Mechanism Design under Costly Signaling: the Value of Non-Coordination*

Yingkai Li† Xiaoyun Qiu‡

Abstract

We study the design of optimal allocation mechanism that uses costly signals from agents as screening devices. The principal is a social planner who aims to minimize total signaling costs given any monotone allocation rule. Payoff equivalence generally fails in these environments, making the details of implementing any allocation rule crucial. In particular, there is a tradeoff between coordinating agents' signals to reduce the costs incurred by losing agents and the risk of information leakage, which encourages potential winners to exert excessive effort based on their recommended signals to compete for the resource. We show that when agents' utilities exhibit decreasing absolute risk aversion for costly signaling, *not* coordinating their signaling choices and requiring them always to always bear the signaling costs regardless of their allocation is optimal for cost saving. In contrast, coordination is optimal under increasing absolute risk aversion. We further characterize the optimal non-coordination mechanism under several canonical costs of signaling. For example, we show that for cost functions reflecting the idea of costly manipulation, the optimal mechanism

*We thank Asher Wolinsky, Bruno Strulovici, and Wojciech Olszewski for advice, helpful conversations, and comments. We also thank Dirk Bergemann, Ian Ball, Eddie Dekel, Piotr Dworczak, Jeff Ely, Kira Goldner, Carl-Christian Groh, Yingni Guo, Marina Halac, Andrei Iakovlev, Annie Liang, Bart Lipman, Brendan Lucier, Deniz Kattwinkel, Joshua Mollner, Kyohei Okumura, Alessandro Pavan, Marcin Pęski, Abhishek Sarkar, James Schummer, Philipp Strack, Matthew Thomas, and participants at the 34th Stony Brook International Conference on Game Theory, NUS Theory Lunch for helpful suggestions, comments, and discussions. Yingkai Li acknowledges financial support from the Sloan Research Fellowship, grant no. FG-2019-12378 and NUS Start-up Grant.

†Department of Economics, National University of Singapore. Email: yk.li@nus.edu.sg

‡Department of Economics, Northwestern University. Email: xiaoyun.qiu@u.northwestern.edu

may exhibit randomization, and the value of randomization is strictly positive in large but finite markets.

Keywords— costly signaling, non-coordination, information leakage, mechanism design without money.

JEL— D47, D61, D82

1 Introduction

In many real-world applications, resources are allocated based on informative signals regarding agents' characteristics. For example, college admissions programs often select talented students based on test scores, such as SAT scores. Procurement processes typically use Request for Proposal (RFP) mechanisms to select qualified vendors.[1] Scientific funding agencies may rely on peer reviews of proposals to identify high-quality projects. Similarly, the public housing system allocates housing to low-income individuals or families based on their financial need or health conditions.

While these signals provide valuable information about the agents for resource allocation, they are also susceptible to manipulation through costly efforts, which may incentivize agents to invest inefficiently in socially wasteful signals. For example, students may fake disabilities to gain extra time on tests, thus achieving higher test scores that do not reflect increased intellectual ability (Sansone and Sansone, 2011).[2] Vendors may create fake companies to qualify for procurement opportunities they would otherwise be ineligible for or falsify documents to misrepresent costs from the source.[3] Applicants for scientific funding may invest effort in overstating the merits of their projects, rather than focusing on developing high-quality projects from the outset (Conix et al., 2021). Individuals applying for public housing may temporarily adopt a low-income lifestyle to gain a favorable position in the government's allocation process.

The main goal of this paper is to design mechanisms that minimize the total costs of signaling while maintaining allocation efficiency. This problem is not trivial because payoff equivalence (c.f., Myerson, 1981) does not hold under the general signaling cost structures

[1]See for instance, the information session from US small business administration, which provides detailed description on how procurement scoring works: `https://www.sba.gov/document/support-small-business-procurement-scorecard-overview`.

[2]Relatedly, on exams intended to test logical reasoning, students can sometimes achieve high scores by learning material by rote or by memorizing answers to past exams, which does not improve their logical reasoning ability.

[3]See for instance the military procurement fraud scheme: `https://www.justice.gov/opa/pr/military-contractors-indicted-7-million-procurement-fraud-scheme`.

we consider in this paper. As a result, different implementation of the same allocation rule could lead to different signaling costs, which are wasteful from the society's perspective. To illustrate the idea, consider a simple example with two agents and one principal. The agents' private types are drawn independently and identically from a uniform distribution F on $[0,1]$. Each agent i with private type θ_i can produce a public signal $s_i \in [0,\infty)$, at a cost $c_i = \max\{0, s_i - \theta_i\}$. That is, each type θ_i can inflate their signal to s_i at a cost $s_i - \theta_i$, while it is costless to produce a signal below their type.

The principal allocates one item between two agents based on their public signals. Specifically, the principal chooses an allocation vector (x_1, x_2), where $0 \le x_i \le 1$ for $i \in \{1,2\}$ and $x_1 + x_2 \le 1$. The agents' value for receiving the item is normalized to 1, so their utility is the probability of receiving the item minus the cost of signaling: $u_i = x_i - c_i$. Suppose the principal aims to implement the efficient allocation, that is, allocating the item to the agent with the higher type, while minimizing the total signaling costs, or equivalently, maximizing the total utilities of the agents.

To implement the efficient allocation, one intuitive approach is to adopt VCG-style mechanisms and use costly signals as instruments, similar to transfers, for screening the agents. The mechanism is structured as follows: only the agent receiving the item is required to generate costly signals. The required signal is set such that the agent's cutoff type for winning the item is indifferent between winning and not winning the item.

Example 1 (VCG-style mechanism)**.** *Each agent i reports a type $\hat{\theta}_i$ to the principal. For each reported type profile $\hat{\boldsymbol{\theta}}$, let $i^* = \arg\max_i \hat{\theta}_i$ be the agent with the higher reported type, and $s_i = 1 + \hat{\theta}_{-i}$ be the signal that agent i has to produce to get the item.*[4]

- *The principal's recommendation is for agent i^* to generate signal s_{i^*}, and for the other agent $i \neq i^*$ to generate signal* 0*.*
- *The principal allocates the item to agent i^* if and only if his signal is at least s_{i^*}. Otherwise the principal keeps the item.*

In this mechanism, by an argument similar to that used for VCG mechanisms, each agent has an incentive to truthfully report their type and produce the recommended signal, ensuring that the efficient allocation is implemented in equilibrium. However, this mechanism does not minimize the total cost of signaling. Specifically, for each agent to have an incentive to truthfully report their type, the higher-type agent must "burn" some utility

[4]The number 1 in the formula for s_i is each agent's valuation of the item. Under this signal recommendation, agent i's utility of winning and losing are $1 + \theta_i - s_i$ and 0 respectively when his type is θ_i. In order for the agent to be indifferent between winning and losing when $\theta_i = \hat{\theta}_{-i}$, we must have $s_i = 1 + \hat{\theta}_{-i}$.

through costly signaling to prove their higher type. More precisely, in expectation, each agent must exert a strictly positive effort of $1 + \mathbf{E}\big[\theta_{(2)} - \theta_{(1)} \,|\, \theta_{(2)} \leq \theta_{(1)}\big] = \frac{2}{3}$, where $\theta_{(2)}$ is the lower type and $\theta_{(1)}$ is the higher type.

However, we show that there exists a non-coordination mechanism, specifically a winner-takes-all (WTA) contest, which enables the principal to implement the efficient allocation without incurring any positive signaling costs.

Example 2 (WTA contest)**.** *The principal commits to a contest rule under which she allocates the item to the agent who generates the highest signal. Under this rule, it is an equilibrium for each agent to use the strategy* $s_i(\theta_i) = \theta_i$.

The WTA contest can also be represented as a direct mechanism where each agent is recommended a signal $s_i(\theta_i) = \theta_i$ after reporting their type. This is a non-coordination mechanism, where the recommended signal for each agent is a deterministic function of their own type and does not depend on the types or reports of other agents. In this mechanism, conditional on the other agents reporting truthfully and following the signal recommendation, each agent i's payoff from producing a signal $s_i \geq \theta_i$ is $F(s_i) - (s_i - \theta_i) = \theta_i$. Therefore, any deviation is not profitable for the agents, and truthfully reporting and following the recommendation not to produce costly signals is indeed an equilibrium. In both mechanisms above, it is easy to verify that the efficient allocation rule is implemented, while the non-coordination mechanism strictly outperforms the VCG-style mechanism in terms of cost efficiency.

Intuitively, one might conjecture that the VCG-style mechanism would outperform the non-coordination mechanism in terms of signaling costs. This is because the VCG-style mechanism only requires the agent to bear the signaling cost when they win the item, while the non-coordination mechanism imposes this cost on agents regardless of whether they win. However, our example suggests the opposite.

The correct intuition is that, in the VCG-style mechanism, coordination of signaling choices leads to information leakage. Upon receiving information or a signal recommendation indicating they are potential winners, lower-type agents may have a strong incentive to deviate by generating a high signal to secure the item. To deter such behavior, the signal required from the winner must be sufficiently high, leading to increased costs for the agent. In contrast, in non-coordination mechanisms, agents remain unaware of the realized types and signals of others. This uncertainty reduces the incentive for lower-type agents to mimic higher-type agents, as they cannot be sure of winning. Moreover, agents still bear the signaling cost even if they lose the item. As a result, the signal required from the high-type agent to deter the low-type agent can be kept low, reducing overall signaling costs. In Section 3,

we show that this intuitive idea extends beyond the simple example illustrated above, provided the agent exhibits constant absolute risk aversion with respect to the cost of signaling. Specifically, we show that under these utility assumptions, non-coordination mechanisms maximize the agents' utilities when implementing any feasible monotone allocation.

Given that non-coordination mechanisms are optimal in terms of cost savings, we now focus on characterizing the optimal non-coordination mechanisms. For any implementable allocation rule, the choice of the non-coordination mechanism is essentially unique, except for the choice of expected utility for the lowest type of each agent. Thus, our main focus here is to characterize the optimal mechanism that strikes the best balance between allocation efficiency and agents' utilities.

Characterizing the optimal non-coordination mechanism for general signaling cost structures remains challenging due to the non-trivial feasibility constraints on allocations (see Section 4.2 for detailed discussions). For tractability, we focus primarily on two canonical forms of utility functions. The first case considers utility functions that are multiplicatively separable. Specifically, each agent i's utility for receiving allocation x_i while producing signal s_i is given by $u_i(x_i, s_i, \theta_i) = v_i(x_i, \theta_i) - \frac{h_i(s_i)}{g_i(\theta_i)}$. In this special case, the problem reduces to the standard money-burning problem, and the optimal mechanism can be characterized using tools from the literature (Hartline and Roughgarden, 2008; Akbarpour et al., 2024).

A more interesting case involves utility functions that capture costly falsification. That is, each agent i's utility for receiving allocation x_i while producing signal s_i is given by $u_i(x_i, s_i, \theta_i) = x_i - \eta \cdot (s_i - \theta_i)^+$ where η captures the agent's ability to falsify the signal. For this type of falsification cost, the optimal mechanisms can still be quite complex, especially with arbitrary type distribution primitives. Nevertheless, we show that the optimal mechanism in symmetric environments can be illustrated as follows. The type space of each agent is partitioned into disjoint intervals, with each interval belonging to one of the following three categories: (1) the *no-tension* interval, (2) the *no-effort* interval, and (3) the *efficient* interval. Types in the no-effort interval are pooled together and assigned a coarse ranking, while types in the other two intervals maintain their strict order. Items are then allocated efficiently according to this ranking. Under this mechanism, any agent with a type in the first two intervals has no incentive to exert effort, while any agent with a type in the third interval exerts positive effort.

Intuitively, the principal always aims to allocate items efficiently, provided that this does not incentivize unnecessary effort. This is achievable in the no-tension interval, making it optimal in that range. However, if allocating the item efficiently encourages an agent to misreport their type as higher, it induces costly effort. In such cases, depending on the principal's weights for agent utilities in the objective function, it may remain optimal

to allocate efficiently while inducing effort (as in the efficient interval). Alternatively, the principal may randomize the allocation so that the marginal benefit of exerting effort equals the marginal cost, completely eliminating the agents' incentives to produce costly signals (this occurs in the no-effort interval). Interestingly, our result shows that these extreme treatments are sufficient to achieve the optimal outcome, and partial randomizations that induce smaller but strictly positive signaling costs are never required in the optimal non-coordination mechanism. Finally, note that although these three categories of intervals account for all possible outcomes in optimal mechanisms, each category may consist of countably many intervals. The order in which these intervals appear also depends on the shape of the type distribution and other primitives, such as the number of available items.

Our paper provides a sharper characterization for large markets with a large number of agents. We consider two cases. In the first case, the principal has only one item to allocate.[5] This assumption is relevant for applications such as prestigious scholarships and research funding, where the number of winners is small relative to the number of applicants. In this case, we show that as the number of agents increases, the structure of the optimal mechanism converges to that of a winner-takes-all (WTA) contest (equivalently, there exists an efficient interval that converges to the entire type space) However, the principal's expected payoff under the WTA contest does not converge to the optimal payoff in the large-market limit. The intuition is that, for any sufficiently large but finite number of agents, the WTA contest puts excessive pressure on agents with types close to the highest in the support in order to win the item, leading to significant costly efforts from those types. In contrast, by introducing a small but non-empty no-effort interval in the optimal mechanism for these high types, which randomizes the allocation, the principal can significantly reduce the cost of signaling with only a negligible loss in efficiency, as all these high types are almost equally qualified for receiving the item.[6]

In the second case, we consider a scenario where the number of items grows proportionally with the number of agents. This model is more suited for applications such as college admissions and government benefit programs like public housing or food subsidies, where a significant fraction of the agents receive an item. In this case, if the items were allocated efficiently, all agents with types above a certain cutoff would receive an item. We find that in the optimal mechanism, the principal randomizes the allocation for types around the cutoff (i.e., the "middle" types) to eliminate their incentives to exert costly effort. This increases the expected utilities of both the "middle" types and slightly higher types (those whose

[5]The idea extends easily to a constant number of items.

[6]Our result on payoff non-convergence contrasts with findings in the large contest literature, where a continuum of agents is used to approximate a finite market (e.g., Olszewski and Siegel, 2016).

types are above the "middle" types, but not significantly above the cutoff), at the cost of only slightly reduces matching efficiency for the "middle" types. Our finding is reminiscent of Director's law, which suggests that public programs are often designed primarily to benefit the middle classes. It also aligns with the empirical results of Krishna et al. (2022), who use data from Turkey to show that randomly allocating college seats to low-scoring students reduces stress for all students.[7]

1.1 Related Work

Our paper is closely related to the literature on costly signaling and money burning (e.g., Chakravarty and Kaplan, 2006; Hartline and Roughgarden, 2008; Condorelli, 2012; Finkelstein and Notowidigdo, 2019; Akbarpour et al., 2024; Yang et al., 2024). Prior work often relies on a linear structure of the payoff functions, either in transfers (e.g., Hartline and Roughgarden, 2008) or in costly signals or costly ordeals (e.g., Akbarpour et al., 2024; Yang et al., 2024). In contrast, our paper considers a general cost structure and shows that the standard payoff equivalence fails, and that the implementation of allocation rules and the (non-)coordination of costly signaling choices play a crucial role in reducing the cost of screening. This observation is absent under restrictive linear cost structures, and our paper provides a tractable framework for analyzing these novel economic effects in screening problems.

Our paper is also conceptually related to the literature on falsification and costly lying (e.g., Green and Laffont, 1986; Hardt et al., 2016; Perez-Richet and Skreta, 2022, 2023, 2024). Green and Laffont (1986) focus on auction settings where the lying costs are either zero or infinite. The cost structures in other papers are more general, while the underlying design problems differ more significantly. For instance, Hardt et al. (2016) focus on classification problems in a machine learning context, whereas Perez-Richet and Skreta (2022) concentrate on test design problems. The paper most closely related to ours is the contemporary work by Perez-Richet and Skreta (2024). They focus on a single-agent model and show that the optimal mechanism is score-based when the falsification cost is linear or quadratic.[8] Their notion of score-based shares similar features with our non-coordination mechanism, as both require deterministic signal recommendations. However, the principal's objective in their paper differs from ours, and insights regarding whether coordination is beneficial

[7] Our prediction aligns with their findings, as the "low" types in our model correspond to students who drop out due to having no chance at college admissions. The random allocation for low-scoring students mirrors the randomization for "middle" types in our model.

[8] They also show that for general cost structures, the optimal mechanism is score-based if it can be implemented as a deterministic mechanism. However, whether a deterministic mechanism is optimal remains unclear except in special cases, such as linear or quadratic costs.

under costly signaling cannot be obtained from a single-agent model.[9]

The costly signaling aspect of our paper resembles the literature on signaling (e.g., Spence, 1973) and gaming (e.g., Frankel and Kartik, 2019; Ball, 2024), which studies manipulative behaviors in signaling games. The main distinction is that in signaling games, there is a competitive market that pays each agent a wage corresponding to their estimated type, whereas we adopt a mechanism design perspective in these markets and characterize the optimal mechanisms for allocating resources using costly signals as screening devices.

In our paper, the principal's main objective is to design a mechanism that implements a specific allocation rule in the least costly way. Our results extend naturally if the principal has a preference over allocations and seeks to maximize the weighted averages of the payoffs from allocations and the agents' utilities. If the weights on the agents' utilities are zero, this reduces to the classical screening problem of maximizing the principal's expected payoff (e.g., Laffont and Martimort, 2009). Conversely, if the weights on the payoffs from allocations are zero, this reduces the problem to mechanisms with pure redistributive concerns (e.g., Dworczak et al., 2021; Akbarpour et al., 2024). Our results imply that in all these environments, under general conditions on cost functions, mechanisms without coordination remain weakly optimal, which greatly simplifies the optimal design problems since payoff equivalence fails due to the absence of quasilinear transfers.

The non-coordination mechanisms in our paper can be implemented as a coarse ranking contest, which can be viewed as a generalization of classic contest formats adopted in the literature that are based on strict rankings, such as all-pay contests (Liu and Lu, 2017), Lazear–Rosen contests (Lazear and Rosen, 1981), and Tullock contests (Fu and Wu, 2019). The optimality of contests among general mechanisms has also been established in Zhang (2024), which relies on payoff equivalence results. The main differences are that payoff equivalence fails in our setting and that they focus on effort-maximizing mechanisms, whereas we focus on reducing effort. Additionally, our analysis of optimal contests in large markets sharply contrasts with the results in large contests by Olszewski and Siegel (2020), as our paper provides an environment where the optimal payoff in a finite large contest cannot be approximated by the limiting case.

[9]Additionally, from a technical perspective, Perez-Richet and Skreta (2024) derive their results by explicitly characterizing the optimal single-agent mechanism. In contrast, we provide general sufficient conditions on the cost structures, including linear or quadratic costs as special cases, showing that the optimal multi-agent mechanism does not coordinate on costly signaling choices, even in settings where characterizing the optimal may be intractable.

2 Model

The principal (she) wishes to allocate k identical items to $n > k$ heterogeneous agents (he). An allocation $\boldsymbol{x} = (x_i)_{i=1}^n$ is a vector of probabilities such that $0 \leq x_i \leq 1$ for each i, and $\sum_{i=1}^n x_i \leq k$.[10] Let $X \subseteq [0,1]^n$ be the space of feasible allocations.

Agents' information and payoffs. Each agent i has a private type θ_i drawn independently from a publicly known distribution F_i supported on $\Theta_i = [\underline{\theta}_i, \bar{\theta}_i] \subseteq \mathbb{R}_+$. We denote the distribution over type profile $\boldsymbol{\theta}$ as $\boldsymbol{F}$. The principal cannot directly observe the agents' private types, but she can base her allocation decision on their public signals. Specifically, each agent i can generate a costly public signal $s_i \in S_i = \mathbb{R}_+$ and the utility of the agent is $u_i(x_i, s_i, \theta_i)$ for receiving an item with probability $x_i \in [0,1]$ when generating signal s_i. We assume that the agent's utility satisfies the von Neumann-Morgenstern expected utility representation.

Assumption 1 (monotonicity)**.** *For any agent i, the utility function u_i is continuous in all of its coordinates, bounded in value, strictly increasing in x_i, weakly decreasing in s_i and weakly increasing in θ_i.*

The interpretation of the monotonicity of the utility function is that the agent always strictly prefers a higher probability of allocation. Moreover, the signals are costly for the agents to generate, and the cost of signal is weakly increasing in the signal realization. The increase in utility from the increase in private type may come from two parts. First, the agent's value for the item may be weakly higher for a higher type. Second, the cost of generating a higher signal may be weakly lower for a higher type. In the main result of this paper, we do not distinguish these two forces in the utility function.

We assume that each agent's utility function satisfies the weak single-crossing property as in Milgrom and Shannon (1994).

Assumption 2 (weak single-crossing property (wSCP))**.** *The utility function u satisfies* weak single-crossing property (wSCP) *if for any types $\theta' > \theta$, any signals $s' > s$ and any allocation probabilities $x' > x$,*

1. *$u(x', s', \theta) - u(x, s, \theta) > 0$ implies that $u(x', s', \theta') - u(x, s, \theta') > 0$;*
2. *$u(x', s', \theta) - u(x, s, \theta) \geq 0$ implies that $u(x', s', \theta') - u(x, s, \theta') \geq 0$.*

[10] Our paper focuses on the interpretation where items are indivisible and x_i represents allocation probabilities. All the results extend naturally by considering divisible items and interpreting x_i as fractional allocations.

Given any distribution G over allocations and signal realizations, the utility of the agent i with type θ_i given distribution G is denoted by $u_i(G, \theta_i)$. We assume that the agent has expected utility representation. Note that our single-crossing condition is only imposed on deterministic signal realizations. In general, single-crossing condition fails when randomization is taken into consideration. The utility function illustrated in the introduction is an example where Assumption 2 holds while the single-crossing condition fails for randomized signals.

General Mechanisms. We focus on direct mechanisms, in which the principal can elicit the types from all the agents, and can make signal recommendations based on the aggregated report before the agents choose their signals. Formally, the timeline for a direct mechanism is as follows:

(1) The principal commits to a signal recommendation policy $\tilde{s} : \Theta \to \Delta(S)$ and an allocation rule $y : \Theta \times S \to X$.

(2) Each agent i reports type $\hat{\theta}_i$ to the principal and receives signal recommendation $\tilde{s}_i(\hat{\boldsymbol{\theta}})$. Then each agent i chooses signal $s'_i \in \{\tilde{s}_i(\hat{\boldsymbol{\theta}}), 0\}$.[11]

(3) The principal observes the signal profile $\boldsymbol{s}'$, and each agent i receives an item with probability $y_i(\hat{\boldsymbol{\theta}}, \boldsymbol{s}')$.

Given any direct mechanism, the *interim allocation* of agent i with private type θ_i is denoted as

$$Q_i(\theta_i) = \mathbf{E}_{\boldsymbol{\theta}_{-i}}\left[\mathbf{E}_{s_i \sim \tilde{s}_i(\theta_i, \boldsymbol{\theta}_{-i})}[y_i(\theta_i, \boldsymbol{\theta}_{-i}, s_i, \tilde{\boldsymbol{s}}_{-i}(\theta_i, \boldsymbol{\theta}_{-i}))]\right]$$

and the *interim utility* is denoted as

$$U_i(\theta_i) = \mathbf{E}_{\boldsymbol{\theta}_{-i}}\left[\mathbf{E}_{s_i \sim \tilde{s}_i(\theta_i, \boldsymbol{\theta}_{-i})}[u_i(y_i(\theta_i, \boldsymbol{\theta}_{-i}, s_i, \tilde{\boldsymbol{s}}_{-i}(\theta_i, \boldsymbol{\theta}_{-i})), s_i, \theta_i)]\right].$$

A direct mechanism is incentive compatible (IC) if each agent has a weakly higher utility for truthful reporting, i.e., for any agent i and any types θ_i, θ'_i, Each agent has a weak incentive

[11] We impose interim participation constraints for agents, i.e., agents can choose to walk away after seeing the signal recommendation. Our formulation is without loss of generality under the interim participation constraints since the principal can partially enforce the recommendation by allocating no items to any agent who chooses a signal different from the recommended one. Here each agent essentially has two choices: following the recommendation ($s'_i = \tilde{s}_i(\hat{\boldsymbol{\theta}})$) or opting out ($s'_i = 0$). An alternative formulation is to only consider the ex ante participation constraint where each agent only has the option to opt out at the beginning of the mechanism, and they are forced to follow the principal's recommendation after participation. This distinction is not crucial for our analysis, and the main results of the paper hold for both cases.

to truthfully report his own type and follow the signal recommendation:

$$U_i(\theta_i) \geq \mathbf{E}_{\boldsymbol{\theta}_{-i}}\left[\mathbf{E}_{s_i\sim\tilde{\boldsymbol{s}}(\theta_i',\boldsymbol{\theta}_{-i})}\left[\max\{0, u_i(y_i(\theta_i',\boldsymbol{\theta}_{-i}, s_i, \tilde{\boldsymbol{s}}_{-i}(\theta_i',\boldsymbol{\theta}_{-i})), s_i, \theta_i)\}\right]\right]. \tag{IC}$$

By the revelation principle (Myerson, 1982), it is without loss to focus on direct mechanisms that are incentive compatible. For the rest of the paper, we refer to those mechanisms as direct mechanisms without explicitly mentioning incentive compatibility when there is no ambiguity.

Non-coordination Mechanisms. We now introduce the non-coordination mechanisms. In non-coordination mechanisms, the principal does not provide signal recommendations to the agents based on the aggregated reported types. Instead, the principal commits to a signal-based allocation rule $\boldsymbol{x} : S \to X$, which maps each realized signal profile to a (randomized) allocation. After the signal-based allocation rule is announced, each agent chooses his signal strategy to maximize their expected utilities. It is worth emphasizing that the effort choice of a given agent is not correlated with the realized types or effort choices of any other agents, i.e., there is no coordination.

We denote the strategy of each agent i by $s_i : \Theta_i \to \Delta(S_i)$. Given the strategy profile $\boldsymbol{s} = (s_i)_{i=1}^n$ and the realized signal profile $\boldsymbol{s}(\boldsymbol{\theta}) = (s_i(\theta_i))_{i=1}^n$, the item is distributed according to the allocation rule $\boldsymbol{x}(\boldsymbol{s}(\boldsymbol{\theta}))$. Formally, the timeline is as follows:

(1) The principal commits to a signal-based allocation rule $\boldsymbol{x} : S \to X$.

(2) Each agent i, with type θ_i, generates a costly signal $s_i(\theta_i)$.

(3) The principal observes the signal profile s, and each agent i receives an item with probability $x_i(\boldsymbol{s}(\boldsymbol{\theta}))$.

Here the signal-based allocation rule defines a game for the agents. We provide a direct mechanism implementation for non-coordination mechanisms when this game has a pure strategy equilibrium. The timeline of a direct non-coordination mechanism is illustrated as follows.

(1) The principal commits to a signal recommendation policy $\tilde{s}_i : \Theta_i \to S_i$ for each agent i and allocation rule $y : S \to X$.

(2) Each agent i reports type $\hat{\theta}_i$ to the principal and receives signal recommendation $\tilde{s}_i(\hat{\theta}_i)$. Then each agent i chooses signal $s_i' \in \{\tilde{s}_i(\hat{\theta}_i), 0\}$.

(3) The principal observes the signal profile $\boldsymbol{s}'$, and each agent i receives an item with probability $y_i(\boldsymbol{s}')$.

The main difference between a general direct mechanism and a direct non-coordination mechanism is in step (1), where the signal recommendation for each agent in the direct non-coordination mechanism only depends on each agent's own reported type, not the entire type profile. This restriction prevent the coordination of the choices of costly signals. Additionally, the allocation function in the direct non-coordination mechanism only depends on the realized signal profile, not the reported types.

Applying a similar argument as the revelation principle, it is straightforward to show that any non-coordination mechanism with a pure strategy equilibrium can be implemented as a direct non-coordination mechanism, and vice versa. Similar result is also observed in the contemporary work by Perez-Richet and Skreta (2024) in the special case of single-agent environments.

Interim approach. The main analyses of the paper highly rely on the interim approach where we focus on interim allocation rule and interim utilities instead of ex post ones. In particular, we are interested in understanding whether a given interim allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ is *feasible* and *implementable*.

Definition 1. *Given any interim allocation rule profile $\boldsymbol{Q} = (Q_i)_{i=1}^n$ where $Q_i : \Theta_i \to [0,1]$, $\boldsymbol{Q}$ is* feasible *if there exists an ex-post allocation rule $\boldsymbol{q}$ such that $Q_i(\theta_i) = \mathbf{E}_{\theta_{-i}}[q_i(\theta_i, \theta_{-i})]$ for any agent i and type θ_i. We say an interim allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ is* feasible *if $\boldsymbol{Q}$ is feasible.*

In general, not all interim allocation rules are feasible. The set of feasible interim allocations has been characterized in Border (1991); Che et al. (2013). See Lemma 3 in Appendix A.2 for more details.

Definition 2. *An interim allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ is* implementable by a direct mechanism *if there exists a direct mechanism with signal recommendation policy $\tilde{\boldsymbol{s}}$ and allocation rule $\boldsymbol{y}$ such that for any agent i and type θ_i, the consistency condition holds. I.e.,*

$$\begin{aligned} Q_i(\theta_i) &= \mathbf{E}_{\boldsymbol{\theta}_{-i}}\left[\mathbf{E}_{s_i \sim \tilde{s}_i(\theta_i, \boldsymbol{\theta}_{-i})}[y_i(\theta_i, \boldsymbol{\theta}_{-i}, s_i, \tilde{\boldsymbol{s}}_{-i}(\theta_i, \boldsymbol{\theta}_{-i}))]\right], \qquad \text{(consistency)} \\ U_i(\theta_i) &= \mathbf{E}_{\boldsymbol{\theta}_{-i}}\left[\mathbf{E}_{s_i \sim \tilde{s}_i(\theta_i, \boldsymbol{\theta}_{-i})}[u_i(y_i(\theta_i, \boldsymbol{\theta}_{-i}, s_i, \tilde{\boldsymbol{s}}_{-i}(\theta_i \boldsymbol{\theta}_{-i})), s_i, \theta_i)]\right]. \end{aligned}$$

We say an interim allocation $\boldsymbol{Q}$ is implementable *if there exists an interim utility $\boldsymbol{U}$ such that $(\boldsymbol{Q}, \boldsymbol{U})$ is implementable by a direct mechanism.*

In addition, an interim allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ is *implementable by a direct non-coordination mechanism* if the mechanism specified in Definition 2 is a direct non-coordination mechanism.

Principal's payoff. The principal cares the utilities of the agents since the costly signals are wasteful from a social perspective. Specifically, given any feasible and implementable interim allocation $\boldsymbol{Q}$, the principal want to design a mechanism that implements $\boldsymbol{Q}$ while maximizing the agents' weighted utilities. Moreover, we allow the weighting function to depend on the agents' private types as in the literature with redistribution purposes (e.g., Dworczak et al., 2021; Akbarpour et al., 2024). Specifically, there exists a weighting function $w_i(\theta_i) \in [0,1]$ such that the principal's objective is to maximize

$$\sum_{i\in[n]} \mathbf{E}_{\theta_i\sim F_i}[w_i(\theta_i)\cdot U_i(\theta_i)]\,,$$

subject to the constraint of implementing $\boldsymbol{Q}$.

One possible rationale for this objective of the principal is that there is a social benefit of allocating the items to particular types of the agents, and the objective of the principal is to maximizes the weighted average between the social benefits and the agents utilities. Specifically, let W_i be the weighted social benefit function for each agent i and the objective of the principal is to maximize

$$\sum_{i\in[n]} \mathbf{E}_{\theta_i\sim F_i}[W_i(Q_i(\theta_i),\theta_i) + w_i(\theta_i)\cdot U_i(\theta_i)]\,.$$

Similar objective functions have also been studied in the falsification literature (Perez-Richet and Skreta, 2022, 2024). This objective is useful for characterizing the optimal mechanisms. However, the optimality of non-coordination mechanisms (Theorem 1) does not depend on this specific formulation of the principal's objective function.

3 Optimality of Non-coordination Mechanisms

In this section, we provide sufficient conditions that the optimal mechanism that implements any given monotone allocation rule is a non-coordination mechanism.

Definition 3. *Given any distribution G over allocations and signals, letting x_G be the marginal distribution over allocations, the certainty equivalent signal $\Sigma(G,\theta)$ of the distribution G for type θ is defined as the signal such that $u(x_G,\Sigma(G,\theta),\theta) = u(G,\theta)$.*

Assumption 3 (Monotone Certainty Equivalence)**.** *Each agent has weakly monotonicity in certainty equivalence for costly signaling. That is, for any distribution G over allocations and signals, the certainty equivalence $\Sigma(G,\theta_i)$ is weakly increasing in θ_i for all agents i.*

One canonical form of utility function is

$$u_i(G, \theta_i) = \mathbf{E}_{(x,s)\sim G}[v_i(x, \theta_i) - c_i(s, \theta_i)]$$

where V_i is agent i's value for allocation and c_i is the cost of signaling. In our model, a higher type is interpreted as an agent with a higher need for the item, and hence lower wealth. Naturally, a lower wealth agent has a stronger preference for certain outcomes and hence a higher certainty equivalent signal. The monotone certainty equivalence assumption implies when the cost function is convex (which is not required for our general result), the agent's preferences for utility loss from costly signaling resembles weakly decreasing absolute risk aversion (DARA) utilities, which is commonly adopted in economic models (e.g., Arrow, 1965; Pratt, 1976).

Theorem 1 (Optimality of Non-coordination)**.** *Under Assumption 1, 2 and 3, for any interim allocation–utility pair* $(\boldsymbol{Q}, \boldsymbol{U})$ *that is implementable by a direct mechanism where* Q_i *is weakly increasing for all* i*, there exists a interim allocation–utility pair* $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ *such that* $\boldsymbol{Q}^\dagger = \boldsymbol{Q}$*,*

$$U_i^\dagger(\theta_i) \geq U_i(\theta_i), \quad \forall i \leq n, \theta_i \in \Theta_i,$$

and $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ *is implementable by a direct non-coordination mechanism.*

Note that in the introduction, we provided an example where the inequality is strict for all the agents.

Proof. To simplify the exposition, we only present the proof when the type space is finite.[12] We denote the type space as $\Theta_i = \{\hat{\theta}_i^{(0)}, \ldots, \hat{\theta}_i^{(m)}\}$, with $\hat{\theta}_i^{(0)} < \cdots < \hat{\theta}_i^{(m)}$. The proof is constructive. Specifically, in **Step 1**, we will replace the stochastic signal recommendation given in the direct mechanism with a deterministic signal recommendation, which is the certainty equivalent signal as defined in Definition 3. Such a replacement will respect the local (upward) IC constraints but might fail the local (downward) IC constraints. Therefore, in **Step 2**, we will modify the deterministic signal recommendation such that local (downward) IC constraints are also respected. By the wSCP of each agent's utility function (Assumption 2), it is easy to verify that the candidate mechanism we have constructed outperforms

[12]For continuous type space, we can show that for any $\epsilon > 0$, by carefully discretizing the type space and applying the same argument, there exists an allocation–utility pair $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ that is implementable by a direct non-coordination mechanism and $\mathrm{Obj}_\alpha(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger) \geq \mathrm{Obj}_\alpha(\boldsymbol{Q}, \boldsymbol{U}) - \epsilon$. The theorem holds by taking $\epsilon \to 0$ and the observation that the set of payoffs obtainable by non-coordination mechanisms is compact.

the given direct mechanism in terms of designer's objective. Moreover, it is implemented as a direct non-coordination mechanism.

Step 1: *Certainty equivalent signal.* The expected utility of each agent i under any direct mechanism is pinned down by the associated random signal recommendation and the expected allocation. Since type space is discrete, we define the following simplified notations. Fix the direct mechanism that delivers interim allocation $\boldsymbol{Q}$ and interim utility $\boldsymbol{U}$, let $G_i^{(k)}$ be the distribution over signals recommended to agent i who reports type $\hat{\theta}_i^{(k)}$, and let $Q_i^{(k)}$ be the corresponding expected allocation given to agent i who reports type $\hat{\theta}_i^{(k)}$.

Denote the certainty equivalent signal of $G_i^{(k)}$ for type $\hat{\theta}_i^{(j)}$ as

$$s_i^{(k,j)} \triangleq \Sigma(G_i^{(k)}, \hat{\theta}_i^{(j)}), \quad 0 \leq k, j \leq m.$$

Moreover, let $s_i^{(k)} := s_i^{(k,k)}$, for all $0 \leq k \leq m$.

First we replace the randomized signal recommendation $G_i^{(k)}$ with its certainty equivalence $s_i^{(k)}$ for each type $\hat{\theta}_i^{(k)}$ while maintaining the same allocation rule. After such a replacement, Assumption 3 guarantees that when type $\hat{\theta}_i^{(k-1)}$ deviates to the adjacent type $\hat{\theta}_i^{(k)}$, he is weakly worse under the certainty equivalence $s_i^{(k)}$ compared to under the original randomized signal recommendation. Specifically, agent type $\hat{\theta}_i^{(k-1)}$'s utility for misreporting as type $\hat{\theta}_i^{(k)}$ in the direct mechanism is at least[13]

$$u_i\left(Q_i^{(k)}, G_i^{(k)}, \hat{\theta}_i^{(k-1)}\right) = u_i\left(Q_i^{(k)}, s_i^{(k,k-1)}, \hat{\theta}_i^{(k-1)}\right) \geq u_i\left(Q_i^{(k)}, s_i^{(k)}, \hat{\theta}_i^{(k-1)}\right),$$

where the equality follows from the definition of certainty equivalent signal, and the inequality holds since (1) $s_i^{(k,j)}$ is increasing in j by Assumption 3; and (2) the utility function is weakly increasing in the signal. Since the utility for misreporting as type $\hat{\theta}_i^{(k)}$ is weakly lower and the utility for truthfully reporting as type $\hat{\theta}_i^{(k-1)}$ remains unchanged, the agent has no incentives for local upward deviation.

Step 2: *Downward adjustments for global incentive.* Step 1 has shown that by replacing the randomized recommendation with its certainty equivalence, the local upward incentive constraints hold. However, the local downward incentive constraints may be violated. In this step, we decrease the recommended signals so that the local (downward) incentive compatible constraints can be restored whenever they are violated under the deterministic

[13]Notice that the agent can potentially adopt double deviation strategies in the direct mechanism based on the signal realizations, i.e., after misreporting, the agent can walk away based on the signal realization.

signal constructed in step 1. The single-crossing property will then show that the global incentives is implied by the local incentives.

First note that since $Q_i^{(k)}$ is weakly increasing in k and the local upward incentive constraints hold, the monotonicity of the utility function implies that the certainty equivalent signals $s_i^{(k)}$ are also weakly increasing in k.

We proceed by induction. Let $\hat{s}_i^{(0)} = s_i^{(0)}$. For any $1 \leq k \leq m$, let $\hat{s}_i^{(k)} = s_i^{(k)}$ if the local downward incentive constraint holds for type $\hat{\theta}_i^{(k)}$. If the local downward incentive constraint is violated, the monotonicity and continuity of the cost function implies that there exists a signal $\hat{s}_i^{(k)} \in [\hat{s}_i^{(k-1)}, s_i^{(k)}]$ such that the local downward incentive constraint bind for type $\hat{\theta}_i^{(k)}$, i.e.,

$$u_i\left(Q_i^{(k)}, \hat{s}_i^{(k)}, \hat{\theta}_i^{(k)}\right) = u_i\left(Q_i^{(k-1)}, \hat{s}_i^{(k-1)}, \hat{\theta}_i^{(k)}\right).$$

Notice that if $Q_i^{(k)} > Q_i^{(k-1)}$, then $\hat{s}_i^{(k)} > \hat{s}_i^{(k-1)}$. In this case, even after the modification of the signal recommendation, the local upward incentive constraint holds due to the single-crossing assumption (Assumption 2). If $Q_i^{(k)} = Q_i^{(k-1)}$, then $\hat{s}_i^{(k)} = \hat{s}_i^{(k-1)}$. In this case local upward incentive constraint holds trivially.

Let $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ be the interim allocation–utility pair that is induced by deterministic signal recommendations $\hat{s}_i^{(k)}$. Notice that the recommended signals are adjusted downwards compared to the certainty equivalent signal for the stochastic signal recommendation given in the direct mechanism. As a result, the interim utilities for all types under the newly constructed mechanism that induces $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ weakly improves upon the interim utilities induced by the direct mechanism $(\boldsymbol{Q}, \boldsymbol{U})$, i.e.,

$$U_i^\dagger(\theta_i) \geq U_i(\theta_i), \quad \forall i \leq n, \theta_i \in \Theta_i.$$

Finally, weak single-crossing property in the utility functions (Assumption 2) also guarantees that global incentive constraints hold if the agent is forced to generate the recommended signal when misreporting her type, i.e., when interim participation constraints are not imposed. Note that the deviating utility for not following the recommendation is at most 0 because the recommended signal is deterministic, and the interim utilities of the agents are weakly positive after the modification. Therefore, the participation constraints are satisfied as well. □

3.1 Discussion on Monotone Allocations

In this section, we have focused on mechanisms with monotone allocations. Restricting to monotone interim allocations is a practical assumption because of fairness concerns. In many applications involving government allocating resources, it may be perceived as unfair to provide higher allocations to lower types (see for instance, Gershkov et al., 2022). For example, in college admissions, students with greater talent should have a higher probability of being admitted to a school. In government subsidy programs, people with lower income, or greater financial need, should have a higher probability of receiving subsidies.

In general, by considering direct mechanisms, the principal can implement a broader set of allocation rules. In particular, the principal can implement non-monotone allocations with direct mechanisms. Note that in contrast, with the weak single-crossing condition (Assumption 2), only monotone allocations are implementable given direct non-coordination mechanisms. The main intuition is that given any two deterministic signal recommendations and associated allocations to the agent, the preference over these two recommendations are monotone over types, while the monotonicity in preference fails with randomized recommendations. Here we provide a simple example of implementable non-monotone allocation.

Example 3. *Consider a simplified single-agent setting where the agent's type θ is drawn from a uniform distribution on $[0,2]$, and the agent's utility function is $x-(s-\theta)^+$. Consider two options for the agents. The first option has a deterministic signal recommendation. That is, the agent receives an item with probability $\frac{1}{2}$ if the generated signal is $\frac{2}{3}$. The second option has a random signal recommendation. With probability $\frac{1}{2}$, the agent receives an item with probability $\frac{1}{2}$ if the generated signal is $\frac{1}{2}$, and with the rest probability, the agent receives an item with probability 1 if the generated signal is 2.*

It is easy to verify that if the agent has type $\theta \in [\frac{1}{3}, \frac{3}{2}]$, the agent prefers the deterministic signal recommendation and if the agent has type $\theta \in [0, \frac{1}{3}] \cup [\frac{3}{2}, 2]$, the agent prefers the randomized signal recommendation. This is an implementable non-monotone allocation.

In other applications without the exogenous requirement on monotone allocations, we can still show that restricting attention to monotone allocations is without loss of optimality for broad classes of signaling costs and objective functions. See Appendix B for detailed discussions.

3.2 Information Leakage

The main intuition behind the optimality of non-coordination mechanisms is that any form of coordinating leaks information to the agents. This in effect produces a random recom-

mendation to the agent based on the leaked information. Under the assumption of weak monotonicity in certainty equivalence for the cost functions, such random allocation can be improved by using deterministic recommendations, where the agent is ignorant of any detailed information regarding the environments, and make the effort choices solely based on his private type and the prior distributional knowledge.

This intuition can be easily extended to richer models beyond our current framework. For example, in markets where there is uncertainty in supply and demand (e.g., Kilian, 2009), i.e., the number of resources and number of participants are drawn from a distribution where the realization is unobserved to the agents, our results indicate that in such environments, it is optimal for the designer to disclose no information regarding either the supply or the demand to the agents before asking the agents to make their effort choices. Similarly, when the resources to be allocated are differentiated in qualities and there is uncertainty in qualities (e.g., Akerlof, 1970), it is also optimal to run non-coordination mechanisms without revealing information about the true qualities.

3.3 Implementation as Coarse Ranking Contests

We show that in symmetric environments, any symmetric direct non-coordination mechanism has an indirect implementation as the coarse ranking contest. Specifically, in coarse ranking contest, there may exist segments of signals that are pooled and assigned the same rank. The items are allocated to the k agents with the highest coarse rankings, with ties broken uniformly at random. This generalizes the concept of a contest as commonly presented in the literature.

Definition 4 (coarse ranking)**.** *Given any countable set of disjoint open intervals* $\{(\underline{s}^{(j)}, \bar{s}^{(j)})\}_{j=1}^{\infty}$ *whose union is a subset of the type space, the* coarse ranking *of agent* i *under the signal profile* $\boldsymbol{s} = (s_1, \ldots, s_n)$ *is*

$$r_i(\boldsymbol{s}) = \left| \left\{ i' \neq i, 1 \leq i' \leq n : s_{i'} > \bar{s}^{(j)} \right\} \right|,$$

and the number of ties for agent i *is*

$$z_i(\boldsymbol{s}) = \left| \left\{ i' \neq i, 1 \leq i' \leq n : \bar{s}^{(j')} = \bar{s}^{(j)} \right\} \right| + 1.$$

Here, for any signal s_i*,* j *is the index such that* $s_i \in (\underline{s}^{(j)}, \bar{s}^{(j)})$*, if such a* j *exists (i.e., if* s_i *falls into one of the intervals defined). If no such* j *exists (i.e., if* s_i *lies outside all the intervals defined), then (slightly overloading the notation) we let* $\bar{s}^{(j)} = \underline{s}^{(j)} = s_i$*. We also call the pair of functions* $(\mathbf{r}, \mathbf{z})$ *a* coarse ranking*.*

Intuitively, each interval in the set $\{(\underline{s}^{(j)}, \bar{s}^{(j)})\}_{j=1}^{\infty}$ specifies a region of signals that are

pooled and assigned the same (coarse) ranking. Outside the closure of the union of these intervals, signals are ranked strictly. In the special case where $\{(\underline{s}^{(j)}, \bar{s}^{(j)})\}_{j=1}^{\infty}$ is empty, the definition of a coarse ranking coincides with the usual definition of a strict ranking, where the agents' ranks are given by the order of their signals in the given signal profile.

Our generalization of strict rankings to coarse rankings leads to a larger class of contest rules, defined as follows. For any coarse ranking $(\mathbf{r}, \mathbf{z})$ and any agent i, the induced (coarse ranking) contest rule is

$$\tilde{x}_i(\boldsymbol{s}; \mathbf{r}, \mathbf{z}) = \begin{cases} 1, & k \geq r_i(\boldsymbol{s}) + z_i(\boldsymbol{s}), \\ \frac{k - r_i(\boldsymbol{s})}{z_i(\boldsymbol{s})}, & k \in (r_i(\boldsymbol{s}), r_i(\boldsymbol{s}) + z_i(\boldsymbol{s})), \\ 0, & k \leq r_i(\boldsymbol{s}). \end{cases}$$

Definition 5 (coarse ranking contest rule)**.** *A mapping profile $\boldsymbol{x} : S \to X$ is a* coarse ranking contest rule *if there exists a coarse ranking $(\mathbf{r}, \mathbf{z})$ such that $x_i(\boldsymbol{s}) = \tilde{x}_i(\boldsymbol{s}; \mathbf{r}, \mathbf{z})$ for each i.*

The class of coarse ranking contest rules, which is a subset of the class of all mappings from the signal space to the allocation space, also includes the class of contest rules that allocate items (prizes) to agents based on the strict ranking of their signals such as all-pay contests (Fu and Wu, 2019). The proof of the proposition is provided in Appendix A.1.

Proposition 1. *In symmetric environments, any symmetric interim allocation–utility pair (Q, U) that is implementable by a direct non-coordination mechanism has an indirect implementation that is a randomization over the coarse ranking contests.*

3.4 Suboptimality of Non-coordination

In this section, we have shown that non-coordination mechanisms are optimal if the utility function satisfies weak monotonicity in certainty equivalence for costly signaling. In this part, we will provide a partial converse showing that direct mechanisms can strictly outperform non-coordination mechanisms when the assumption is violated.

Assumption 4. *The utility function satisfies strictly decreasing in certainty equivalence if for any non-degenerate distribution G over signals, the certainty equivalence $\Sigma(G, \theta)$ is strictly decreasing in θ.*

Assumption 5. *For any agent i, the utility function u_i is continuous in all of its coordinates, bounded in value, strictly increasing in x_i and θ_i, and strictly decreasing in s_i.*

Proposition 2 (Sub-optimality of non-coordination)**.** *For any utility function that satisfies Assumption 2, 4 and 5, there exists a distribution $\boldsymbol{F}$ such that for any $(\boldsymbol{Q}, \boldsymbol{U})$ with a strictly increasing $\boldsymbol{Q}$ that is implemented by a non-coordination mechanism, there exists another interim utility $\boldsymbol{U}^\dagger$ such that $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ is implemented by a direct mechanism and outperforms it, i.e., $U_i^\dagger(\theta_i) \geq U_i(\theta_i)$ for all agents i and type θ_i, and there exists a type such the inequality is strict.*

Proof. Given any utility function of the agents, consider a type distribution where F_i has binary support for all agent i, denoted by $\{\underline{\theta}_i, \bar{\theta}_i\}$ where $\underline{\theta}_i < \bar{\theta}_i$. Given any strictly increasing interim allocation Q_i, let $\underline{s}_i < \bar{s}_i$ be the recommended signals for agents with types $\underline{\theta}_i$ respectively $\bar{\theta}_i$ in the non-coordination mechanism. It is easy to verify that in the optimal non-coordination mechanism, $\underline{s}_i = 0$.

Since the utility function is continuous and strictly increasing in s and $\bar{s}_i$ is strictly positive, there exists a non-degenerate distribution G_i over signal recommendations such that agent with type $\bar{s}_i$ is indifferent between signal recommendation G_i and deterministic signal recommendation $\bar{s}_i$. Moreover, by Assumption 4, the utility of type $\underline{s}_i$ for misreporting as type $\bar{s}_i$ is strictly lower given G_i compared to deterministic signal recommendation $\bar{s}_i$. Therefore, there exists G_i' that first order stochastically dominates G_i such that the utility of type $\underline{s}_i$ for obtaining signal recommendation G_i' and $\bar{s}_i$ are the same. By offering signal recommendation $\underline{s}_i$ to type $\underline{\theta}_i$ with interim allocation $Q_i(\underline{\theta}_i)$, and offering randomized signal recommendation G_i' to type $\bar{\theta}_i$ with interim allocation $Q_i(\bar{\theta}_i)$ in the general mechanism, the utility of all types weakly improves, and the utility of type $\bar{\theta}_i$ strictly improves. □

4 Characterization of Optimal Mechanism

In this section, we characterize the optimal non-coordination mechanism when the principal optimizes the weighted average between the social welfare and the agents' utilities, i.e.,

$$\sum_{i\in[n]} \mathbf{E}_{\theta_i\sim F_i}[W_i(Q_i(\theta_i), \theta_i) + w_i(\theta_i)\cdot U_i(\theta_i)]\,.$$

Note that characterizing the optimal mechanisms in general even with the restriction to non-coordination mechanism is challenging without any further structural assumption on the agents' utility functions. Therefore, for tractability reasons, in this section, we will focus on utility functions that are additively separable, i.e., for any agent i, there exists a valuation function $v_i(x_i, \theta_i)$ and a cost function $c_i(s_i, \theta_i)$ such that the utility function is

$$u_i(s_i, s_i, \theta_i) = v_i(x_i, \theta_i) - c_i(s_i, \theta_i).$$

Moreover, we will consider two special forms of cost functions. In the first case, the signal is mutiplicatively separable in type and effort, i.e., $c_i(s_i, \theta_i) = \frac{s_i}{\theta_i}$. This is an assumption commonly seen in the signaling literature and the contest design literature (e.g., Spence, 1973; Fang et al., 2020). In the second case, we assume that each agent's signal can be costly inflated upon his type by manipulation effort, i.e., $c_i(s_i, \theta_i) = (s_i - \theta_i)^+$. This assumption is more commonly found in papers dealing with falsification (e.g., Perez-Richet and Skreta, 2022).

Symmetric environment. To simplify the exposition, in the rest of the paper we assume that the agents are ex-ante homogeneous, i.e., $\Theta_i = \Theta = [\underline{\theta}, \bar{\theta}]$ and $F_i = F$ for all i. In addition, we assume that the density function f_i exists for all i, and $f_i(\theta_i) > 0$ for any $\theta_i \in [\underline{\theta}_i, \bar{\theta}_i]$.

4.1 Multiplicatively Separable Cost

We first consider the case where the costs are multiplicatively separable.

Definition 6 (multiplicatively separable cost)**.** *For any agent i, there exists an increasing function $h_i(s_i)$ and a decreasing function $g_i(\theta_i)$ such that $c_i(s_i, \theta_i) = \frac{h_i(s_i)}{g_i(\theta_i)}$ for any $\theta_i \in \Theta_i$ and $s_i \in S_i$.*

With multiplicatively separable cost (Definition 6), the utility of agent with type θ can be represented as

$$u_i(x_i, s_i, \theta_i) = v_i(x_i, \theta_i) - \frac{h_i(s_i)}{g_i(\theta_i)} = \frac{1}{g_i(\theta_i)} \left(v_i(x_i, \theta_i) \cdot g_i(\theta_i) - h_i(s_i) \right).$$

Therefore, the behavior of the agent is equivalent to one with utility function $v_i(x_i, \theta_i) \cdot g_i(\theta_i) - h_i(s_i)$. By viewing $h_i(s_i)$ as the quasilinear transfers in the model, solving the optimal mechanism boils down to the classical single-dimensional screening with quasilinear transfer under single-crossing properties. The characterization of the optimal mechanism is standard in this case, and similar characterizations can be found in Hartline and Roughgarden (2008); Akbarpour et al. (2024).

4.2 Costly Falsification

We consider the case where it is costly for the agents to mimic higher types. The cost for downward deviation is zero. This captures, for example, applications where allocations are based on agents' performance evaluations, and where agents can easily reduce their performance.

Definition 7 (costly falsification)**.** *For any agent i, any $\theta_i \in \Theta_i$ and $s_i \in S_i$, $u_i(x_i, s_i, \theta_i) = x_i - \eta \cdot (s_i - \theta_i)^+$.*

In this case, since the utilities are still separable, one naive approach is to treat the allocation x as the quasilinear transfers in standard screening models, with the costly signals viewed as the allocations. However, this approach does not work immediately since we have a non-trivial feasibility constraint on the allocations (see Lemma 3 for characterizations of the interim feasibility constraints), and this would translate into a non-trivial feasibility constraint on the transfers in the otherwise standard screening model. To resolve this complication, instead of transforming our problem in the way described above, we directly characterize the optimal mechanism using optimal control.

In this section, we focus on mechanisms with monotone allocation rules. This is shown to be without loss of optimality in Appendix B for utility functions that include the one described in Definition 7. Under the restriction of monotone allocations, Theorem 1 implies that it is without loss to focus on non-coordination mechanisms. We characterize the optimal non-coordination mechanism for a simpler objective of

$$\mathrm{Obj}_\alpha(\boldsymbol{Q}, \boldsymbol{U}) = \mathbf{E}_{\boldsymbol{\theta}}\left[\alpha \cdot \sum_i \theta_i \cdot Q_i(\theta_i) + (1-\alpha) \cdot \sum_i U_i(\theta_i)\right]. \tag{1}$$

In this objective function, all agents are treated equally. The parameter $\alpha \in [0,1]$ captures the relative weights between the allocation efficiency and the agents' utilities.

Symmetric mechanism. Note that finding the optimal mechanism is not a convex program. This is because, even though the objective function is linear, the incentive constraints (IC) are not convex with the cost function defined in Definition 7. Specifically, a convex combination of two allocation–utility pairs may violate the (IC) constraints. Nonetheless, as we show in the following lemma, the optimal non-coordination mechanism for this non-convex optimization problem is always symmetric in symmetric environments.

Lemma 1. *The optimal non-coordination mechanism is symmetric for any $\alpha \in [0,1]$.*

If we restrict our attention to symmetric mechanism, the problem of designing the optimal mechanism reduces to the single-agent optimization problem for any particular agent i. When there is no ambiguity, we omit the subscript i from the notation for this single-agent problem; we use the interim allocation rule Q and the utility function U for a single agent to refer to the interim allocation profile and the interim utility profile, respectively. Let $Q_{\mathrm{E}}(\theta)$ be the interim allocation rule maximizing matching efficiency, i.e., the efficient allocation rule. The optimization problem can then be reformulated as follows:

$$\begin{aligned} \hat{V}_\alpha = \sup_{Q,U} \quad & \mathbf{E}_\theta[\alpha \cdot \theta \cdot Q(\theta) + (1-\alpha) \cdot U(\theta)] \\ \text{s.t.} \quad & Q \text{ is monotone,} \\ & \int_\theta^{\bar\theta} Q(z)\,\mathrm{d}F(z) \le \int_\theta^{\bar\theta} Q_{\mathrm{E}}(z)\,\mathrm{d}F(z) \quad \forall \theta \in [\underline{\theta}, \bar\theta], \\ & (Q,U) \text{ satisfies (IC).} \end{aligned} \tag{$\hat{\mathcal{P}}_\alpha$}$$

The second inequality constraint is the interim feasibility constraint ($\widehat{\mathrm{IF}}$) characterized in Che et al. (2013). See Lemma 3 for a detailed discussion.

Optimal mechanisms. Let (Q_α, U_α) be the optimal solution for Problem $(\hat{\mathcal{P}}_\alpha)$.[14] The following theorem implies that the optimal allocation partitions the type space into three types of intervals.

Theorem 2. *For any $\alpha \in (0,1)$, the optimal non-coordination mechanism (Q_α, U_α) defines an interval partition $\{(\underline{\theta}^{(j)}, \bar\theta^{(j)})\}_{j=1}^\infty$ of the type space.*[15] *For any $j \ge 1$, the interval $(\underline{\theta}^{(j)}, \bar\theta^{(j)})$ belongs to exactly one of the following three regions:*[16]

(1) It belongs to the no-tension region *if $Q_\alpha(\theta) = U_\alpha(\theta) = Q_{\mathrm{E}}(\theta)$ and $U'_\alpha(\theta) < \eta$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar\theta^{(j)})$.*

(2) It belongs to the no-effort region *if $Q_\alpha(\theta) = U_\alpha(\theta)$ and $U'_\alpha(\theta) = \eta$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar\theta^{(j)})$, and*

$$\int_{\underline{\theta}^{(j)}}^{\bar\theta^{(j)}} Q_\alpha(\theta)\,\mathrm{d}F(\theta) = \int_{\underline{\theta}^{(j)}}^{\bar\theta^{(j)}} Q_{\mathrm{E}}(\theta)\,\mathrm{d}F(\theta).$$

(3) It belongs to the efficient region *if $Q_\alpha(\theta) = Q_{\mathrm{E}}(\theta) > U_\alpha(\theta)$ and $U'_\alpha(\theta) = \eta$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar\theta^{(j)})$.*

For arbitrary n, k and distribution F, the allocation rule can be quite complex even with our characterization, as both the number of intervals in each region and their ordering may exhibit complex dependencies on the shape of the efficient allocation rule and the coefficient α. In Section 5.1, we derive a sharper characterization of the optimal mechanism in large markets.

[14] The existence of an optimal allocation rule is guaranteed by the compactness of the constraint set and the continuity of the objective functional.

[15] If the partition is finite, say, consisting of only K disjoint intervals, then $\underline{\theta}^{(j)} = \bar\theta^{(j)}$ for all $j > K$.

[16] The definitions of the interim allocation and utility on the cutoff points $\{\underline{\theta}^{(j)}\}_{j=1}^\infty$ do not affect the objective value.

Table 1: Regions of the type space for the optimal mechanism

	(IC) binds	($\widehat{\text{IF}}$) binds	effort	Q_α
no-tension region	$\times$	✓	$= 0$	$= Q_{\text{E}}$
no-effort region	✓	$\times$	$= 0$	$\neq Q_{\text{E}}$
efficient region	✓	✓	> 0	$= Q_{\text{E}}$

The proof of Theorem 2, given in Appendix A.2, uses tools from optimal control. Intuitively, we can view the principal's problem as an optimization problem with two constraints, ($\widehat{\text{IF}}$) and (IC). Under optimality, either one of the constraints binds or neither of them binds. When ($\widehat{\text{IF}}$) binds, the optimal allocation rule and the efficient allocation rule coincide. If the slope of the efficient allocation rule is no larger than the marginal cost, the efficient allocation rule can be implemented when no agent exerts effort. This happens in the no-tension region. However, if the slope of the efficient allocation rule is larger than the marginal cost, (IC) requires that agents exert positive effort. This happens in the efficient region. In the second case, the principal can also consider the option of not allocating the items efficiently, i.e., letting ($\widehat{\text{IF}}$) be slack, so that agents have no incentive to exert effort, i.e., (IC) binds, which would imply that the optimal allocation and utility coincide. This happens in the no-effort region. Table 1 summarizes these possibilities.

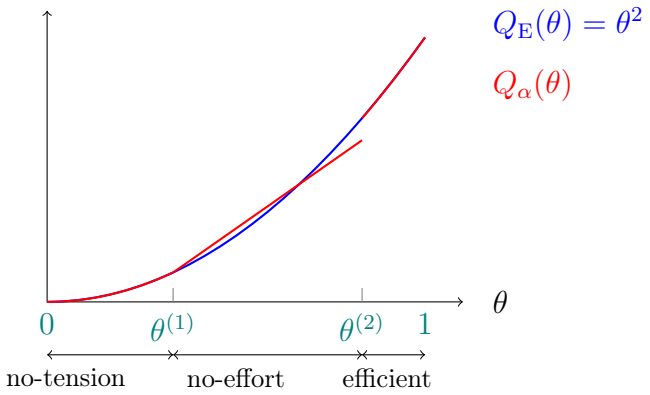
5 Large Contests

In many applications of interest, the number of participating agents is large. In this section, we show that optimal non-coordination mechanisms in such settings exhibit particularly simple structures. To simplify the exposition, we make the following assumption (on top of symmetric environment assumption) throughout the section. However, the main economic insights extend without this assumption.

Assumption 6 (continuity)**.** *There exist* $\underline{\beta}_1, \bar{\beta}_1, \beta_2 \in (0, \infty)$ *such that* $f(\theta) \in [\underline{\beta}_1, \bar{\beta}_1]$ *and* $f'(\theta) \geq -\beta_2$ *for any type* $\theta \in [\underline{\theta}, \bar{\theta}]$.

Note that, as shown in Section 3.3, any symmetric non-coordination mechanism can also be implemented as a coarse ranking contest in arbitrary symmetric environments. Therefore, in the context of this section, we often refer to the non-coordination mechanism as the contest.

Figure 1: Optimal interim allocation rule under convex $Q_{\mathrm{E}}(\theta)$



†Suppose $n = 2$, $k = 1$, $F(\theta) = \theta^2$, $\theta \in [0, 1]$, and $\eta = 1$. In this example, the interim efficient allocation rule is $Q_{\mathrm{E}}(\theta) = F^{n-1}(\theta) = \theta^2$, i.e., the highest type gets the item, and $Q_\alpha(\theta)$ is the optimal interim allocation rule.

5.1 Scarce Resources

In some applications, such as the awarding of prestigious fellowships to university students, the competition is fierce and the ratio of the number of competing agents to the number of items available is large. In this subsection, we study the model of Section 2 in the special case where $k = 1$ and the number of agents n is very large.[17] For sufficiently large n, the efficient allocation rule becomes convex, which simplifies the characterization of the optimal contest. The proofs of the results in this subsection are provided in Appendix A.3.

Lemma 2. *Let $k = 1$. Under Assumption 6, there exists a positive integer N such that for any $n \geq N$, the efficient allocation rule $Q_{\mathrm{E}}(\theta)$ is convex in θ.*

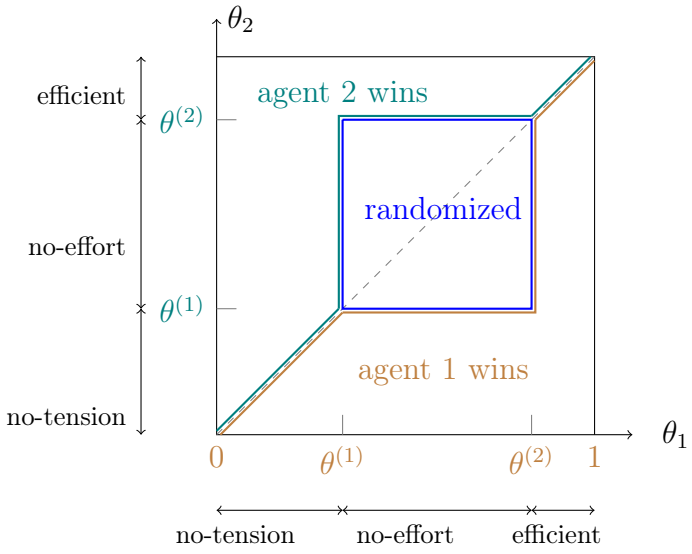
Convex efficient allocation. First consider the case when the efficient allocation rule is convex. The optimal contest is then as follows.

Proposition 3. *Suppose $Q_{\mathrm{E}}(\theta)$ is convex in θ. For any $\alpha \in (0, 1)$, there exist cutoff types $\underline{\theta} \leq \theta^{(1)} \leq \theta^{(2)} \leq \bar{\theta}$ such that in the type space of each agent in the optimal contest Q_α, the interval $(\underline{\theta}, \theta^{(1)})$ is the no-tension region, $(\theta^{(1)}, \theta^{(2)})$ is the no-effort region, and $(\theta^{(2)}, \bar{\theta})$ is the efficient region.*

Figure 1 illustrates the optimal interim allocation rule arising from a convex efficient allocation rule in an example with two agents. Figure 2 illustrates the corresponding ex post rule that maps type profiles to allocations. Intuitively, when the efficient allocation

[17] When k is a small constant greater than 1, the analysis is significantly more involved. We omit this case here since the economic insights it yields are similar.

Figure 2: Implementation of the optimal allocation rule



†Suppose $n = 2$, $k = 1$, $F(\theta) = \theta^2$, $\theta \in [0, 1]$. When both agents produce signals in $(\theta^{(1)}, \theta^{(2)})$, the item is allocated randomly, but the agent with the higher signal has a higher probability (which is strictly less than 1) of getting the item. When at least one agent produces a signal outside $(\theta^{(1)}, \theta^{(2)})$, the item is allocated to the agent with a higher signal.

is convex, its derivative crosses η from below only once. Therefore, for low types, there is no tension: the derivative of the efficient allocation is small enough so that, in the optimal contest, the item can be allocated efficiently without any effort on the part of the agents. For high types, since the change in the efficient allocation is large, the incentive constraints bind and the interim utility must be linear. Moreover, in order for the interim allocation to be interim feasible, in the region where the utility is linear, the no-effort region must occur before the efficient region, not the other way around.

It is interesting to note that in the optimal contest when the efficient allocation rule is convex, there is distortion for middle types but not for high or low types. This stands in sharp contrast to the classical auction design setting, where distortions typically occur for low types.

Convergence results. Using Lemma 2 and Proposition 3, we can immediately characterize the optimal contest for the allocation of scarce resources across a large number of agents. Moreover, we show that as the number of agents increases, the no-tension region converges to the full type space. Since the contest format in the no-tension region is WTA, this implies that in the limit, the format of the optimal contest is essentially WTA.

Theorem 3 (convergence of contest format)**.** *Let $k = 1$. Under Assumption 6, for any*

$\alpha \in (0,1)$, there exists N such that for any $n \geq N$, the optimal contest takes the form described in Proposition 3. Moreover, as n goes to infinity, the no-tension region converges to the entire type space.

Given this convergence result, it may appear tempting to use the WTA contest as an approximation of the optimal contest for a large finite number of agents. However, as shown in Theorem 4 below, the principal's payoff under the WTA contest does not converge to her payoff under the optimal contest as the number of agents increases. This is because in the optimal contest, by randomizing the allocation for a small range of high types, the principal can significantly increase the agents' expected utilities while keeping the loss in matching efficiency small.

For any interim allocation rule Q that is implementable by a contest, by Proposition 4, there exists a unique interim utility U with $U(\underline{\theta}) = Q(\underline{\theta})$ such that (Q, U) is implementable by a contest and yields a weakly higher payoff for the principal than any other mechanism (Theorem 1). Denote this payoff by $V_\alpha(Q)$, i.e.,

$$V_\alpha(Q) = \sup\{\mathrm{Obj}_\alpha(Q,U) : U(\underline{\theta}) = Q(\underline{\theta}) \text{ and } (Q,U) \text{ is implementable by a contest}\}.$$

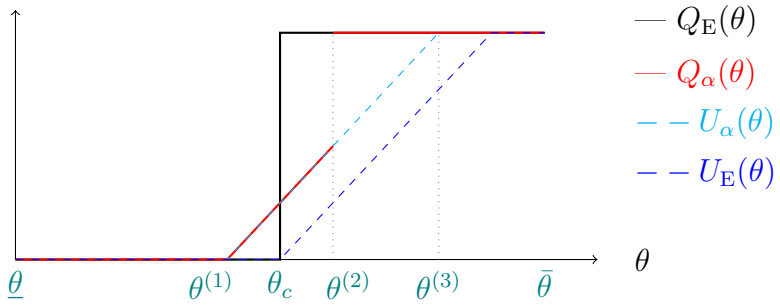
Theorem 4 (non-convergence in payoffs)**.** *Let $k = 1$. Under Assumption 6, for any $\alpha \in (0,1)$ and any sufficiently small $\epsilon > 0$, there exists $N_{F,\epsilon}$ such that for any finite $n > N_{F,\epsilon}$, the ratio between the principal's payoff in the optimal contest and her payoff in the WTA contest is at least $\delta \triangleq \frac{(\bar{\theta}-\epsilon)\cdot\alpha+1-\alpha}{\bar{\theta}\cdot\alpha+(1-\alpha)(1-\frac{1}{e}+\epsilon)} > 1$; that is, $\frac{V_\alpha(Q_{\alpha,n})}{V_\alpha(Q_{\mathrm{E},n})} \geq \delta$.*

Note that our framework enables us to completely characterize the optimal contest for a large but finite number of agents. For comparison, Olszewski and Siegel (2016, 2020) approximate equilibria in contests using a continuum model. Our finding that the optimal contest format converges to the WTA format is consistent with the results of Olszewski and Siegel (2016), but the non-convergence of the optimal payoff (Theorem 4) stands in contrast to their work. The non-convergence in payoff result highlights the importance of randomizing the allocations for top types in practical applications to reduce the cost of wasteful signals.

5.2 Large-Scale Economy

In applications such as college admissions and affordable housing programs, the resources to be allocated are not necessarily scarce. To model such situations, consider a setting with n agents and $0 < k < n$ items, and replicate both the agents and the items $z \in \mathbb{N}_+$ times. The parameter z captures the scale of the economy. As the scale z goes to infinity, the

Figure 3: Optimal allocation and utility for large-scale economy in the limit



†Types in $(\theta^{(1)}, \theta^{(2)})$ are called middle types; in the optimal contest, their utilities are higher than under efficient allocation, because they do not exert effort. Types above $\theta^{(2)}$ are called high types; in the optimal contest their utilities are weakly greater, or in some cases strictly greater, because they each exert less effort.

efficient allocation rule in this setting converges to the cutoff rule, under which the items are allocated to the top $\frac{k}{n}$ of the types. Efficient allocation hence creates strong incentives for types close to the cutoff to exert wasteful effort. Theorem 5 shows that in the optimal contest, to eliminate these incentives, the principal randomizes the allocation for types close to the cutoff. Let θ_c be the cutoff type, defined by $\Pr[\theta \geq \theta_c] = \frac{k}{n}$.

Theorem 5. *Under Assumption 6, for any $\alpha \in (0,1)$ and any fixed integers $n > k > 0$, there exists Z such that for any integer $z \geq Z$, in a setting with $z \cdot n$ agents and $z \cdot k$ items, there exist cutoff types $\underline{\theta} \leq \theta^{(1)} < \theta_c < \theta^{(2)} \leq \theta^{(3)} \leq \bar{\theta}$ such that in the type space of each agent in the optimal contest, the intervals $(\underline{\theta}, \theta^{(1)})$ and $(\theta^{(3)}, \bar{\theta})$ comprise the no-tension region, $(\theta^{(1)}, \theta^{(2)})$ is the no-effort region, and $(\theta^{(2)}, \theta^{(3)})$ is the efficient region.*

Intuitively, in the limit, the efficient allocation rule converges to a step function, with only types above θ_c receiving the items. The interim utility under efficient allocation is thus represented by the blue curve in Figure 3. In order to increase the weighted average between the matching efficiency and the sum of the expected utilities, the principal can randomize the allocation for types around the cutoff θ_c, i.e., within $(\theta^{(1)}, \theta^{(2)})$. This leads to an efficiency loss of at most $\theta^{(2)} - \theta^{(1)}$ when an item is allocated inefficiently, but increases the utilities for all types within $(\theta^{(1)}, \theta^{(3)})$. When $\theta^{(1)}$ is sufficiently close to $\theta^{(2)}$, the increase in expected utility is significantly larger than the efficiency loss. Therefore, the principal can increase her payoff by randomizing on $(\theta^{(1)}, \theta^{(2)})$. Finally, for types that are sufficiently low or sufficiently high, it is easy to verify that both the matching efficiency and the sum of the agents' utilities are maximized under efficient allocation. In Appendix A.4, we show that this intuition applies when the scale of the economy is finite but sufficiently large.

Our result is reminiscent of Director's law, which states that public programs tend to be designed primarily to target the middle classes. Specifically, although the principal cares

about the utilities of all of the agents, the optimal contest gives preferential treatment to the middle types, in the sense that they obtain higher utilities than they would in a fully competitive setting where items are allocated efficiently. This is optimal for the principal because it induces the middle types to exert no effort, which weakly (strictly) decreases the effort level for all (some) of the higher types. This reasoning is also in line with the empirical results of Krishna et al. (2022), who show (using data from Turkey) that randomizing the allocation of college seats to students, especially those with low scores, reduces overall student stress.

6 Conclusions

In this paper, we study the design of optimal screening mechanisms for allocating limited resources to multiple agents based on costly signals. We show that, to reduce socially wasteful costs, it is optimal to implement a non-coordination mechanism in which each agent determines their costly signals without knowledge of any additional information, such as the type reports or signal choices of other agents, beyond the prior belief. We further characterize the optimal non-coordination mechanism and show that it exhibits features consistent with findings in real-world applications.

Our paper opens the door to several promising future lines of research. First, it remains unclear how to characterize the optimal mechanism with fully general signaling costs, and novel techniques may be needed to address this challenge. Second, revisiting the literature on money burning and costly signaling in various applications could help identify additional optimal mechanism structures when signaling costs take a general form, as in our paper. Finally, our work connects to the signaling game literature (e.g., Spence, 1973) by replacing a competitive market with a monopolistic market designer. Investigating the middle ground using oligopoly models, where competition exists but is not fully competitive, would be an intriguing next step.

References

Akbarpour, M., Dworczak, P., and Kominers, S. D. (2024). Redistributive allocation mechanisms. *Journal of Political Economy*, 132(6):1831–1875.

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500.

Arrow, K. (1965). Aspects of the theory of risk bearing. *The Theory of Risk Aversion. Helsinki: Yrjo Jahnssonin Saatio.*

Ball, I. (2024). Scoring strategic agents. *Accepted at American Economic Journal: Microeconomics.*

Border, K. C. (1991). Implementation of reduced form auctions: A geometric approach. *Econometrica: Journal of the Econometric Society*, pages 1175–1187.

Chakravarty, S. and Kaplan, T. R. (2006). Manna from heaven or forty years in the desert: Optimal allocation without transfer payments. *Available at SSRN 939389.*

Che, Y.-K., Kim, J., and Mierendorff, K. (2013). Generalized reduced-form auctions: A network-flow approach. *Econometrica*, 81(6):2487–2520.

Clarke, F. (2013). *Functional Analysis, Calculus of Variations and Optimal Control.* Springer.

Condorelli, D. (2012). What money can't buy: Efficient mechanism design with costly signals. *Games and Economic Behavior*, 75(2):613–624.

Conix, S., De Block, A., and Vaesen, K. (2021). Grant writing and grant peer review as questionable research practices. *F1000Research*, 10.

Dworczak, P., Kominers, S. D., and Akbarpour, M. (2021). Redistribution through markets. *Econometrica*, 89(4):1665–1698.

Fang, D., Noe, T., and Strack, P. (2020). Turning up the heat: The discouraging effect of competition in contests. *Journal of Political Economy*, 128(5):1940–1975.

Finkelstein, A. and Notowidigdo, M. J. (2019). Take-up and targeting: Experimental evidence from snap. *The Quarterly Journal of Economics*, 134(3):1505–1556.

Frankel, A. and Kartik, N. (2019). Muddled information. *Journal of Political Economy*, 127(4):1739–1776.

Fu, Q. and Wu, Z. (2019). Contests: Theory and topics. In *Oxford Research Encyclopedia of Economics and Finance.*

Gershkov, A., Moldovanu, B., Strack, P., and Zhang, M. (2022). Optimal insurance: Dual utility, random losses and adverse selection. *American Economic Review (submitted).*

Green, J. R. and Laffont, J.-J. (1986). Partially verifiable information and mechanism design. *Review of Economic Studies*, 53(3):447–456.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. (2016). Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122.

Hartline, J. D. and Roughgarden, T. (2008). Optimal mechanism design and money burning. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, pages 75–84.

Kilian, L. (2009). Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. *American economic review*, 99(3):1053–1069.

Kleiner, A., Moldovanu, B., and Strack, P. (2021). Extreme points and majorization: Economic applications. *Econometrica*, 89(4):1557–1593.

Krishna, K., Lychagin, S., Olszewski, W., Siegel, R., and Tergiman, C. (2022). Pareto improvements in the contest for college admissions. Technical report, National Bureau of Economic Research.

Laffont, J.-J. and Martimort, D. (2009). The theory of incentives: the principal-agent model. In *The theory of incentives*. Princeton university press.

Lazear, E. P. and Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, 89(5):841–864.

Liu, X. and Lu, J. (2017). Optimal prize-rationing strategy in all-pay contests with incomplete information. *International Journal of Industrial Organization*, 50:57–90.

Milgrom, P. and Shannon, C. (1994). Monotone comparative statics. *Econometrica: Journal of the Econometric Society*, pages 157–180.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73.

Myerson, R. B. (1982). Optimal coordination mechanisms in generalized principal–agent problems. *Journal of mathematical economics*, 10(1):67–81.

Olszewski, W. and Siegel, R. (2016). Large contests. *Econometrica*, 84(2):835–854.

Olszewski, W. and Siegel, R. (2020). Performance-maximizing large contests. *Theoretical Economics*, 15(1):57–88.

Perez-Richet, E. and Skreta, V. (2022). Test design under falsification. *Econometrica*, 90(3):1109–1142.

Perez-Richet, E. and Skreta, V. (2023). Fraud-proof non-market allocation mechanisms. *Working paper*.

Perez-Richet, E. and Skreta, V. (2024). Score-based mechanisms. *arXiv preprint arXiv:2403.08031*.

Pratt, J. W. (1976). Risk aversion in the small and in the large. *Econometrica*, 44(2):420.

Sansone, R. A. and Sansone, L. A. (2011). Faking attention deficit hyperactivity disorder. *Innovations in Clinical Neuroscience*, 8(8):10.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3):355–374.

Yang, F., Dworczak, P., and Akbarpour, M. (2024). Comparison of screening devices. *Revise & Resubmit, Journal of Political Economy*.

Zhang, M. (2024). Optimal contests with incomplete information and convex effort costs. *Theoretical Economics*, 19(1):95–129.

FOR ONLINE PUBLICATION

A Omitted Proofs

A.1 Proofs for Additional Discussions

Proof of Proposition 1. For any symmetric interim allocation–utility pair (Q, U) that is implementable by a non-coordination mechanism, by definition, there exists a mapping from the signal space to the allocation space $x(\hat{s})$ specifying the allocation for each agent given the generated signal $\hat{s}$.[18] The distribution over types $F(\theta)$ induces a distribution over signals; call it $\hat{F}(s)$. Similarly, a feasibility constraint on the allocation rule defined in the signal space may be induced by ($\widehat{\text{IF}}$) based on $\hat{F}$. Such operations are valid since the recommended signal is a non-decreasing function of the type. By Theorems 1 and 2 in Kleiner et al. (2021), any monotone feasible allocation rule x can be written as a convex combination of the extreme points. Using the construction of the extreme points in Theorem 3 of Kleiner et al. (2021), one can easily verify that the extreme points are the coarse ranking contest rules defined in Definition 5. Hence x can be expressed as a randomization over the coarse ranking contest rules. Notice that the above operations do not affect the agents' incentives by preserving the distribution over outcomes; hence the expected utility of each agent in the coarse ranking contest is still U. □

A.2 Proofs for Costly Falsification

Lemma 3 (Che et al., 2013)**.** *Given a set $\boldsymbol{A} = \prod_{i=1}^{n} A_i \subset \Theta$, let $w(\boldsymbol{\theta}, \boldsymbol{A}) = |\{i : \theta_i \in A_i\}|$ be the number of agents whose type θ_i is in A_i.*[19] *The interim allocation rule $\boldsymbol{Q}$ is interim feasible if and only if*

$$\sum_i \int_{A_i} Q_i(\theta_i)\,\mathrm{d}F_i(\theta_i) \le \int_A \min\{k, w(\boldsymbol{\theta}, \boldsymbol{A})\}\,\mathrm{d}\boldsymbol{F}(\theta) \qquad \forall \boldsymbol{A} = \prod_{i=1}^{n} A_i \subset \Theta. \tag{IF}$$

[18] Notice that such a mapping might not exist if (Q, U) is implementable by a direct mechanism that is not a non-coordination mechanism.

[19] Here, $|\cdot|$ denotes the cardinality of a set.

Moreover, for monotone allocations in symmetric environments, (IF) *is equivalent to the following:*[20]

$$\int_{\theta}^{\bar{\theta}} Q(z)\,\mathrm{d}F(z) \leq \int_{\theta}^{\bar{\theta}} Q_{\mathrm{E}}(z)\,\mathrm{d}F(z), \quad \forall \theta \in [\underline{\theta}, \bar{\theta}], \tag{$\widehat{\mathrm{IF}}$}$$

where $Q_{\mathrm{E}}(\theta) = \sum_{j=0}^{k-1} \binom{n-1}{j} \cdot (1-F(\theta))^{j} \cdot F^{n-1-j}(\theta)$ *is the interim allocation rule for allocating* k *items efficiently.*

Incentive compatibility. First we characterize the incentive compatibility conditions in any direct mechanism that implements a monotone allocation rule.

Lemma 4. *An interim allocation–utility pair* $(\boldsymbol{Q}, \boldsymbol{U})$ *with monotone* $\boldsymbol{Q}$ *is implementable by a non-coordination mechanism if and only if* $\boldsymbol{Q}$ *is interim feasible, and for any agent* i *with type* θ_i,[21]

$$\textit{(1)}\ \ U_i'(\theta_i) \in [0, \eta]; \quad \textit{(2)}\ \ U_i(\theta_i) \leq Q_i(\theta_i); \quad \textit{(3)}\ \ U_i'(\theta_i) = \eta \ \textit{if}\ U_i(\theta_i) < Q_i(\theta_i). \tag{IC}$$

The idea behind Lemma 4 is as follows. For any interim allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ that is implementable by a non-coordination mechanism, there is a level constraint and a slope constraint on interim utility. The level constraint is intuitive: because the effort costs are non-negative, each agent's utility is bounded above by his allocation. The slope constraint says that the marginal increase in the interim utility is bounded above by the marginal cost of effort; if this were not the case, then a low type would have an incentive to misreport and produce the recommended signal for a higher type. Finally, if the level constraint is slack at any type θ, the equilibrium effort for type θ must be strictly positive. In order to eliminate the incentives for higher types to deviate to θ, the slope constraint must be binding at type θ.

Proof of Lemma 4. We will prove each direction of the if-and-only-if condition separately.

Only if: If $(\boldsymbol{Q}, \boldsymbol{U})$ is implementable by a non-coordination mechanism, there exist a signal recommendation policy $\hat{\boldsymbol{s}}$ and an allocation rule $\boldsymbol{x}$ that induce $(\boldsymbol{Q}, \boldsymbol{U})$. The allocation rule $\boldsymbol{Q}$ satisfies interim feasibility, because it is induced by the ex-post allocation rule $q_i(\boldsymbol{\theta}) =$

[20] In a symmetric environment, by a slight abuse of notation, we use $F = F_i$ for all i to denote each agent's type distribution.

[21] The function U_i may not be differentiable everywhere. For any type θ_i such that U_i is not differentiable at θ_i, we let $U_i'(\theta_i)$ denote any subgradient (or simply the left and right derivative) of the function U_i. It is not hard to show that U_i is a monotone function and hence is differentiable almost everywhere.

$x_i(\hat{s}_i(\theta_i), \hat{\boldsymbol{s}}_{-i}(\theta_{-i}))$ for all i and $\boldsymbol{\theta}$. Notice that for any signal recommendation policy $\hat{\boldsymbol{s}}$ and allocation rule $\boldsymbol{x}$ implementing $(\boldsymbol{Q}, \boldsymbol{U})$, it is without loss to assume that any realization of the recommendation $\hat{s}_i(\theta_i)$ is weakly higher than θ_i. This is because weakly increasing a signal recommendation below θ_i does not induce type θ_i to exert additional effort, but weakly decreases other types' incentives for deviation.

For any agent i and any pair of types $\theta_i < \theta_i'$, let s_i' be the largest signal realization given $\hat{s}_i(\theta_i')$. Thus we have $s_i' \geq \theta_i'$. Note that agent i with type θ_i' obtains utility $U_i(\theta_i')$ from choosing signal s_i', as he must be indifferent for all his signal realizations. Therefore, agent i's utility from reporting signal s_i' when his type is θ_i is

$$\begin{aligned}\mathbf{E}_{\theta_{-i}}\big[x_i(s_i', \hat{s}_{-i}(\theta_{-i}))\big] - \eta \cdot e(s_i', \theta_i) &= \mathbf{E}_{\theta_{-i}}\big[x_i(s_i', \hat{s}_{-i}(\theta_{-i}))\big] - \eta \cdot e(s_i', \theta_i') - \eta \cdot (\theta_i' - \theta_i) \\ &= U_i(\theta_i') - \eta \cdot (\theta_i' - \theta_i).\end{aligned}$$

Since his utility from deviating in his choice of signal is weakly lower, we have

$$U_i(\theta_i) \geq U_i(\theta_i') - \eta \cdot (\theta_i' - \theta_i).$$

By rearranging the terms and taking the limit as $\theta_i' \to \theta_i$, we obtain $U_i'(\theta) \leq \eta$. Similarly, let s_i be the largest signal realization given $\hat{s}_i(\theta_i)$. We have

$$\begin{aligned}U_i(\theta_i') &\geq \mathbf{E}_{\theta_{-i}}[x_i(s_i, \hat{s}_{-i}(\theta_{-i}))] - \eta \cdot e(s_i, \theta_i') \\ &\geq \mathbf{E}_{\theta_{-i}}[x_i(s_i, \hat{s}_{-i}(\theta_{-i}))] - \eta \cdot e(s_i, \theta_i) = U_i(\theta_i).\end{aligned}$$

Again by rearranging the terms and taking the limit as $\theta_i' \to \theta_i$, we have $U_i'(\theta) \geq 0$. Hence $U_i'(\theta_i) \in [0, \eta]$ for any type θ_i.

Finally, as the effort is non-negative, the interim allocation must be weakly larger than the interim utility. When the inequality is strict, the agent must choose signal realizations strictly higher than his type with positive probability. In this case, we have $s_i > \theta_i$. For any type $\theta_i' \in (\theta_i, s_i)$, we have

$$\begin{aligned}U_i(\theta_i') &\geq \mathbf{E}_{\theta_{-i}}[x_i(s_i, \hat{s}_{-i}(\theta_{-i}))] - \eta \cdot e(s_i, \theta_i') \\ &= \mathbf{E}_{\theta_{-i}}[x_i(s_i, \hat{s}_{-i}(\theta_{-i}))] - \eta \cdot e(s_i, \theta_i) + \eta \cdot (\theta_i' - \theta_i) \\ &= U_i(\theta_i) + \eta \cdot (\theta_i' - \theta_i).\end{aligned}$$

By rearranging the terms and taking the limit as $\theta_i' \to \theta_i$, we obtain $U_i'(\theta_i) \geq \eta$. Since we also know that $U_i'(\theta_i) \leq \eta$, both inequalities must be equalities and hence $U_i'(\theta_i) = \eta$.

If: Since $\boldsymbol{Q}$ is interim feasible, there exists an ex-post allocation rule $\boldsymbol{q}$ that implements $\boldsymbol{Q}$. Consider the signal recommendation policy $\hat{\boldsymbol{s}}$ where $\hat{s}_i(\theta_i) = \theta_i + \frac{1}{\eta}(Q_i(\theta_i) - U_i(\theta_i))$ for any agent i with type θ_i. It is easy to verify that $\hat{s}_i(\theta_i)$ is monotone in θ_i, since $Q_i'(\theta_i) \geq 0$ and $U_i'(\theta_i) \leq \eta$. Let $\theta_i(s_i)$ be the inverse of function $\hat{s}_i$.[22] Consider the allocation rule $\boldsymbol{x}$ where $x_i(\boldsymbol{s}) = q_i(\boldsymbol{\theta}(\boldsymbol{s}))$ for all agents i. We show that $\hat{\boldsymbol{s}}$ and $\boldsymbol{x}$ implement $(\boldsymbol{Q}, \boldsymbol{U})$.

First, by our construction, when all agents follow the recommendations, the interim allocation and the interim utility coincide with $\boldsymbol{Q}$ and $\boldsymbol{U}$, respectively. Thus, it is sufficient to show that the agents have weak incentives to follow the recommendations. In particular, if agent i with type θ_i deviates to reporting type $\theta_i' > \theta_i$, his utility from deviation is

$$Q_i(\theta_i') - \eta \cdot e(\hat{s}_i(\theta_i'), \theta_i) = U_i(\theta_i') - \eta \cdot (\theta_i' - \theta_i) \leq U_i(\theta_i),$$

where the last inequality holds because the derivative of U is always at most η. We now analyze the incentives for downward deviation in three cases. If the deviation type $\theta_i' < \theta_i$ satisfies $Q_i(\theta_i') = U_i(\theta_i')$, the utility from deviation is

$$Q_i(\theta_i') - \eta \cdot e(\hat{s}_i(\theta_i'), \theta_i) = U_i(\theta_i') \leq U_i(\theta_i).$$

If the deviation type $\theta_i' < \theta_i$ satisfies $Q_i(\theta_i') > U_i(\theta_i')$, let $\theta_i^\dagger > \theta_i'$ be the smallest type such that $Q_i(\theta_i^\dagger) = U_i(\theta_i^\dagger)$. If $\theta_i \geq \theta_i^\dagger$, the utility from deviation is

$$Q_i(\theta_i') - \eta \cdot e(\hat{s}_i(\theta_i'), \theta_i) \leq Q_i(\theta_i^\dagger) = U_i(\theta_i^\dagger) \leq U_i(\theta_i).$$

If $\theta_i < \theta_i^\dagger$, the derivative of U for any type between θ_i' and θ_i must be constant and equal to η. Hence the utility from deviation is

$$Q_i(\theta_i') - \eta \cdot e(\hat{s}_i(\theta_i'), \theta_i) \leq U_i(\theta_i') + \eta \cdot (\theta_i - \theta_i') = U_i(\theta_i).$$

Combining these inequalities, we conclude that none of the agents have any incentive to deviate from the recommendations. □

Payoff equivalence. From Lemma 4 we can establish Proposition 4, which says that for any allocation–utility pair implementable by a non-coordination mechanism, the utility function for each agent is uniquely pinned down by the interim allocation, up to the choice of the utility for the lowest type.

[22] Note that $\hat{s}_i(\theta_i)$ is only weakly monotone. When there are multiple types θ_i with the same signal recommendation s_i, we map s_i randomly to those types according to the type distribution F_i.

Proposition 4. *Fix any monotone and interim feasible allocation rule $\boldsymbol{Q}$, and any $\{\underline{u}_i\}_{i=1}^n$ such that $\underline{u}_i \leq Q_i(\underline{\theta}_i)$ for all i. There exists a unique interim utility profile $\boldsymbol{U}$ with $U_i(\underline{\theta}_i) = \underline{u}_i$ for all i such that $(\boldsymbol{Q}, \boldsymbol{U})$ is implementable by a non-coordination mechanism. Moreover, for any interim allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ that is implementable by a non-coordination mechanism, we have the following:*

- *If $U_i(\underline{\theta}_i) > U_i^\dagger(\underline{\theta}_i)$ for any agent i, then $U_i(\theta_i) \geq U_i^\dagger(\theta_i)$ for every agent i and every type θ_i.*
- *If $U_i(\underline{\theta}_i) = U_i^\dagger(\underline{\theta}_i)$ for any agent i, then $U_i(\theta_i) = U_i^\dagger(\theta_i)$ for every agent i and every type θ_i.*

In the classical mechanism design setting, payoff equivalence means that once the allocation is determined, the curvature of the utility function is fixed, and the utility function can only be shifted by a constant determined by the utility of the lowest type. In our setting, for a fixed allocation rule, shifting the utility of the lowest type does not shift the utilities for all types by the same constant. We illustrate this in Figure 4. For any agent i, if the utility of the lowest type is lower than the interim allocation of the lowest type, or if the derivative of the interim allocation is larger than the parameter η, then (IC) implies that the interim utility U_i must be a straight line with derivative η until U_i intersects Q_i (in the example in Figure 4, the intersection occurs at type $\theta_i^{(1)}$). Then U_i coincides with Q_i until the derivative of Q_i exceeds η. In a setting with discrete types, one could apply this reasoning to recursively pin down the interim utility for all types. Unfortunately, the recursive argument fails to work when the type space is continuous and we provide a formal proof to circumvent this technicality.

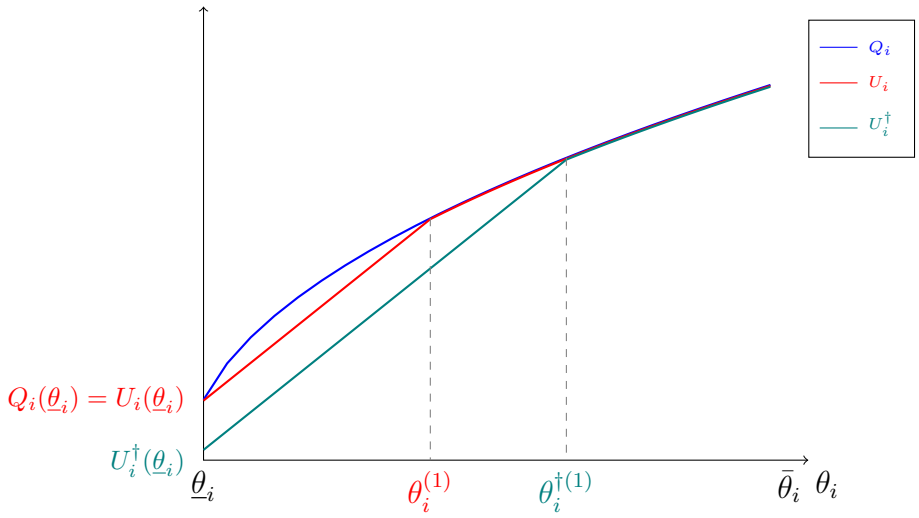
It is immediate from Proposition 4 that the lowest type exerts zero effort in the optimal non-coordination mechanism, because if we set the utility of the lowest type equal to the interim allocation, then the utilities of all higher types are weakly increased.

Proof of Proposition 4. For any agent i, given any monotone and interim feasible allocation $\boldsymbol{Q}$, and $\underline{u}_i \leq Q_i(\underline{\theta}_i)$, let

$$U_i(\theta_i) = \min\left\{\underline{u}_i + \eta(\theta_i - \underline{\theta}_i),\ \inf_{\theta_i' \leq \theta_i} Q_i(\theta_i') + \eta(\theta_i - \theta_i')\right\}. \tag{2}$$

Notice that $U_i(\theta_i) \leq Q_i(\theta_i)$, because $\inf_{\theta_i' \leq \theta_i} Q_i(\theta_i') + \eta(\theta_i - \theta_i') \leq Q_i(\theta_i)$ for all θ_i. For those types θ_i such that $U_i(\theta_i) < Q_i(\theta_i)$, by the definition of U_i, there exists some $\theta_i' < \theta_i$ such that $U_i(\theta_i) = \min\{\underline{u}_i + \eta(\theta_i - \underline{\theta}_i),\ Q_i(\theta_i') + \eta(\theta_i - \theta_i')\}$, implying that $U_i'(\theta_i) = \eta$. For those types θ_i such that $U_i(\theta_i) = Q_i(\theta_i)$, by definition, $Q_i(\theta_i') + \eta(\theta_i - \theta_i') \geq Q_i(\theta_i)$ for all $\theta_i' < \theta_i$.

Figure 4: Illustration of Proposition 4



†Both U_i and $U_i^\dagger$ implement the allocation rule Q_i, but U_i gives agent i a higher utility and hence is the better implementation from the principal's point of view. Moreover, U_i and $U_i^\dagger$ do not differ by a constant as in the standard payoff equivalence result (where the constant would equal $U_i(\underline{\theta}_i) - U_i^\dagger(\underline{\theta}_i)$). However, by the construction provided in the proof of Proposition 4, U_i is uniquely identified by the allocation rule and $U_i(\underline{\theta}_i)$.

This can happen only when $Q_i'(\theta_i) < \eta$, implying that $U_i'(\theta_i) < \eta$. Hence $(\boldsymbol{Q}, \boldsymbol{U})$ satisfies (IC) and is implementable by a non-coordination mechanism.

Next we show that $\boldsymbol{U}$ is the unique utility profile such that $(\boldsymbol{Q}, \boldsymbol{U})$ is implementable by a non-coordination mechanism, given utilities $\{\underline{u}_i\}_{i=1,\ldots,n}$ for the lowest types. Suppose $\boldsymbol{U}^\dagger$ is a different utility profile such that $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ is implementable by a non-coordination mechanism and $U_i^\dagger(\underline{\theta}_i) = \underline{u}_i$ for all i. Then (IC) implies that $U_i^\dagger(\theta_i) \leq Q_i(\theta_i)$ for all θ_i.

Suppose there exists θ_i such that $U_i^\dagger(\theta_i) > U_i(\theta_i)$. This is only possible if $U_i(\theta_i) < Q_i(\theta_i)$ and there exists some $\theta_i' < \theta_i$ such that $U_i(\theta_i) = Q_i(\theta_i') + \eta(\theta_i - \theta_i') < U_i^\dagger(\theta_i)$. However, this implies that in the direct mechanism $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$, agent i with type θ_i' has an incentive to misreport his type as θ_i. This contradicts the assumption that $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ is implementable by a non-coordination mechanism.

Now suppose there exists θ_i such $U_i^\dagger(\theta_i) < U_i(\theta_i)$. This is only possible if $U_i^\dagger(\theta_i) < Q_i(\theta_i)$. Let $\theta_i' = \underline{\theta}_i$ if $U_i^\dagger(\theta_i) < Q_i(\theta_i)$ for all θ_i. Otherwise, let $\theta_i' = \sup\{z \leq \theta_i : U_i^\dagger(\theta_i') = Q_i(\theta_i')\}$. In both cases, by (IC), we have $U_i^\dagger(\theta_i) = U_i^\dagger(\theta_i') + \eta(\theta_i - \theta_i')$. In the case where $\theta_i' = \underline{\theta}_i$, we have

$$U_i^\dagger(\theta_i) = U_i^\dagger(\theta_i') + \eta(\theta_i - \theta_i') < U_i(\theta_i) \leq \underline{u}_i + \eta(\theta_i - \theta_i'),$$

implying that $U_i^\dagger(\underline{\theta}_i) < \underline{u}_i$, a contradiction. In the case where $\theta_i' > \underline{\theta}_i$, we can similarly

infer that $U_i^\dagger(\theta_i') < Q_i(\theta_i')$, which is again a contradiction. Hence, for any interim allocation rule $\boldsymbol{Q}$, if there exists an interim utility $\boldsymbol{U}$ such that $(\boldsymbol{Q}, \boldsymbol{U})$ is implementable by a non-coordination mechanism, then $\boldsymbol{U}$ is uniquely pinned down by $\boldsymbol{Q}$ and the utility profile for the lowest types, and it is given by the expression (2).

Finally, for any $\boldsymbol{U}^\dagger$ such that $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ is implementable by a non-coordination mechanism, if $U_i^\dagger(\underline{\theta}_i) < U_i(\underline{\theta}_i)$ for all i, then by (2) we must have $U_i^\dagger(\theta_i) \le U_i(\theta_i)$ for all θ_i. □

Optimality of symmetric mechanisms.

Proof of Lemma 1. Consider a relaxed problem $(\mathcal{P}'_\alpha)$ where, instead of the (IC) constraints, we only require that $U_i'(\theta_i) \in [0, \eta]$ and $U_i(\theta_i) \le Q_i(\theta_i)$ for any agent i with type θ_i. Note that this is a convex constraint, and hence the relaxed problem is a convex problem. Thus, there exists a symmetric optimal solution $(\boldsymbol{Q}, \boldsymbol{U})$ for Problem $(\mathcal{P}'_\alpha)$ if the environment is symmetric. Moreover, as $\boldsymbol{U}$ is maximized given the derivative constraint and the upper bound of $\boldsymbol{Q}$, the allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ also satisfies the (IC) constraints by the proof of Proposition 4. Therefore, $(\boldsymbol{Q}, \boldsymbol{U})$ is also feasible and hence is an optimal solution for Problem $(\hat{\mathcal{P}}_\alpha)$. □

Characterization of the optimal mechanism. To simplify the notation in the later analysis, given the partition of the type space, we add a degenerate interval $\underline{\theta}^{(0)} = \bar{\theta}^{(0)} = \bar{\theta}$.

Proof of Theorem 2. By Lemma 4, the optimal utility function U_α must be continuous with subgradient between 0 and η, and $Q_\alpha(\theta) = U_\alpha(\theta)$ if $Q'_\alpha(\theta) < \eta$. Therefore, we can partition the type space into countably many disjoint intervals $\{(\underline{\theta}^{(j)}, \bar{\theta}^{(j)})\}_{j=1}^{\infty}$, each of which falls into one of the following three categories:

Case 1: $Q_\alpha(\theta) = U_\alpha(\theta)$ and $U'_\alpha(\theta) < \eta$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$.

Case 2: $Q_\alpha(\theta) = U_\alpha(\theta)$ and $U'_\alpha(\theta) = \eta$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$.

Case 3: $Q_\alpha(\theta) > U_\alpha(\theta)$ and $U'_\alpha(\theta) = \eta$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$.

For any interim allocation rule Q, let $\mathcal{Q}(\theta) = \int_\theta^{\bar{\theta}} Q(t)\,\mathrm{d}F(t)$. Notice that $\mathcal{Q}(\theta)$ is a continuous function.

Lemma 5. *If Q is optimal, then $\mathcal{Q}(\theta) - \int_\theta^{\bar{\theta}} Q_{\mathrm{E}}(t)\,\mathrm{d}F(t) < 0$ implies*

(A) $U(\theta) = Q(\theta)$, and

(B) either $U'(\theta) = \eta$ or $U'(\theta) = 0$.

Corollary 1. *If (Q,U) is optimal, then $Q(\theta) > U(\theta)$ implies $\mathcal{Q}(\theta) - \int_\theta^{\bar\theta} Q_{\mathrm{E}}(t)\,\mathrm{d}F(t) = 0$ and $Q(\theta) = Q_{\mathrm{E}}(\theta)$ almost everywhere.*

Corollary 2. *If (Q,U) is optimal, then $0 < U'(\theta) < \eta$ implies $\mathcal{Q}(\theta) - \int_\theta^{\bar\theta} Q_{\mathrm{E}}(t)\,\mathrm{d}F(t) = 0$ and $Q(\theta) = Q_{\mathrm{E}}(\theta)$ almost everywhere.*

Corollary 1 implies that in Case 3, we have $Q_\alpha(\theta) = Q_{\mathrm{E}}(\theta) > U_\alpha(\theta)$. Thus Case 3 will correspond to the efficient region.

The analysis of Case 1 is decomposed into two subcases:

Case 1a: $Q_\alpha(\theta) = U_\alpha(\theta)$ and $U'_\alpha(\theta) \in (0,\eta)$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar\theta^{(j)})$.

Case 1b: $Q_\alpha(\theta) = U_\alpha(\theta)$ and $U'_\alpha(\theta) = 0$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar\theta^{(j)})$.

By Corollary 2, in Case 1a, we have $\mathcal{Q}(\theta) - \int_\theta^{\bar\theta} Q_{\mathrm{E}}(t)\,\mathrm{d}F(t) = 0$ and $Q_\alpha(\theta) = Q_{\mathrm{E}}(\theta)$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar\theta^{(j)})$. Therefore, Case 1a corresponds to the no-tension region. Moreover, we show that Case 1b cannot occur (the proof is deferred to the end of the section). Thus Case 1 gives the no-tension region.

Lemma 6. *Case 1b does not occur in the optimal solution.*

Finally, for any interval j that corresponds to Case 2, if $\underline{\theta}^{(j)} > \underline{\theta}$, since the integration constraint $(\widehat{\mathrm{IF}})$ binds for all types within each interval under any of the two other cases, it must also bind for both endpoints of the interval j; hence

$$\int_{\underline{\theta}^{(j)}}^{\bar\theta^{(j)}} Q_\alpha(\theta)\,\mathrm{d}F(\theta) = \int_{\underline{\theta}^{(j)}}^{\bar\theta^{(j)}} Q_{\mathrm{E}}(\theta)\,\mathrm{d}F(\theta).$$

If $\underline{\theta}^{(j)} = \underline{\theta}$, then $(\widehat{\mathrm{IF}})$ also binds at $\underline{\theta}$, since otherwise we could increase the allocation and utility for a sufficiently small region above type $\underline{\theta}$ without violating feasibility, which would contradict the optimality of the solution. Hence the above equality again holds, and Case 2 corresponds to the no-effort region. □

Proof of Lemma 5. Consider the following relaxation of Problem $(\hat{\mathcal{P}}_\alpha)$:

$$\begin{aligned} \sup_{Q,U} \quad & \mathbf{E}_\theta[\alpha\cdot\theta\cdot Q(\theta) + (1-\alpha)\cdot U(\theta)] \\ \text{s.t.} \quad & \int_\theta^{\bar\theta} Q(\theta)\,\mathrm{d}F(z) \le \int_\theta^{\bar\theta} Q_{\mathrm{E}}(z)\,\mathrm{d}F(z) \quad \forall\theta\in[\underline{\theta},\bar\theta], \\ & U(\theta) \le Q(\theta), \qquad 0 \le U'(\theta) \le a. \end{aligned} \tag{$\hat{\mathcal{R}}_\alpha$}$$

Here we have omitted the monotonicity constraint on the allocation, as doing so does not affect the optimal solution (Theorem 1).

Define $\mathcal{Q}(\theta) = \int_\theta^{\bar\theta} Q(t)\,\mathrm{d}F(t)$ and $\mathcal{Q}'(\theta) = -Q(\theta)f(\theta)$. The relaxed problem can be rewritten as follows:

$$\begin{aligned}
\sup_{\mathcal{Q},U} \quad & \int_{\underline{\theta}}^{\bar\theta} -\alpha\cdot\theta\cdot\mathcal{Q}'(\theta) + (1-\alpha)\cdot U(\theta)\cdot f(\theta)\,\mathrm{d}\theta \\
\text{s.t.} \quad & \mathcal{Q}(\theta) \le \int_\theta^{\bar\theta} F^{n-1}(t)\,\mathrm{d}F(t), && \lambda(\theta), \\
& U(\theta)f(\theta) + \mathcal{Q}'(\theta) \le 0, && \gamma(\theta), \\
& 0 \le U'(\theta), && \kappa_1(\theta), \\
& U'(\theta) \le a, && \kappa_2(\theta).
\end{aligned}$$

The Lagrange multipliers $\lambda(\theta), \gamma(\theta), \kappa_1(\theta), \kappa_2(\theta)$ are non-negative. The Lagrangian is given by

$$\begin{aligned}
\hat{\mathcal{L}}(\mathcal{Q},\mathcal{Q}',U,U',\lambda,\gamma,\kappa_1,\kappa_2) = & -[\alpha\theta\cdot\mathcal{Q}'(\theta) - (1-\alpha)\cdot U(\theta)f(\theta) \\
& + \lambda(\theta)(\mathcal{Q}(\theta) - \int_\theta^{\bar\theta} Q_{\mathrm{E}}(t)\,\mathrm{d}F(t)) \\
& + \gamma(\theta)(U(\theta)f(\theta) + \mathcal{Q}'(\theta)) \\
& + \kappa_1(\theta)(U'(\theta) - a) - \kappa_2(\theta)U'(\theta)].
\end{aligned}$$

The solution of the problem satisfies the following conditions:

(1) The Euler–Lagrange conditions,[23]

$$\frac{\partial\hat{\mathcal{L}}}{\partial\mathcal{Q}} - \frac{\mathrm{d}}{\mathrm{d}\theta}\frac{\partial\hat{\mathcal{L}}}{\partial\mathcal{Q}'} = 0 \Leftrightarrow \quad \lambda(\theta) - (\alpha + \gamma'(\theta)) = 0 \tag{EL-1}$$

and

$$\frac{\partial\hat{\mathcal{L}}}{\partial U} - \frac{\mathrm{d}}{\mathrm{d}\theta}\frac{\partial\hat{\mathcal{L}}}{\partial U'} = 0 \Leftrightarrow \quad (\gamma(\theta) - (1-\alpha))f(\theta) - \kappa'(\theta) = 0, \tag{EL-2}$$

where $\kappa(\theta) = \kappa_1(\theta) - \kappa_2(\theta)$, hold whenever they are well-defined.

[23] We are looking for piecewise continuous solutions (the state variables are continuous and the control variables are piecewise continuous), since, in principle, the allocation $Q(\theta)$ may be merely piecewise continuous and not continuous, while $U(\theta)$ is continuous but its derivative might not be. The necessary conditions should be the integral form of the Euler–Lagrange conditions, together with the Erdmann–Weierstrass corner conditions (cf. Clarke, 2013). However, the latter have no bite here, and we can use the usual form of the Euler–Lagrange conditions, since they do not involve the state variables or the controls. Notice, though, that the Lagrange multiplier $\gamma(\theta)$ could potentially be PC^1.

(2) The complementary slackness conditions hold:

$$\lambda(\theta)(\mathcal{Q}(\theta) - \int_{\theta}^{\bar{\theta}} F^{n-1}(t)\,\mathrm{d}F(t)) = 0, \quad \lambda(\theta) \geq 0, \tag{CS-1a}$$

$$\gamma(\theta)[U(\theta)f(\theta) + \mathcal{Q}'(\theta)] = 0, \quad \gamma(\theta) \geq 0, \tag{CS-1b}$$

$$\kappa_1(\theta)[U'(\theta) - a] = 0, \quad \kappa_1(\theta) \geq 0, \tag{CS-1c}$$

$$\kappa_2(\theta)U'(\theta) = 0, \quad \kappa_2(\theta) \geq 0. \tag{CS-1d}$$

Suppose $\mathcal{Q}(\theta) - \int_{\theta}^{\bar{\theta}} Q_{\mathrm{E}}(t)\,\mathrm{d}F(t) < 0$. We show that the following two conditions hold for the optimal solution:

- $U(\theta) = Q(\theta)$. (By (CS-1a), $\lambda(\theta) = 0$ holds in an interval. From (EL-1), we have $\gamma'(\theta) = -\alpha$. Hence $\gamma(\theta)$ cannot be a constant in this interval; in particular, $\gamma(\theta) \neq 0$ except for at most one point. Combined with (CS-1b), this further implies that $U(\theta)f(\theta) + \mathcal{Q}'(\theta) = 0$, i.e., $U(\theta) = Q(\theta)$.)

- Either $U'(\theta) = a$ or $U'(\theta) = 0$. (Reasoning similarly as for the previous condition, we have that $\gamma(\theta) \neq 1 - \alpha$ except for at most one point, which combined with (EL-2) implies that $\kappa'(\theta) \neq 0$ except for at most one point. This means that $\kappa(\theta)$ is not a constant; in particular, it is not zero. The result follows from applying (CS-1c) and (CS-1d).) □

Proof of Corollary 1. The contrapositive of Lemma 5 is also true: $Q(\theta) > U(\theta)$ implies $\mathcal{Q}(\theta) - \int_{\theta}^{\bar{\theta}} Q_{\mathrm{E}}(t)\,\mathrm{d}F(t) = 0$. By rearranging the terms and taking the derivative with respect to θ, we have $Q(\theta) = Q_{\mathrm{E}}(\theta)$ almost everywhere. □

Proof of Lemma 6. Suppose Case 1b occurs. In this case, since U is a continuous function, $Q_\alpha(\theta) = U_\alpha(\theta) = U_\alpha(\bar{\theta}^{(j)})$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$. Let j' be the index of the interval such that $\bar{\theta}^{(j)} = \underline{\theta}^{(j')}$. We consider three possible situations for interval j':

- Interval j' belongs to Case 1. In this case, the integration constraint ($\widehat{\mathrm{IF}}$) binds at $\underline{\theta}^{(j')}$, and $U(\bar{\theta}^{(j)}) = U(\underline{\theta}^{(j')}) = Q_{\mathrm{E}}(\underline{\theta}^{(j')})$. Therefore, there exists a constant $\epsilon > 0$ such that ($\widehat{\mathrm{IF}}$) is violated at type $\bar{\theta}^{(j)} - \epsilon$, a contradiction.

- Interval j' belongs to Case 2. In this case, ($\widehat{\mathrm{IF}}$) does not bind at type $\underline{\theta}^{(j')}$. Suppose otherwise; then we must have $Q_{\mathrm{E}}(\underline{\theta}^{(j')}) \leq U(\bar{\theta}^{(j)})$ in order for ($\widehat{\mathrm{IF}}$) to hold for type $\underline{\theta}^{(j')} + \epsilon$ given sufficiently small $\epsilon > 0$. However, this would imply that ($\widehat{\mathrm{IF}}$) is violated for type $\bar{\theta}^{(j)} - \epsilon$ given sufficiently small $\epsilon > 0$.

  Next we consider two cases for interval j.

– $\underline{\theta}^{(j)} > \underline{\theta}$. In this case, $(\widehat{\text{IF}})$ cannot bind at any type $\theta \in [\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$. This is because if it binds at θ, then $Q(\theta) = U(\theta) > Q_{\text{E}}(\theta)$. By the continuity of U and Q_{E}, and the constraint that $Q \geq U$, there exists a constant $\epsilon > 0$ such that $(\widehat{\text{IF}})$ is violated at $\theta - \epsilon$. Thus, there exist $\epsilon, \delta > 0$ such that for any type $\theta \in [\underline{\theta}^{(j)} - \epsilon, \bar{\theta}^{(j)} + \epsilon]$,

$$\int_{\theta}^{\bar{\theta}} Q(z) \, \mathrm{d}F(z) \leq \int_{\theta}^{\bar{\theta}} Q_{\text{E}}(z) \, \mathrm{d}F(z) - \delta.$$

Moreover, we can select ϵ to be sufficiently small to satisfy the additional condition that $Q'(\theta) \leq \eta$ for any type $\theta \in [\underline{\theta}^{(j)} - \epsilon, \bar{\theta}^{(j)} + \epsilon]$. Given a parameter θ^*, let $Q^{\ddagger}$ be the allocation such that

(1) $Q^{\ddagger}(\theta) = Q(\underline{\theta}^{(j)} - \epsilon)$ for any type $\theta \in [\underline{\theta}^{(j)} - \epsilon, \theta^*]$;

(2) $Q^{\ddagger}(\theta) = Q(\underline{\theta}^{(j)} - \epsilon) + \eta \cdot (\theta - \theta^*)$ for any type $\theta \in (\theta^*, \theta^* + \frac{1}{\eta} \cdot Q(\bar{\theta}^{(j)} + \epsilon) - Q(\underline{\theta}^{(j)} - \epsilon))$;

(3) $Q^{\ddagger}(\theta) = Q(\bar{\theta}^{(j)} + \epsilon)$ for any type $\theta \in [\theta^* + \frac{1}{\eta} \cdot Q(\bar{\theta}^{(j)} + \epsilon) - Q(\underline{\theta}^{(j)} - \epsilon), \bar{\theta}^{(j)} + \epsilon]$.

The parameter θ^* is chosen so that

$$\int_{\underline{\theta}^{(j)} - \epsilon}^{\bar{\theta}^{(j)} + \epsilon} Q^{\ddagger}(z) \, \mathrm{d}F(z) = \int_{\underline{\theta}^{(j)} - \epsilon}^{\bar{\theta}^{(j)} + \epsilon} Q(z) \, \mathrm{d}F(z).$$

It is easy to verify that

$$\int_{\underline{\theta}^{(j)} - \epsilon}^{\bar{\theta}^{(j)} + \epsilon} z \cdot Q^{\ddagger}(z) \, \mathrm{d}F(z) > \int_{\underline{\theta}^{(j)} - \epsilon}^{\bar{\theta}^{(j)} + \epsilon} z \cdot Q(z) \, \mathrm{d}F(z),$$

since $Q^{\ddagger}$ shifts allocation probabilities from low types to high types compared to Q. Therefore, given a sufficiently small constant $\hat{\delta} > 0$, consider another allocation–utility pair $(Q^{\dagger}, U^{\dagger})$ such that

(1) $Q^{\dagger}(\theta) = Q(\theta)$ and $U^{\dagger}(\theta) = U(\theta)$ for any type $\theta \notin [\underline{\theta}^{(j)} - \epsilon, \bar{\theta}^{(j)} + \epsilon]$;

(2) $Q^{\dagger}(\theta) = (1 - \hat{\delta}) \cdot Q(\theta) + \hat{\delta} \cdot Q^{\ddagger}(\theta)$ and $U^{\dagger}(\theta) = (1 - \hat{\delta}) \cdot U(\theta) + \hat{\delta} \cdot Q^{\ddagger}(\theta)$ for any type $\theta \in [\underline{\theta}^{(j)} - \epsilon, \bar{\theta}^{(j)} + \epsilon]$.

The new allocation–utility pair $(Q^{\dagger}, U^{\dagger})$ is feasible and strictly improves the objective value, a contradiction to the optimality of (Q, U).

– $\underline{\theta}^{(j)} = \underline{\theta}$. The proof for this case is similar. The only difference is that we can change the allocation and utility within interval j without worrying about the continuity of the utility function for lower types. Therefore, using a similar

construction for $Q^\ddagger$ and $(Q^\dagger, U^\dagger)$, restricted to the interval $[\underline{\theta}^{(j)}, \bar{\theta}^{(j)} + \epsilon]$ for sufficiently small $\epsilon > 0$, we can again show that the allocation–utility pair (Q, U) that contains Case 1b is not optimal.

- Either interval j' belongs to Case 3, or $\underline{\theta}^{(j')}$ is the highest possible type $\bar{\theta}$. In either case, for the integration constraint $(\widehat{\mathrm{IF}})$ to be satisfied within interval j, both the efficient allocation Q_{E} and the interim allocation Q must be strictly above the utility at $\underline{\theta}^{(j')}$. Therefore, the allocation within interval j can be increased, relative to allocations above $\underline{\theta}^{(j')}$, without violating the monotonicity. Again we use a similar construction for $Q^\ddagger$ and $(Q^\dagger, U^\dagger)$, restricted to the interval $[\underline{\theta}^{(j)} - \epsilon, \bar{\theta}^{(j)}]$ for sufficiently small $\epsilon > 0$. Here we add the further operation of increasing the utility $U^\dagger$ for types above $\underline{\theta}^{(j')}$ to maintain the monotonicity of the utility function; this only increases the objective value. Thus, the allocation–utility pair (Q, U) that contains Case 1b is not optimal. □

A.3 Proofs for Scarce Resource

Proof of Lemma 2. Taking the second-order derivative gives us

$$\begin{aligned} Q''_{\mathrm{E}} &= (F^{n-1})'' = ((n-1)F^{n-2} \cdot f)' = (n-1)((n-2)F^{n-3} \cdot f^2 + F^{n-2} \cdot f') \\ &\geq (n-1)F^{n-3}((n-2)\underline{\beta}_1^2 - F \cdot \beta_2) \geq 0 \end{aligned}$$

when $n \geq N \geq 2 + \frac{\beta_2}{\underline{\beta}_1^2}$. □

Proof of Proposition 3. By Theorem 2, there exists a partition of the type space $\{(\underline{\theta}^{(j)}, \bar{\theta}^{(j)})\}_{j=1}^{\infty}$ such that each interval belongs to one of the three cases. It is sufficient to show that the order of the three cases on the type space cannot be changed in the optimal non-coordination mechanism.

First we show that for j such that interval j is in the no-tension region, it is optimal for all intervals containing types below $\underline{\theta}^{(j)}$ to be in the no-tension region as well. The main reason is that by the convexity of the efficient allocation rule, for any type $\theta \leq \underline{\theta}^{(j)}$, $Q'_{\mathrm{E}}(\theta) \leq Q'_{\mathrm{E}}(\underline{\theta}^{(j)}) \leq \eta$. Therefore, if we set $U_\alpha(\theta) = Q_\alpha(\theta) = Q'_{\mathrm{E}}(\theta)$, the resulting non-coordination mechanism is feasible and trivially maximizes the objective value.

Let $\theta^{(1)}$ be the supremum of the set of all types θ lying in the no-tension region. The argument in previous paragraph shows that the whole interval $(\underline{\theta}, \theta^{(1)})$ is in the no-tension region. Moreover, by Theorem 2, $Q_\alpha(\theta^{(1)}) = U_\alpha(\theta^{(1)}) = Q_{\mathrm{E}}(\theta^{(1)})$, and for any $\theta \geq \theta^{(1)}$ we have $U'_\alpha(\theta) = \eta$. Now we consider two cases:

- If $Q'_{\mathrm{E}}(\theta^{(1)}) \geq \eta$, then by the convexity of the efficient allocation rule, $Q_{\mathrm{E}}(\theta) > U_\alpha(\theta)$ for any type $\theta > \theta^{(1)}$, which implies that

$$\int_{\underline{\theta}^{(j)}}^{\bar{\theta}^{(j)}} U_\alpha(\theta)\,\mathrm{d}F(\theta) < \int_{\underline{\theta}^{(j)}}^{\bar{\theta}^{(j)}} Q_{\mathrm{E}}(\theta)\,\mathrm{d}F(\theta)$$

  for any interval j with types above $\theta^{(1)}$. In this case, $\theta^{(1)} = \theta^{(2)}$ and the no-effort region does not exist.

- If $Q'_{\mathrm{E}}(\theta^{(1)}) < \eta$, then $Q_{\mathrm{E}}(\theta) < U_\alpha(\theta)$ for any type θ sufficiently close to $\theta^{(1)}$. Therefore, for j such that $\underline{\theta}^{(j)} = \theta^{(1)}$, interval j must be in the no-effort region. Let $\theta^{(2)} = \bar{\theta}^{(j)}$. Note that for the integration constraint to be satisfied in interval j, we must have $Q_{\mathrm{E}}(\theta^{(2)}) \geq U_\alpha(\theta^{(2)})$ and $Q'_{\mathrm{E}}(\theta^{(2)}) \geq \eta$. Therefore, for any type $\theta > \theta^{(2)}$, we have $Q_{\mathrm{E}}(\theta) > U_\alpha(\theta)$; hence any interval above type $\theta^{(2)}$ is in the efficient region. □

Proof of Theorem 3. By Lemma 2, for sufficiently large n, the efficient allocation rule is convex. Therefore, the interim allocation rule of the optimal non-coordination mechanism takes the form described in Proposition 3.

Let $Q_{\alpha,n}(\theta)$ and the $Q_{\mathrm{E},n}(\theta)$ be the optimal interim allocation rule in a non-coordination mechanism and efficient allocation rule respectively with $n < \infty$ agents. For any finite n, we have that

$$\frac{1}{n} \geq \int_{\theta_n^{(1)}}^{\bar{\theta}} Q_{\mathrm{E},n}(\theta)\,\mathrm{d}F(\theta) \geq \int_{\theta_n^{(1)}}^{\bar{\theta}} \left(\eta \cdot (\theta - \theta_n^{(1)}) + Q_{\mathrm{E},n}(\theta_n^{(1)})\right)\,\mathrm{d}F(\theta).$$

The first inequality holds because the ex-ante probability that a given agent gets the item is at most $\frac{1}{n}$, and the second inequality holds because the efficient allocation majorizes the interim allocation, since the latter is again at least the interim utility. Since $Q_{\mathrm{E},n}(\theta_n^{(1)})$ is non-negative, we have that

$$\int_{\theta_n^{(1)}}^{\bar{\theta}} (\theta - \theta_n^{(1)})\,\mathrm{d}F(\theta) \leq \frac{1}{n\eta}$$

for any n. Note that $\frac{1}{n\eta}$ converges to 0 as n approaches infinity. In order for the inequality to hold, $\theta_n^{(1)}$ must also converge to $\bar{\theta}$ as n approaches infinity. □

Proof of Theorem 4. First we present Lemma 7, whose proof is given later in this section. Lemma 7 says that given the efficient allocation rule, the sum of the expected utilities of the agents is small compared to the best scenario, i.e., the scenario in which the highest type gets the item without exerting effort, which is 1.

Lemma 7. *For any $\epsilon > 0$, there exists $N_0 \geq 1$ such that for any $n \geq N_0$, we have $n \cdot \mathbf{E}_{\theta\sim F}[U_{\mathrm{E},n}(\theta)] \leq 1 - \frac{1}{e} + \epsilon$.*

Intuitively, this means that competition is high among agents with sufficiently high types. Thus agents with high types need to exert high effort to ensure a large allocation, leading to a utility loss relative to the first-best utility. By applying Lemma 7, we obtain an upper bound on the performance of the WTA contest. That is, for any $\epsilon > 0$, there exists N_0 such that for any $n \geq N_0$, we have

$$\begin{aligned} n \cdot V_\alpha(Q_{\mathrm{E},n}) &= n\alpha \cdot \mathbf{E}_{\theta\sim F}[\theta \cdot Q_{\mathrm{E},n}(\theta)] + n(1-\alpha) \cdot \mathbf{E}_{\theta\sim F}[U_{\mathrm{E},n}(\theta)] \\ &\leq \alpha \cdot \bar{\theta} + (1-\alpha) \cdot \left(1 - \frac{1}{e} + \epsilon\right). \end{aligned}$$

The inequality holds by Lemma 7 and the fact that the upper bound on the type of the agent winning the item is $\bar{\theta}$.

Next we provide a lower bound on the performance of the optimal contest. In particular, for any n large enough, consider a feasible allocation

$$Q_n(\theta) = \begin{cases} Q_{\mathrm{E},n}(\theta) & \text{if } \theta \leq \hat{\theta}_n, \\ \eta \cdot (\theta - \hat{\theta}_n) + Q_{\mathrm{E},n}(\hat{\theta}_n) & \text{if } \theta > \hat{\theta}_n, \end{cases}$$

such that $\mathbf{E}_{\theta\sim F}[Q_n(\theta)] = \mathbf{E}_{\theta\sim F}[Q_{\mathrm{E},n}(\theta)] = \frac{1}{n}$. Let $U_n(\theta) = Q_n(\theta)$. Notice that (Q_n, U_n) satisfies the (IC) constraints. Moreover, $Q_n(\theta)$ induces no effort and hence $\mathbf{E}_{\theta\sim F}[U_n(\theta)] = \frac{1}{n}$. In the following lemma (proved at the end of this section), we show that the matching efficiency of the given allocation rule converges to the optimal welfare when the number of agents is sufficiently large.

Lemma 8. *For any $\epsilon > 0$, there exists N_1 such that for any $n \geq N_1$, $n \cdot \mathbf{E}_{\theta\sim F}[\theta \cdot Q_n(\theta)] \geq \bar{\theta} - \epsilon$.*

Therefore, there exists N_1 such that for any $n \geq N_1$, we have

$$\begin{aligned} n \cdot V_\alpha(Q_{\alpha,n}) &\geq n \cdot \alpha \mathbf{E}_{\theta\sim F}[\theta \cdot Q_n(\theta)] + n \cdot (1-\alpha)\mathbf{E}_{\theta\sim F}[U_n(\theta)] \\ &\geq \alpha(\bar{\theta} - \epsilon) + 1 - \alpha. \end{aligned}$$

Finally, for any $\epsilon > 0$, letting $N = \max\{N_0, N_1\}$, we can combine the inequalities above to obtain

$$\frac{V_\alpha(Q_{\alpha,n})}{V_\alpha(Q_{\mathrm{E},n})} \geq \frac{(\bar{\theta} - \epsilon) \cdot \alpha + 1 - \alpha}{\bar{\theta} \cdot \alpha + (1-\alpha)(1 - \frac{1}{e} + \epsilon)}$$

for any $n \geq N$. □

Proof of Lemma 7. Let n be a sufficiently large number so that $Q_{\mathrm{E},n}$ is convex, and let $\theta_n^\dagger$ be the cutoff type such that in the incentive-compatible implementation of efficient allocation, agents with any type $\theta > \theta_n^\dagger$ exert costly effort, i.e., $Q'_{\mathrm{E},n}(\theta_n^\dagger) = \eta$. In other words, $(n-1) \cdot F^{n-2}(\theta_n^\dagger) \cdot f(\theta_n^\dagger) = \eta$. Rearranging the terms, we have

$$F^{n-2}(\theta_n^\dagger) = \frac{\eta}{(n-1) \cdot f(\theta_n^\dagger)}.$$

Note that by Assumption 6, the right-hand side is bounded below by $\frac{\eta}{(n-1)\cdot\bar{\beta}_1}$. Therefore, for any $\epsilon_0 > 0$, there exists N_0 such that for any $n \geq N_0$, we have

$$F(\theta_n^\dagger) \geq \left(\frac{\eta}{(n-1)\cdot\bar{\beta}_1}\right)^{\frac{1}{n-2}} \geq 1 - \epsilon_0.$$

Since the density is bounded below by $\underline{\beta}_1$, we have that $\theta_n^\dagger \geq \bar{\theta} - \frac{\epsilon_0}{\underline{\beta}_1}$. For any $\epsilon_1 > 0$, let N_1 be an integer such that $\frac{\eta}{(n-1)\cdot f(\theta_n^\dagger)} \leq \epsilon_1$ for any $n \geq N_1$. The expected utility of an agent with type $\bar{\theta}$ is

$$U_{\mathrm{E},n}(\bar{\theta}) = F^{n-1}(\theta_n^\dagger) + \eta(\bar{\theta} - \theta_n^\dagger) \leq F^{n-2}(\theta_n^\dagger) + \frac{\eta \cdot \epsilon_0}{\underline{\beta}_1} \leq \epsilon_1 + \frac{\eta \cdot \epsilon_0}{\underline{\beta}_1}.$$

Let $\theta_n^\ddagger$ be the type such that $F(\theta_n^\ddagger) = 1 - \frac{1}{n}$. There exists N_2 such that $\theta_n^\ddagger \geq \theta_n^\dagger$ for any $n \geq N_2$. For any $\epsilon > 0$, let $\epsilon_1 = \frac{\epsilon}{2}$, $\epsilon_0 = \frac{\epsilon\underline{\beta}_1}{2\eta}$, and $N = \max\{N_0, N_1, N_2\}$. For any $n \geq N$, the expected effort of any agent is at least his effort from types above $\theta_n^\ddagger$, which is bounded below by

$$(1 - F(\theta_n^\ddagger)) \cdot (Q_{\mathrm{E},n}(\theta_n^\ddagger) - U_{\mathrm{E},n}(\bar{\theta})) \geq \frac{1}{n}\left(\frac{1}{e} - \epsilon_1 + \frac{\eta \cdot \epsilon_0}{\underline{\beta}_1}\right) = \frac{1}{n}\left(\frac{1}{e} - \epsilon\right).$$

Since the item is always allocated in equilibrium, the total utility is

$$n \cdot \mathbf{E}_{\theta\sim F}[U_{\mathrm{E},n}(\theta)] \leq 1 - \frac{1}{e} + \epsilon. \qquad \square$$

Proof of Lemma 8. Note that compared to the efficient allocation $Q_{\mathrm{E},n}$, the chosen allocation rule Q_n only randomizes the allocation for types between $\hat{\theta}_n$ and $\bar{\theta}$. Therefore, we have

$$n \cdot \mathbf{E}_{\theta\sim F}[\theta \cdot Q_n(\theta)] \geq n \cdot \mathbf{E}_{\theta\sim F}[\theta \cdot Q_{\mathrm{E},n}(\theta)] - (\bar{\theta} - \hat{\theta}_n).$$

As in the proof of Theorem 3, we can show that $\lim_{n\to\infty} \hat{\theta}_n = \bar{\theta}$. By taking the limit of the

above inequality, we have that

$$\lim_{n\to\infty} n \cdot \mathbf{E}_{\theta\sim F}[\theta \cdot Q_n(\theta)] \geq \lim_{n\to\infty} n \cdot \mathbf{E}_{\theta\sim F}[\theta \cdot Q_{\mathrm{E},n}(\theta)] = \bar{\theta}.$$

Thus, for any $\epsilon > 0$, there exists N_1 such that for any $n \geq N_1$, $n \cdot \mathbf{E}_{\theta\sim F}[\theta \cdot Q_n(\theta)] \geq \bar{\theta} - \epsilon$. $\square$

A.4 Proof of Large Scale Economy

It is tempting to conjecture that when z is large enough, $Q_{\mathrm{E},z}(\theta)$ has an S shape (i.e., it is convex for small θ and concave for large θ), which would naturally imply the order of the intervals as stated in our result. However, this is not true in general.[24] To circumvent this inconvenience, note that for any small constant ϵ_0, when z is large enough, the interim efficient allocation has small slope (smaller than the marginal cost of effort η) outside the small interval $(\theta_c - \epsilon_0, \theta_c + \epsilon_0)$ centered at θ_c. Moreover, since the value of the efficient allocation changes a lot in this small interval, the agents will exert high effort in equilibrium if the items are allocated efficiently, leading to low expected utility for types around θ_c. We show that in the optimal contest, the principal randomizes the allocation around θ_c. In particular, the no-effort region, where the allocation is randomized, will cover the whole interval $(\theta_c - \epsilon_0, \theta_c + \epsilon_0)$. Since the derivative of the efficient allocation outside this region is at most η, the principal's objective value is maximized by the efficient allocation. We provide the formal proof below.

Proof of Theorem 5. Since the distribution is continuous, the probability there is a tie for any two distinct types is 0. Therefore, given the scale parameter z, the interim efficient allocation is

$$Q_{\mathrm{E},z}(\theta) = \mathbf{Pr}\big[\theta_{(nz-kz:nz-1)} \leq \theta\big] = \sum_{j=0}^{zk-1} \binom{zn-1}{j} \cdot (1-F(\theta))^j \cdot (F(\theta))^{zn-1-j},$$

where $\theta_{(nz-kz:nz-1)}$ is the $(nz-kz)$th order statistic, i.e., the $(nz-kz)$th smallest value in a sample of $nz-1$ observations, and the binomial coefficient $\binom{n}{k}$ is defined by $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

[24] The second-order derivative of the allocation is

$$Q''_{\mathrm{E},z}(\theta) = (zn-1) \cdot \binom{zn-2}{zk-1} (1-F(\theta))^{zk-2} \cdot (F(\theta))^{z(n-k)-2} \cdot \big(f^2(\theta)(z(n-k)-1-(zn-2)F(\theta)) + f'(\theta)(1-F(\theta))F(\theta)\big).$$

No matter how large the parameter z is, for types within $(\theta_c - \epsilon_0, \theta_c + \epsilon_0)$, the sign of the second-order derivative may change multiple times.

Recall that θ_c is the cutoff type such that $1-F(\theta_c)=\frac{k}{n}$. The derivative of the allocation is

$$Q'_{\mathrm{E},z}(\theta)=f(\theta)\cdot(zn-1)\cdot\binom{zn-2}{zk-1}(1-F(\theta))^{zk-1}\cdot(F(\theta))^{z(n-k)-1}.$$

Note that $\binom{zn-2}{zk-1}(1-F(\theta))^{zk-1}\cdot(F(\theta))^{z(n-k)-1}$ is the probability that the binomial random variable $B(zn-2,1-F(\theta))$ equals $zk-1$. When $1-F(\theta)<\frac{k}{n}$, this probability becomes exponentially small as zn increases, which implies that $\lim_{z\to\infty}Q'_{\mathrm{E},z}(\theta)=0$. Therefore, for any $\epsilon_0>0$, there exists Z_0 such that for any $z\geq Z_0$, for any type $\theta\notin[\theta_c-\epsilon_0,\theta_c+\epsilon_0]$,

$$Q'_{\mathrm{E},z}(\theta)\leq\eta.$$

Again by Hoeffding's inequality, for any $\epsilon_1>0$, there exists Z_1 such that for any $z\geq Z_1$,

$$Q_{\mathrm{E},z}(\theta)\leq\epsilon_1$$

for any type $\theta\leq\theta_c-\epsilon_0$ and

$$Q_{\mathrm{E},z}(\theta)\geq 1-\epsilon_1$$

for any type $\theta\geq\theta_c+\epsilon_0$. Intuitively, this is because $\lim_{z\to\infty}Q_{\mathrm{E},z}(\theta)$ is a step function, i.e.,

$$\lim_{z\to\infty}Q_{\mathrm{E},z}(\theta)=\begin{cases}0 & \text{if } \theta<\theta_c,\\ 1 & \text{if } \theta\geq\theta_c.\end{cases}$$

Let $\tilde{\theta}^{(1)}\triangleq\theta_c-\epsilon_0-\sqrt{\frac{8\epsilon_0\bar{\beta}_1}{\eta\underline{\beta}_{-1}}}$.

Lemma 9. *For sufficiently large z, in the optimal contest $(Q_{\alpha,z},U_{\alpha,z})$, we have $U_{\alpha,z}(\tilde{\theta}^{(1)})>Q_{\mathrm{E},z}(\tilde{\theta}^{(1)})$.*

We defer the proof of the lemma to the end of this section. Note that in the optimal contest $(Q_{\alpha,z},U_{\alpha,z})$, $U_{\alpha,z}(\tilde{\theta}^{(1)})>Q_{\mathrm{E},z}(\tilde{\theta}^{(1)})$ implies that type $\tilde{\theta}^{(1)}$ must belong to a no-effort interval. Let $\theta^{(1)}<\tilde{\theta}^{(1)}<\theta^{(2)}$ be the endpoints of this no-effort interval. Let Θ_+ be the set of types in $(\theta^{(1)},\theta^{(2)})$ such that $Q_{\mathrm{E},z}(\theta)>\hat{Q}_{\alpha,z}(\theta)$, and let Θ_- be the set of types in $(\theta^{(1)},\theta^{(2)})$ such that $Q_{\mathrm{E},z}(\theta)<Q_{\alpha,z}(\theta)$. Since the integration constraint binds within $(\theta^{(1)},\theta^{(2)})$, we have that

$$\int_{\Theta_+}(Q_{\mathrm{E},z}(\theta)-Q_{\alpha,z}(\theta))\,\mathrm{d}F(\theta)+\int_{\Theta_-}(Q_{\mathrm{E},z}(\theta)-Q_{\alpha,z}(\theta))\,\mathrm{d}F(\theta)=0.$$

Note that

$$\int_{\Theta_-} (Q_{\mathrm{E},z}(\theta) - Q_{\alpha,z}(\theta))\,\mathrm{d}F(\theta) \leq -\int_{\theta^{(1)}}^{\theta_c-\epsilon_0} (\eta(\theta-\theta^{(1)})-\epsilon_1)\,\mathrm{d}F(\theta)$$
$$\leq -\underline{\beta}_1 \cdot \left(\frac{\eta}{2}\cdot(\theta_c-\epsilon_0-\theta^{(1)})^2 - \epsilon_1\cdot(\theta_c-\epsilon_0-\theta^{(1)})\right).$$

Similarly,

$$\int_{\Theta_+} (Q_{\mathrm{E},z}(\theta) - Q_{\alpha,z}(\theta))\,\mathrm{d}F(\theta) \leq \int_{\theta_c-\epsilon_0}^{\theta^{(2)}} 1\,\mathrm{d}F(\theta) \leq \bar{\beta}_1\cdot(\theta^{(2)}-\theta_c+\epsilon_0).$$

Combining the inequalities above, for sufficiently small $\epsilon_1 \leq \frac{\eta}{4}(\theta_c-\epsilon_0-\theta^{(1)})$, we must have

$$\theta^{(2)} \geq \theta_c-\epsilon_0+\frac{\eta\cdot\underline{\beta}_1}{4\bar{\beta}_1}\cdot(\theta_c-\epsilon_0-\theta^{(1)})^2 \geq \theta_c+\epsilon_0.$$

We obtain the last inequality simply by substituting the bound for $\theta^{(1)}$. This implies that in the optimal contest, the no-effort region $(\theta^{(1)},\theta^{(2)})$ covers the whole interval $(\theta_c-\epsilon_0,\theta_c+\epsilon_0)$. Note that since the derivative of the efficient allocation outside the no-effort region $(\theta^{(1)},\theta^{(2)})$ is at most η, the principal's objective is maximized by the efficient allocation. In particular, let $\theta^{(3)} \geq \theta^{(2)}$ be the type such that the linear extension of the utility function within the no-tension region intersects the efficient allocation rule. Then the interval $(\theta^{(2)},\theta^{(3)})$ is the efficient region, and the union of $(\underline{\theta},\theta^{(1)})$ and $(\theta^{(3)},\bar{\theta})$ is the no-tension region. □

Proof of Lemma 9. It is sufficient to show that any contest $(\tilde{Q}_{\alpha,z},\tilde{U}_{\alpha,z})$ such that $\tilde{U}_{\alpha,z}(\tilde{\theta}^{(1)}) \leq Q_{\mathrm{E},z}(\tilde{\theta}^{(1)})$ cannot be an optimal contest. We prove this by contradiction: given such a contest, we construct a contest $\hat{Q}_{\alpha,z},\hat{U}_{\alpha,z}$ that yields a higher objective value.

Let $\epsilon_0,\epsilon_1,\epsilon_2>0$ be any numbers such that the following hold:[25]

$$0<\epsilon_0\leq\min\left\{\frac{\underline{\beta}_1}{10\eta\cdot\bar{\beta}_1},\epsilon_2^4\right\},\qquad \epsilon_0+2\sqrt{\frac{8\epsilon_0\bar{\beta}_1}{\eta\underline{\beta}_1}}\leq\epsilon_2,$$
$$0<\epsilon_1\leq\min\{0.01,\epsilon_2^4\},\qquad 0<\epsilon_2<\frac{\underline{\beta}_1}{10\eta\cdot\bar{\beta}_1},$$
$$\alpha\bar{\beta}_1\cdot\left((\epsilon_2+\epsilon_0)^2\cdot\frac{\bar{\beta}_1}{\underline{\beta}_1}+\epsilon_0+\epsilon_2\right)^2<\frac{1}{2\eta}(1-\alpha)\cdot\left(\eta\left(\epsilon_2-\epsilon_0-\sqrt{\frac{8\epsilon_0\bar{\beta}_1}{\eta\underline{\beta}_1}}\right)-\epsilon_1\right).$$

Let $\hat{\theta}^{(1)}\triangleq\theta_c-\epsilon_2$. By our choice of ϵ_0, we have $\hat{\theta}^{(1)}<\tilde{\theta}^{(1)}$.

[25]Notice that these inequalities can hold at the same time: if one chooses ϵ_0 and ϵ_1 that are "small" compared to ϵ_2, for example, $\epsilon_0=o(\epsilon_2^4)$ and $\epsilon_1=o(\epsilon_2^4)$, then the last inequality holds because the left-hand side is of higher order than the right-hand side.

Consider a contest $(\hat{Q}_{\alpha,z}, \hat{U}_{\alpha,z})$ characterized by three cutoffs $\hat{\theta}^{(1)} < \hat{\theta}^{(2)} \leq \hat{\theta}^{(3)}$ such that the union of $(\underline{\theta}, \hat{\theta}^{(1)})$ and $(\hat{\theta}^{(3)}, \bar{\theta})$ is the no-tension region, $(\hat{\theta}^{(1)}, \hat{\theta}^{(2)})$ is the no-effort region, and $(\hat{\theta}^{(2)}, \hat{\theta}^{(3)})$ is the efficient region.

Step 1: In this step, we will show that if $\hat{\theta}^{(1)}$ is chosen so that $\theta_c - \frac{\underline{\beta}_1}{10\eta \cdot \bar{\beta}_1} \leq \hat{\theta}^{(1)}$,[26] then the integration constraint for the no-effort interval imposes an upper bound on the length of the no-effort interval, i.e., $\hat{\theta}^{(2)} \leq \tilde{\theta}$, where $\tilde{\theta} \triangleq \theta_c + \epsilon_0 + \frac{2\eta \cdot \bar{\beta}_1}{\underline{\beta}_1} \cdot (\theta_c + \epsilon_0 - \hat{\theta}^{(1)})^2$.

Let $\hat{\Theta}_+$ be the set of types in $(\hat{\theta}^{(1)}, \hat{\theta}^{(2)})$ such that $Q_{\mathrm{E},z}(\theta) > \hat{Q}_{\alpha,z}(\theta)$, and let $\hat{\Theta}_-$ be the set of types in $(\hat{\theta}^{(1)}, \hat{\theta}^{(2)})$ such that $Q_{\mathrm{E},z}(\theta) < \hat{Q}_{\alpha,z}(\theta)$. Since the integration constraint binds within $(\hat{\theta}^{(1)}, \hat{\theta}^{(2)})$, we have that

$$\begin{aligned} 0 &= \int_{\hat{\Theta}_+} (Q_{\mathrm{E},z}(\theta) - \hat{Q}_{\alpha,z}(\theta)) \, \mathrm{d}F(\theta) + \int_{\hat{\Theta}_-} (Q_{\mathrm{E},z}(\theta) - \hat{Q}_{\alpha,z}(\theta)) \, \mathrm{d}F(\theta) \\ &\geq \int_{\theta_c + \epsilon_0}^{\hat{\theta}^{(2)}} (1 - 2\epsilon_1 - \eta(\theta - \hat{\theta}^{(1)})) \, \mathrm{d}F(\theta) - \int_{\hat{\theta}^{(1)}}^{\theta_c + \epsilon_0} \eta(\theta - \hat{\theta}^{(1)}) \, \mathrm{d}F(\theta). \end{aligned}$$

By our choice of $\hat{\theta}^{(1)}$ and ϵ_0, ϵ_1, we have $1 - 2\epsilon_1 - \eta(\theta - \hat{\theta}^{(1)}) \geq \frac{1}{2}$ for any type $\theta \leq \tilde{\theta}$. Therefore,

$$\begin{aligned} &\int_{\theta_c + \epsilon_0}^{\tilde{\theta}} (1 - 2\epsilon_1 - \eta(\theta - \hat{\theta}^{(1)})) \, \mathrm{d}F(\theta) - \int_{\hat{\theta}^{(1)}}^{\theta_c + \epsilon_0} \eta(\theta - \hat{\theta}^{(1)}) \, \mathrm{d}F(\theta) \\ \geq &\frac{\underline{\beta}_1}{2}(\tilde{\theta} - \theta_c - \epsilon_0) - \eta \cdot \bar{\beta}_1 (\theta_c + \epsilon_0 - \hat{\theta}^{(1)})^2 \geq 0. \end{aligned}$$

Combining the above two inequalities, we get the desired bound on $\hat{\theta}^{(2)}$.

Step 2: Next we utilize the upper bound to show that the objective value from the contest $\hat{Q}_{\alpha,z}, \hat{U}_{\alpha,z}$ is higher than that from the contest $\tilde{Q}_{\alpha,z}, \tilde{U}_{\alpha,z}$ with $\tilde{U}_{\alpha,z}(\tilde{\theta}^{(1)}) \leq \tilde{Q}_{\alpha,z}(\tilde{\theta}^{(1)})$. Note that $Q_{\mathrm{E},z}$ and $\hat{Q}_{\alpha,z}(\theta)$ coincide at any type θ outside the no-effort region. Therefore, the loss in efficiency compared to the efficient allocation rule is

$$\begin{aligned} &\alpha \cdot \int_{\hat{\theta}^{(1)}}^{\hat{\theta}^{(2)}} \theta \cdot Q_{\mathrm{E},z} \, \mathrm{d}F(\theta) - \alpha \cdot \int_{\hat{\theta}^{(1)}}^{\hat{\theta}^{(2)}} \theta \cdot \hat{Q}_{\alpha,z}(\theta) \, \mathrm{d}F(\theta) \\ =&\alpha \cdot \int_{\hat{\Theta}_+} \theta \cdot (Q_{\mathrm{E},z}(\theta) - \hat{Q}_{\alpha,z}(\theta)) \, \mathrm{d}F(\theta) - \alpha \cdot \int_{\hat{\Theta}_-} \theta \cdot (Q_{\mathrm{E},z}(\theta) - \hat{Q}_{\alpha,z}(\theta)) \, \mathrm{d}F(\theta) \\ \leq&\alpha \cdot (\hat{\theta}^{(2)} - \hat{\theta}^{(1)}) \cdot \int_{\hat{\Theta}_+} (Q_{\mathrm{E},z}(\theta) - \hat{Q}_{\alpha,z}(\theta)) \, \mathrm{d}F(\theta) \\ \leq&\alpha \cdot (\hat{\theta}^{(2)} - \hat{\theta}^{(1)}) \cdot (F(\hat{\theta}^{(2)}) - F(\hat{\theta}^{(1)})) \leq \alpha \bar{\beta}_1 \cdot (\hat{\theta}^{(2)} - \hat{\theta}^{(1)})^2, \end{aligned}$$

[26] Such a choice is possible because by the choice of $\epsilon_0, \epsilon_1, \epsilon_2$, we have $\theta_c - \frac{\underline{\beta}_1}{10\eta \cdot \bar{\beta}_1} \leq \tilde{\theta}^{(1)}$.

where the second inequality holds because the interim allocations are bounded within $[0,1]$, and the last inequality holds by the continuity assumption (Assumption 6).

Moreover, note that the utility $\tilde{U}_{\alpha,z}$ increases at a rate of at most η after type $\tilde{\theta}^{(1)}$, while the utility $\hat{U}_{\alpha,z}$ increases at a rate of η within the interval $(\tilde{\theta}^{(1)}, \hat{\theta}^{(3)})$. Therefore, the gain in utility is at least

$$\begin{aligned}
&(1-\alpha)\cdot\int_{\tilde{\theta}^{(1)}}^{\hat{\theta}^{(3)}}\hat{U}_{\alpha,z}(\theta)\,\mathrm{d}F(\theta)-(1-\alpha)\cdot\int_{\tilde{\theta}^{(1)}}^{\hat{\theta}^{(3)}}\tilde{U}_{\alpha,z}(\theta)\,\mathrm{d}F(\theta)\\
\geq&(1-\alpha)\cdot(F(\hat{\theta}^{(3)})-F(\tilde{\theta}^{(1)}))\cdot(\hat{U}_{\alpha,z}(\tilde{\theta}^{(1)})-\tilde{U}_{\alpha,z}(\tilde{\theta}^{(1)}))\\
\geq&(1-\alpha)\cdot(F(\hat{\theta}^{(3)})-F(\tilde{\theta}^{(1)}))\cdot(\eta\cdot(\tilde{\theta}^{(1)}-\hat{\theta}^{(1)})-\epsilon_1)\\
\geq&\frac{1}{2\eta}(1-\alpha)\cdot\underline{\beta}_1\cdot(\eta\cdot(\tilde{\theta}^{(1)}-\hat{\theta}^{(1)})-\epsilon_1).
\end{aligned}$$

Since the matching efficiency in the contest $(\tilde{Q}_{\alpha,z},\tilde{U}_{\alpha,z})$ is bounded above by the efficient allocation rule, combining the inequalities, we have that

$$\begin{aligned}
&\mathrm{Obj}_\alpha(\tilde{Q}_{\alpha,z},\tilde{U}_{\alpha,z})-\mathrm{Obj}_\alpha(\hat{Q}_{\alpha,z},\hat{U}_{\alpha,z})\\
&\leq\alpha\bar{\beta}_1\cdot(\hat{\theta}^{(2)}-\hat{\theta}^{(1)})^2-\frac{1}{2\eta}(1-\alpha)\cdot\underline{\beta}_1\cdot(\eta\cdot(\tilde{\theta}^{(1)}-\hat{\theta}^{(1)})-\epsilon_1)\\
&\leq\alpha\bar{\beta}_1\cdot\left((\epsilon_2+\epsilon_0)^2\cdot\frac{\bar{\beta}_1}{\underline{\beta}_1}+\epsilon_0+\epsilon_2\right)^2-\frac{1}{2\eta}(1-\alpha)\cdot\left(\eta\left(\epsilon_2-\epsilon_0-\sqrt{\frac{8\epsilon_0\bar{\beta}_1}{\eta\underline{\beta}_1}}\right)-\epsilon_1\right)<0.
\end{aligned}$$

The last inequality comes from the choice of $\epsilon_0,\epsilon_1,\epsilon_2$. Therefore, the contest $(\tilde{Q}_{\alpha,z},\tilde{U}_{\alpha,z})$ is not optimal. □

B Optimality of Monotone Allocations

Note that there exist direct mechanisms that implement non-monotone interim allocations (see Example 3). This is because the agent's utility does not satisfy the single-crossing property in when the mechanism provides randomized signal recommendation. In this section, we show that under an additional assumption, which we call "no concave crossing," non-monotone interim allocation rules are not optimal for the objective of maximizing weighted average between welfare and agents utilities.

Assumption 7 (no-concave-crossing)**.** *For any agent i, consider any allocation and deterministic signal recommendation (x_i,s_i), and any allocation and stochastic signal recommen-*

dation (x_i', D_i)*, if there exist two types* $\hat{\theta}_i < \hat{\theta}_i'$ *such that*

$$u_i(x_i, s_i, \hat{\theta}_i) \le \mathbf{E}_{s\sim D_i}\left[u_i(x_i', s, \hat{\theta}_i)\right]$$
$$u_i(x_i, s_i, \hat{\theta}_i') \le \mathbf{E}_{s\sim D_i}\left[u_i(x_i', s, \hat{\theta}_i')\right],$$

then for any $\theta_i \in [\hat{\theta}_i, \hat{\theta}_i']$*, we have the following* $u_i(x_i, s_i, \theta_i) \le \mathbf{E}_{s\sim D_i}[u_i(x_i', s, \theta_i)]$.

Note that the above condition only requires no-concave-crossing between the utility curve generated by a deterministic recommendation and the one generated by a general randomized recommendation. The condition usually fails if we consider two randomized recommendations. It is always satisfied if the cost function is linear or quadratic.

For any non-monotone allocation, we can consider its monotone rearrangement which swaps the allocation among types such that the resulting allocation is monotone. The following assumption assumes that such rearrangement always benefits the principal. This assumption implies that the principal benefits from allocating the resource to higher types than lower types.

Assumption 8 (assortative matching)**.** *For any interim allocation* $\boldsymbol{Q}$ *and* $Q^\dagger$ *that is a monotone rearrangement of* $\boldsymbol{Q}$*, we have*

$$\sum_{i\in[n]} \mathbf{E}_{\theta_i\sim F_i}[W_i(Q_i(\theta_i), \theta_i)] \le \sum_{i\in[n]} \mathbf{E}_{\theta_i\sim F_i}\left[W_i(Q_i^\dagger(\theta_i), \theta_i)\right].$$

Finally, to simplify the exposition, we assume that the type distribution has finite support.

Assumption 9. *The type space of agent* i *is discrete and finite; that is, it has the form* $\Theta_i = \{\hat{\theta}_i^{(0)}, \dots, \hat{\theta}_i^{(m)}\}$*, with* $\hat{\theta}_i^{(0)} < \cdots < \hat{\theta}_i^{(m)}$.

Theorem 6. *Under Assumptions 7, 8 and 9, for any interim allocation–utility pair* $(\boldsymbol{Q}, \boldsymbol{U})$ *that is implementable by a direct mechanism, there exists* $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ *with monotone* $\boldsymbol{Q}^\dagger$ *that is implementable by a non-coordination mechanism and yields a weakly higher objective value.*

Proof. Let $\boldsymbol{Q}^\dagger$ be a monotonic rearrangement of $\boldsymbol{Q}$ that is feasible and weakly improves matching efficiency. We will construct a $\boldsymbol{U}^\dagger$ such that $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ is implementable by a non-coordination mechanism and show that $U_i^\dagger(\hat{\theta}_i^{(k)}) \ge U_i(\hat{\theta}_i^{(k)})$ for all i and all $k \in \{0, \dots, m\}$, i.e., $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ weakly improves all agents' utilities for all types compared to the mechanism $(\boldsymbol{Q}, \boldsymbol{U})$.

We prove the above claim by induction.

$k = 0$**:** Let $U_i^\dagger(\hat{\theta}_i^{(0)}) = Q_i^\dagger(\hat{\theta}_i^{(0)})$. Since $\boldsymbol{Q}^\dagger$ is a monotone rearrangement of $\boldsymbol{Q}$, and $\hat{\theta}_i^{(0)}$ is the lowest type, we have $Q_i(\hat{\theta}_i^{(0)}) \leq Q_i^\dagger(\hat{\theta}_i^{(0)})$. By incentive compatibility, we have $Q_i(\hat{\theta}_i^{(0)}) \geq U_i(\hat{\theta}_i^{(0)})$. Hence we have $U_i^\dagger(\hat{\theta}_i^{(0)}) \geq U_i(\hat{\theta}_i^{(0)})$.

For any $k \geq 1$**:**

Case 1: suppose $Q_i^\dagger(\hat{\theta}_i^{(k)}) \geq Q_i(\hat{\theta}_i^{(k)})$. We can apply the same argument as in Theorem 1 to construct $U_i^\dagger(\hat{\theta}_i^{(k)})$ so that the expected cost of any agent i with type $\hat{\theta}_i^{(k)}$ is weakly lower than the expected cost under $(\boldsymbol{Q}, \boldsymbol{U})$, the direct mechanism we start with, i.e., $U_i^\dagger(\hat{\theta}_i^{(k)}) \geq U_i(\hat{\theta}_i^{(k)})$.

Case 2: suppose $Q_i^\dagger(\hat{\theta}_i^{(k)}) < Q_i(\hat{\theta}_i^{(k)})$.

Let $s_i^{(k)}$ be the signal that type $\hat{\theta}_i^{(k-1)}$ is indifferent between truthfully reporting in the newly constructed mechanism and misreporting as the adjacent higher type $\hat{\theta}_i^{(k)}$ in the newly constructed mechanism, so as to receive$(Q_i^\dagger(\hat{\theta}_i^{(k)}), s_i^{(k)})$, i.e.,

$$U_i^\dagger(\hat{\theta}_i^{(k-1)}) = u_i(\hat{\theta}_i^{(k-1)}; Q_i^\dagger(\hat{\theta}_i^{(k)}), s_i^{(k)}).$$

Let $U_i^\dagger(\hat{\theta}_i^{(k)}) = u_i(\hat{\theta}_i^{(k)}; Q_i^\dagger(\hat{\theta}_i^{(k)}), s_i^{(k)})$. It remains to show that $U_i^\dagger(\hat{\theta}_i^{(k)}) \geq U_i(\hat{\theta}_i^{(k)})$.

Since $\boldsymbol{Q}^\dagger$ is a monotone rearrangement of $\boldsymbol{Q}$, for any i and any type $\hat{\theta}_i^{(k)}$, there must exist $k' > k$ such that $Q_i(\hat{\theta}_i^{(k')}) \leq Q_i^\dagger(\hat{\theta}_i^{(k)})$. Similarly, let $\tilde{s}_i$ be the signal that type $\hat{\theta}_i^{(k-1)}$ is indifferent between truthfully reporting in the newly constructed mechanism and misreporting as the higher type $\hat{\theta}_i^{(k')}$ in the original direct mechanism, so as to receive $(Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i)$, i.e.,

$$U_i^\dagger(\hat{\theta}_i^{(k-1)}) = u_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i). \tag{3}$$

Note that $Q_i^\dagger(\hat{\theta}_i^{(k)}) \geq Q_i(\hat{\theta}_i^{(k')})$ implies $s_i^{(k)} \geq \tilde{s}_i$. Notice that by construction, $s_i^{(k')}$ is the signal recommended to type $\hat{\theta}_i^{(k')}$ under non-coordination mechanism and by single-crossing in utility, we can show that $s_i^{(k')} \geq s_i^{(k)}$. Hence we have $s_i^{(k')} \geq \tilde{s}_i$.

Let $D_i^{(k)}$ be the distribution over signal recommendations to type $\hat{\theta}_i^{(k)}$ under the original mechanism that gives agent i interim utility U_i. We first establish the following two inequalities that describe the preference of type $\hat{\theta}_i^{(k')}$ and type $\hat{\theta}_i^{(k-1)}$.

- type $\hat{\theta}_i^{(k')}$ is weakly better off by receiving the deterministic recommendation $(Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i)$ than misreporting as type $\hat{\theta}_i^{(k)}$ so as to receive $(Q_i(\hat{\theta}_i^{(k)}), D_i^{(k)})$, the potentially stochastic allocation and signal recommendation to type $\hat{\theta}_i^{(k)}$ in the original direct mechanism, i.e.,

  $$u_i(\hat{\theta}_i^{(k')}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i) \geq u_i(\hat{\theta}_i^{(k')}; Q_i(\hat{\theta}_i^{(k)}), D_i^{(k)}).$$

  This is because

$$U_i(\hat{\theta}_i^{(k')}) \geq u_i(\hat{\theta}_i^{(k')}; Q_i(\hat{\theta}_i^{(k)}), D_i^{(k)}), \tag{4}$$

which is implied by incentive compatibility in the original direct mechanism; and

$$u_i(\hat{\theta}_i^{(k')}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i) \geq U_i(\hat{\theta}_i^{(k')}), \tag{5}$$

which is true because (1) by the inductive argument $U_i^\dagger(\hat{\theta}_i^{(k-1)}) \geq U_i(\hat{\theta}_i^{(k-1)})$, and the definition of $\tilde{s}_i$ (Equation 3), $\tilde{s}_i \leq \Sigma(D_i^{(k')}, \hat{\theta}_i^{(k)})$ is smaller than, the certainty equivalent signal for type $\hat{\theta}_i^{(k)}$ given the distribution $D_i^{(k')}$; and (2) $\Sigma(D_i^{(k')}, \hat{\theta}_i^{(k)}) \leq \Sigma(D_i^{(k')}, \hat{\theta}_i^{(k')})$.

- type $\hat{\theta}_i^{(k-1)}$ is weakly better off by receiving the deterministic recommendation $(Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i)$ than misreporting as type $\hat{\theta}_i^{(k)}$ so as to receive $(Q_i(\hat{\theta}_i^{(k)}), D_i^{(k)})$, the potentially stochastic allocation and signal recommendation to type $\hat{\theta}_i^{(k)}$ in the original direct mechanism, i.e.,

  $$u_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i) \geq u_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k)}), D_i^{(k)}).$$

  This is because

  $$u_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i) = U_i^\dagger(\hat{\theta}_i^{(k-1)}) \geq U_i(\hat{\theta}_i^{(k-1)}) \geq u_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k)}), D_i^{(k)}).$$

  The first inequality holds by the induction assumption. The second inequality holds because $u_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k)}), D_i^{(k)})$ is the utility that type $\hat{\theta}_i^{(k-1)}$ obtains by deviating to report type $\hat{\theta}_i^{(k)}$ and always following the signal recommendation.

Combining the two inequalities, since $\hat{\theta}_i^{(k-1)} < \hat{\theta}_i^{(k)} < \hat{\theta}_i^{(k')}$, we immediately obtain from Assumption 7 that

$$u_i(\hat{\theta}_i^{(k)}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i) \geq u_i(\hat{\theta}_i^{(k)}; Q_i(\hat{\theta}_i^{(k)}), D_i^{(k)}). \tag{6}$$

Moreover, since $Q_i^\dagger(\hat{\theta}_i^{(k)}) \geq Q_i(\hat{\theta}_i^{(k')})$, $s_i^{(k)} \geq \tilde{s}_i$, and the utilities of the agent given these two options coincide at type $\hat{\theta}_i^{(k-1)}$, by single-crossing property, we have that for type $\hat{\theta}_i^{(k)}$,

$$u_i(\hat{\theta}_i^{(k)}; Q_i^\dagger(\hat{\theta}_i^{(k)}), s_i^{(k)}) \geq u_i(\hat{\theta}_i^{(k)}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i). \tag{7}$$

By construction, we have

$$U_i^\dagger(\hat{\theta}_i^{(k)}) = u_i(\hat{\theta}_i^{(k)}; Q_i^\dagger(\hat{\theta}_i^{(k)}), s_i^{(k)}),$$

and by definition, we have

$$U_i(\hat{\theta}_i^{(k)}) = u_i(\hat{\theta}_i^{(k)}; Q_i(\hat{\theta}_i^{(k)}), D_i^{(k)}).$$

Combining them with equation 6 and 7, we have

$$U_i^{\dagger}(\hat{\theta}_i^{(k)}) \geq U_i(\hat{\theta}_i^{(k)}). \qquad \square$$