}-v]} [-\nu(\theta)u + \phi(\theta)] \mathrm{d}F(\theta) \\
&= \int_{\theta \in (\mathrm{H}-v, 1]} \phi(\theta) \mathrm{d}F(\theta) \ + \ 0 \\
&> 0.
\end{aligned}$$

Hence, when $P = \kappa$, the sender will be strictly better off by increasing $P$, and thus $p(\theta) \equiv \kappa$ cannot be supported in any optimal direct mechanism. □

Proposition A2 restricts the range of $P$ in (A.6) to $(\kappa, 1]$ whenever the sender can recommend a strictly positive measure of agents in the best situation, which finishes the proof of Theorem 1. Corollary 1.1 follows directly from Proposition A2.