Optimal Screening of Signal-Manipulating Agents via Contests[*]

Yingkai Li[†] Xiaoyun Qiu[‡]

January 15, 2024

Abstract

We study the design of screening mechanisms subject to competition and manipulation. A social planner has limited resources that she will allocate to multiple agents, on the basis of signals which the agents can manipulate through unproductive effort. We show that the welfare-maximizing mechanism takes the form of a contest, and we characterize the optimal contest. We apply our results to two settings: one in which there is only one item to be allocated, and one in which the number of items is proportional to the number of agents. In both settings we show that when there are sufficiently many agents, a winner-takes-all contest is never optimal. In particular, the planner always benefits from randomizing the allocation to some agents.

Keywords— contests, mechanism design without money, signaling, manipulation, competition

1 Introduction

Many real-world scenarios can be modeled as contests in which winners are selected based on a score which, in itself, is an unproductive signal, but which reflects the underlying (unobserved) attributes of interest. For

[*] 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, 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 for helpful suggestions, comments, and discussions. Yingkai Li acknowledges financial support from the Sloan Research Fellowship, grant no. FG-2019-12378.

[†] Cowles Foundation for Research in Economics and Department of Computer Science, Yale University. Email: yingkai.li@yale.edu

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

example, college admissions programs may aim to select talented students based on their test scores (e.g., SAT scores). Government subsidy programs may attempt to identify high-needs individuals based on their credit scores (e.g., FICO scores). Scientific funding agencies may rely on peer review scores to identify high-quality projects.

While these scores provide some information about agents' abilities, they are also open to manipulation via costly effort on the part of the agents. The obvious problem is that this leads agents to invest inefficiently in 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).[1] Companies may falsify their workforce sizes to meet criteria for legal and financial preferential treatment (Askenazy et al., 2022). Applicants for scientific funding may expend effort to overstate the merits of their projects, rather than to develop high-quality projects in the first place (Conix et al., 2021).

The goal of this paper is to characterize the mechanisms that maximize an arbitrary weighted average between *matching efficiency* and the *sum of the agents' utilities*. We take a reduced-form approach to modeling a screening problem with multiple agents where hidden effort is unproductive. In our setting, unlike in the signaling literature and the recent gaming literature (e.g., Spence, 1973; Frankel and Kartik, 2019; Ball, 2019), the mechanism designer has limited resources to allocate, which leads to a non-negligible competition effect among the agents.[2] Therefore, implementing full matching efficiency would encourage undesirable effort on the part of the agents, blurring the accuracy of their signals. The assumption that effort is unproductive and is purely a social waste constitutes the main departure of our setting from the existing literature on contests involving competition among agents. In the latter, effort is often seen as socially beneficial, so that contests are considered desirable insofar as they encourage effort; for example, Lazear and Rosen (1981) establish that contests are better than piece-rate systems at incentivizing effort. In our setting, however, where exerting effort is a rent-seeking act, it is unclear whether contests remain a desirable choice of mechanism.

Our first result provides a theoretical foundation for the pervasive use of the contest format in the real world: it establishes that in our setting, contests are optimal among all mechanisms. We start by considering general (direct) mechanisms, in which the agents report their types (possibly falsely) to the principal, and then the principal makes a signal recommendation to each agent (i.e., a recommendation of how much effort to exert). The signal recommendation for a given agent may depend on the reported types of the other agents; thus, in a general mechanism, the principal can theoretically coordinate the agents' effort choices by carefully choosing her recommendations. We then define a contest as a special case of a general mechanism, in which an agent's signal recommendation depends *only* on his own reported type; in other words, coordination across the agents is ruled out.[3] It turns out, however, that this is not detrimental:

[1] 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.

[2] Moreover, unlike in the classical signaling setting, where it is marginally less costly for high types to exert effort, in our setting there is no marginal cost difference across types. See Section 3 for more detailed discussion.

[3] The Cambridge Dictionary defines a contest as "a competition to do better than other people, especially to win a prize or achieve a position of leadership or power." However, there is no unified definition of a contest in the economics literature. Although the most commonly used structure is that of the winner-takes-all (WTA) contest, richer prize

the coordination available in the more complex general mechanisms is unnecessary, because contests are optimal. In particular, we show that the VCG-format mechanism (Definition 3) is strictly dominated by the winner-takes-all (WTA) contest.

The intuition for this result comes from the way in which these mechanisms (dis)incentivize effort. In a mechanism that is not a contest, each of the principal's signal recommendations takes into account the entire profile of reported types, and so the recommendation to a given agent may reveal ("leak") information about the other agents' types.[4] This makes feasible the following double-deviation strategy: an agent can misreport his type, receive his signal recommendation, and then either (1) follow the recommendation (if it is favorable to him) or (2) opt out (if it is not favorable). In non-contest mechanisms, such double deviation increases an agent's off-path utility. We show that when the effort cost is convex but not too convex, the effect of information leakage outweighs the effort savings made possible by coordination via signal recommendations.[5] A contest prevents this type of double deviation by shutting down the possibility of information leakage. Hence, for a fixed level of matching efficiency, a contest gives agents weakly lower utilities if they deviate by exerting large amounts of effort;[6] therefore, it results in weakly lower effort costs.

Having established that contests are optimal among all mechanisms, we then characterize the optimal contest. Viewed as a direct mechanism, the optimal contest may be described as follows. The type space of each agent is partitioned into three regions: (1) the *no-tension region*, where the optimal allocation rule coincides with the efficient allocation rule, and no type exerts effort; (2) the *no-effort region*, where the optimal allocation rule differs from the efficient allocation rule, and no type has a strict incentive to exert effort (here we select the principal-preferred equilibrium, in which types in the no-effort region do not exert effort);[7] and (3) the *efficient region*, where the optimal allocation rule coincides with the efficient allocation rule, and all types exert positive effort. Intuitively, the principal always wishes to allocate the items efficiently if this does not incentivize effort. However, if implementing the efficient allocation rule makes it too lucrative for an agent to misrepresent himself as a high type, it will induce effort. In such

structures are possible; for example, Moldovanu and Sela (2001) and Olszewski and Siegel (2020) consider contests with multiple and non-identical prizes. Moreover, contests may be stochastic; see e.g. Skaperdas (1996). In computer science, contests are modeled as all-pay auctions, which coincide with our definition (see Chawla et al. (2019) and references therein). In Section 5.1, we explain the connection between our definition and rank-order contests; in particular, we show that our contests can be viewed as *coarse ranking* contests.

[4] For comparison, Ben-Porath et al. (2023) consider a resource allocation problem where agents can exert costly effort to acquire evidence that credibly reveals their types. Therefore, in their setting, effort is desirable for the principal. The optimal mechanism is sequential and features coordination, because this incentivizes the most effort.

[5] The condition of not being too convex rules out, for example, cases where the cost function exhibits kinks. In the main body of the paper, we assume a linear cost function in order to obtain a closed-form characterization. In Appendix B, we show that the main results remain valid as long as the cost function is not too convex.

[6] In the formal proof, we show that any general mechanism that is not a contest is non-optimal. In other words, the optimal mechanism has to be a contest.

[7] These results are stated in terms of *interim*, rather than ex-post, allocation rules. Since we are considering a multi-agent setting, it is not immediately obvious what the ex-post implementation looks like when the no-effort region exists. To illustrate, suppose there is only one item, and under the optimal contest, the no-effort region consists of a single interval. In the ex-post implementation, if the highest two types fall into the no-effort region, then the item is randomly allocated between them. Otherwise, the item is allocated efficiently. See Figure 4 for an example with two agents, one item, and a convex efficient allocation rule.

scenarios, depending on the weight the principal places on the agents' utilities, either it remains optimal for the principal to implement efficient allocation while inducing effort on path (this occurs in the efficient region), or it is optimal for the principal to choose a portion of the type space on which to "flatten" the allocation rule, so as to eliminate agents' incentives to exert effort (this occurs in the no-effort region). We establish that in the optimal contest, these three regions account for all the possible outcomes. However, each region may consist of countably many intervals, and the order in which these intervals appear will depend on the shape of the type distribution and other primitives.

We apply our results to study contests with large numbers of agents. We consider two cases. In the first case, the principal has only one item to allocate. This assumption speaks to applications such as scholarship 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 grows, the format of the optimal contest converges to that of the WTA contest, but the principal's payoff does not converge to her WTA payoff. Our proof relies on the fact that under some mild technical assumptions, when the number of agents is sufficiently large, the interim efficient allocation rule is convex. By specifically characterizing the optimal contest under a convex interim efficient allocation rule, we are able to show that as the number of agents grows, the measure of the no-tension region (where the contest format is WTA) converges to that of the whole type space. However, for any sufficiently large but finite number of agents, the optimal contest also features a small but non-empty no-effort interval; that is, sufficiently high types always have incentives to exert effort, so that full matching efficiency can be achieved only at the price of high effort costs (low agent utilities). By randomizing the allocation on this interval, the principal can always obtain a higher payoff from the optimal contest than she would get from the WTA contest.

Since the principal's payoff in the optimal contest does not converge to that in the WTA contest, she will never use a WTA contest to approximate the optimal contest, no matter how large the number of agents is. Conceptually, this contrasts with the finding of Bulow and Klemperer (1996) that the revenue from implementing the efficient allocation rule with one additional buyer exceeds the revenue under the optimal mechanism.

In the second case we consider, the number of items grows proportionally with the number of agents. This model is better suited for applications such as college admissions and government benefit programs, where a non-negligible fraction of the agents receive an item. In this case, if the items were allocated efficiently, all types above a certain cutoff would get an item. We find that in the optimal contest, the principal randomizes the allocation for types around this cutoff ("middle" types) to eliminate their incentives to exert costly effort. This improves their expected utilities, though at the cost of slightly lowering matching efficiency. Our finding is reminiscent of Director's law, which says that public programs tend to be designed primarily to benefit the middle classes. It is also consistent with the empirical findings 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.

1.1 Related Work

Our paper is related to several strands of the existing literature. First, it complements the seminal work of Lazear and Rosen (1981) by providing a different rationale for the adoption of contests in practice. Lazear and Rosen show that contests perform better in encouraging effort than piece-rate pay, while our paper shows that contests perform at least as well in maximizing welfare as more general mechanisms. Moreover, Lazear and Rosen find that competition is always beneficial, whereas in our setting we show that it is ideal to have only a certain amount of competition, rather than the maximum level.

Second, there is a large literature on characterizing equilibria in contests with various allocation and prize structures (e.g., Barut and Kovenock, 1998; Baye et al., 1993; Che and Gale, 1998; Clark and Riis, 1998; Siegel, 2009, 2010). In these works, effort is assumed to be productive and hence desirable for the principal. The effort-maximizing contest has been considered in several specific settings; for example, the principal may be able to incentivize effort by assigning the agents to sub-contests Moldovanu and Sela (2006), or by increasing or decreasing inequality in the division of the prizes among winners (Fang et al., 2020). Zhang (2023) uses payoff equivalence to show that the effort-maximizing mechanism can be implemented by a contest. By contrast, our paper considers unproductive effort, and we do not impose any particular allocation structure. Our design problem has a completely different objective, namely, to allocate items efficiently while also minimizing wasteful effort costs. Moreover, we provide a tight characterization of the optimal contest when the number of agents is large, whereas in the previous literature on large contests (e.g., Olszewski and Siegel, 2016, 2020), only the limit of the optimal contests is characterized.

The signals in our model can be viewed as messages sent by the agents to the principal, and the corresponding effort costs can be viewed as lying costs. This ties our model to the literature on partial verification, where, however, lying costs are assumed to be either 0 or ∞, depending on the agent's true type and his message (e.g., Green and Laffont, 1986). In our paper we allow for a more general form of lying cost, where the cost of lying can increase with the magnitude of the lie. Furthermore, in contrast to the literature on evidence (e.g., Ben-Porath et al., 2014; Mylovanov and Zapechelnyuk, 2017), our model involves evidence that is not "hard," since an agent can (at a cost) fabricate a signal different from his true type.

As in the literature on money-burning (e.g., Hartline and Roughgarden, 2008; Chawla et al., 2019), the agents in our model can exert costly and socially wasteful effort, which is similar to burning utility. In the money-burning literature, however, the money burned is observable, and the money typically enters an agent's utility in a quasi-linear way that does not depend on the agent's type. In our model, by contrast, the effort expended is unobserved; the signal recommendation has to satisfy an extra incentive compatibility constraint owing to a moral hazard problem; and the signal enters each agent's utility in a type-dependent way.

Our paper is related to the literature on signaling games (e.g., Spence, 1973), but with several key differences. In classical signaling games, there is a competitive market that pays each agent a wage corresponding to his estimated type, whereas in our setting, there is a single designer who has a limited budget to be allocated across several agents. Furthermore, in the classical setting, the single-crossing property holds: effort is assumed to be productive, so the cost of exerting an extra unit of effort is lower for higher agent types.

By contrast, in our setting, we assume that effort is wasteful; therefore, the marginal cost of exerting effort is the same for all types, and the single-crossing property fails. This connects our paper to the recent gaming literature (e.g. Frankel and Kartik, 2019; Ball, 2019), which studies manipulative behaviors in signaling games. We depart from this literature by (1) considering a screening setting with multiple agents and limited resources, so that competition among agents is a non-negligible force, and (2) making different qualitative assumptions on the agents' utilities, which do not satisfy the single-crossing property. In addition, Frankel and Kartik (2019) assume the existence of an order-reversing action, so that there is no separating equilibrium (signal-jamming). In our setting, a separating equilibrium exists, but it is not optimal under the principal's objective.

Our paper is also related to manipulative behaviors in information design problems (e.g. Perez-Richet and Skreta, 2022), in mechanism design problems without monetary transfers (e.g. Perez-Richet and Skreta, 2023), and in classification problems in the machine learning setting (e.g. Hardt et al., 2016). Perez-Richet and Skreta (2022, 2023) focus on fraud-proof mechanisms, where agents have no incentive to exert effort on path; we impose no such restrictions, allowing agents to exert effort on path. Hardt et al. (2016) consider a single-agent classification problem, while we tackle a multi-agent resource allocation problem where agents can manipulate signals. The coordination of efforts among agents, one of the main challenges in our model, is absent in Hardt et al. (2016).

Finally, our paper is related to the literature on strategic communication with lying costs. A major departure from this literature is that we study an allocation problem.

2 An Example

Suppose there are two agents and one principal. Each agent i has a private type θ_i drawn from a uniform distribution $F(\theta) = \theta$ with support on $[0, 1]$. Notice that the private type here reflects the agent's hidden ability, instead of his private valuation of the item to be allocated.[8] Each agent i can privately choose a non-negative effort e_i to produce a public signal $s_i \in [0, \infty)$. The effort is $e_i = \max\{0, s_i - \theta_i\}$. As we will discuss in Section 3, this assumption captures realistic features of agents in our motivating examples. For simplicity, assume that effort cost equals effort level.

The principal has one item to allocate between the two agents based on their public signals; i.e., the principal chooses an allocation vector (x_1, x_2) with $0 \leq x_i \leq 1$, $i \in \{1, 2\}$, and $x_1 + x_2 \leq 1$. Each agent's valuation for the item is 1. His utility is the value of the item he receives minus his effort cost: $u_i = x_i - e_i$. The *matching efficiency* of an allocation (x_1, x_2) is measured by $\theta_1 x_1 + \theta_2 x_2$. Under perfect assortative matching, the agent with the higher type gets the item. The principal values both matching efficiency and minimizing the agents' effort, and she values them equally. That is, her objective is to maximize $\sum_{i=1}^{2}(\theta_i x_i - e_i)$. As we will discuss in Section 3, this objective function models the principal as a benevolent social planner who trades off between matching efficiency and the agents' utilities.

[8] One can think of the type as the cost type, as in the signaling literature, instead of the valuation type as in the mechanism design literature.

Suppose for the moment that the principal cares only about matching efficiency; that is, she wishes to implement the efficient allocation. In the classical auction environment, a natural candidate for implementing the efficient allocation is the second-price auction, in which the highest-type agent wins and pays the threshold price (the second-highest valuation). In our setting, where monetary payments are not involved, we can implement a similar mechanism using signals. In order for the winning agent to be indifferent between winning and losing at the threshold type (the second-highest type), the signal sent by the winning agent must equal not the second-highest type, but rather one (the value of the item) plus the second-highest type. We call this a *second-price-format* mechanism.

Example 1 (second-price-format 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*[9]*, and $s^* = 1 + \hat{\theta}_{-i^*}$ the signal that agent i^* has to produce to get the item.*[10]

- *The principal's recommendation is for agent i^* to generate signal s^*, 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^*. Otherwise the principal keeps the item.*

In this mechanism, by an argument similar to that used for the second-price auction, each agent has an incentive to truthfully report his type and produce the recommended signal. The item is then allocated to the agent with the higher type; that is, the principal achieves the best outcome in terms of matching efficiency. However, this mechanism fails to minimize expected effort, because for each agent to have an incentive to truthfully report his type, the higher-type agent has to "burn" some utility to prove that he has the higher type. More specifically, he has to exert strictly positive effort $(1+\mathbf{E}\left[\theta_{(2)}|\theta_{(2)} \leq \theta_{(1)}\right]-\theta_{(1)}) \in (0,1)$, where $\theta_{(2)}$ is the lower type and $\theta_{(1)}$ is the higher type.

Next we propose another mechanism, a winner-takes-all (WTA) contest, which allows the principal to increase her objective value by minimizing effort costs. A contest is a special mechanism in which the principal allocates the item based *solely* on the ranking of the agents' signals.[11][12]

Example 2 (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 $s_i(\theta_i) = \theta_i$.*

[9] To avoid heavy notations, we omit the dependence of i^* on $\hat{\boldsymbol{\theta}}$ whenever it does not cause confusion.

[10] The number 1 in the formula for s^* is each agent's valuation of the item. Under this recommendation, agent i's utility from reporting $\hat{\theta}_i$, given his own type θ_i and the other agent's reported type $\hat{\theta}_{-i}$, is $\theta_i - \hat{\theta}_{-i}$ if $\hat{\theta}_{-i} < \hat{\theta}_i$ and 0 otherwise. This gives type θ_i a utility of 0 when $\hat{\theta}_{-i} = \hat{\theta}_i$. If the recommended signal were lower, then type θ_i would have a positive utility when $\hat{\theta}_{-i} = \hat{\theta}_i$, which would create an incentive for him to misreport his type as higher. Similar reasoning applies when the recommended signal is higher. Hence, $s^* = 1 + \max_{i \neq i^*} \hat{\theta}_i$ is the only recommendation that works.

[11] Notice that the mechanism in Example 1 is not a contest, because the winning agent's signal depends on the other agent's type.

[12] In the next section, we will define a more general notion of contests.

Given that the other agent is using the strategy $s_{-i}(\theta_{-i}) = \theta_{-i}$, agent i's payoff from exerting effort e_i is $1 \cdot F(\theta_i + e_i) - e_i = \theta_i$. Thus, agent i has no strict incentive to exert effort, and $s_i(\theta_i) = \theta_i$ is a best response. This implies that it is an equilibrium for each agent to produce a signal equal to his own type. Notice that in this contest, as in the second-price-format mechanism of Example 1, the principal allocates the item to the agent with the highest type; furthermore, unlike in Example 1, each agent exerts zero effort in equilibrium. Therefore, the contest maximizes the principal's objective function; in other words, it is optimal.

Moreover, in both the second-price-format mechanism and the contest, the efficient allocation rule is implemented, and the lowest type gets utility zero, since he receives the item with probability zero by exerting effort zero. However, in the contest, agent i's expected utility is $F(\theta_i) = \theta_i$, while in the second-price-format mechanism it is $(\theta_i - \mathbf{E}[\theta_{-i}|\theta_{-i} \leq \theta_i])F(\theta_i)$, which is strictly less than $F(\theta_i)$ for any $\theta_i > 0$. These observations immediately imply that payoff equivalence fails to hold in our setting.

The optimality of contests relative to more general mechanisms can be explained as follows. The role of the agents' effort in our setting is analogous to that of monetary transfers in mechanism design settings. However, because effort is not observed, it is not enforceable; the only observable (and hence contractible) quantities are the agents' signals. The fact that neither type nor effort is observable introduces the possibility of double deviation: for a given recommended signal, an agent could first lie about his type, then exert effort to produce a commensurate signal.

In the second-price-format mechanism, the recommended signal for each agent depends on the other agents' realized types. Thus, from an agent's point of view, the recommended signal is stochastic. This makes the double-deviation strategy attractive: an agent could misreport as a higher type and then, depending on the realized signal recommendation, either produce a signal to match his reported type (if the effort cost needed is not too high) or opt out (if it is). To put it differently, the signal recommendation reveals some information about the other agents' types (a phenomenon we refer to as "information leakage"), which sheds light on whether it will be profitable to pretend to be a high type. The principal must recommend a signal that is high enough to deter such double deviation. This makes it more expensive (in terms of effort costs) to implement a given efficient allocation rule under the second-price-format mechanism than to do so in a scenario where the signal recommendation does not leak information about the other agents' types – i.e., in a contest. Contests rule out information leakage because they require the principal to commit to a contest rule, which specifies a Bayesian game for the agents. Hence each agent chooses a strategy that does not depend on the other agents' types.

It turns out that these observations are not confined to the current simple example. In the following sections, we will show that contests remain optimal in more general settings.[13] In particular, we will show that the second-price-format mechanism is strictly dominated by the WTA contest. In the formal analysis, we will reverse the order of exposition used above. Here is a sketch of our approach using the simple example of this section. To more closely compare the mechanisms described, notice first that the contest

[13] For a general distribution function, it is not necessarily true that contests induce zero effort on path. However, even in cases where contests induce effort, they achieve higher welfare.

can be implemented by a direct mechanism in which each agent i reports his own type as $\hat{\theta}_i$. Given agent i's reported type $\hat{\theta}_i$, the principal recommends that he generate the signal $s_i(\hat{\theta}_i) = \hat{\theta}_i$. The item is allocated to the agent who generates the highest signal. The major difference between the direct mechanisms of Examples 1 and 2 is that in the contest, an agent's signal recommendation is based solely on his own type. It turns out that the converse is also true: for each member of the smaller class of mechanisms under which an agent's signal recommendation depends only on his type (and no one else's), there is indeed an indirect implementation that is consistent with what we usually call a contest, i.e., a game that allocates items to agents based on the ranking of their signals.

The rest of the paper is organized as follows. In Section 3, we formally define our model. In Section 4, we define first the class of general mechanisms we consider, then a smaller class in which each agent's signal recommendation depends only on his own type. We say mechanisms in this smaller class are *implementable by contests*, or we simply call them *contests*, for the time being without justification. We then prove that the optimal mechanism (out of all general mechanisms) falls within the class of contests.

In Section 5 we precisely characterize the optimal contest, and in Section 5.1 we verify that the mechanisms we call contests are consistent with the class of games usually called contests in the literature. In Section 6 we apply our main results to large (but finite) contests. In Section 7 we briefly discuss several generalizations of our results. Section 8 provides concluding remarks. The appendix contains all proofs omitted from the main body of the paper.

3 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 $\sum_{i=1}^n x_i \leq k$. Let $X \subseteq [0,1]^n$ be the space of all such (randomized) allocations. Each agent i has a private type θ_i drawn independently from a publicly known distribution F_i supported on $\Theta_i = [\underline{\theta}_i, \bar{\theta}_i] \in \mathbb{R}_+$. Denote the distribution over type profile $\boldsymbol{\theta}$ by $\boldsymbol{F}$. For simplicity, we assume that the density function f_i exists for all i, and $f_i(\theta_i) > 0$ for any $\theta_i \in [\underline{\theta}_i, \bar{\theta}_i]$.

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 public signal $s_i \in S_i = \mathbb{R}_+$ by expending effort $e(s_i, \theta_i) = (s_i - \theta_i)^+ := \max\{0, s_i - \theta_i\}$. In other words, each agent's signal is the sum of his type and the effort he chooses to expend, assuming the desired signal is higher than his type. This specification captures the feature that an agent can generate any signal lower than his own type for free; for example, a talented student can easily pretend to be less talented by purposely giving wrong answers on an exam. The assumption also captures the phenomenon of head starts in contests (see Siegel, 2014); for instance, more talented students can achieve high scores with less effort than less talented students. Unlike in the classical signaling setting, in our model, having a higher type does not make effort more productive in terms of generating signal; that is, we do not model the signal as the product of effort and type. We argue that our formulation better matches applications where gaming effort and innate ability are orthogonal. For instance, in certain school subjects, high test scores can be achieved either through logical reasoning ability or through rote memorization of the material, and regardless of which skill the test is intended to measure, improving one does not affect the other. Applicants for research funding may expend effort to produce longer or better-written grant applications, which does not increase the quality or originality of the underlying project.[14]

Assume that agent i's effort cost is $\eta \cdot e(s_i, \theta_i)$, where $\eta > 0$, representing each agent's marginal cost of effort (i.e., his ability to manipulate his signal), is publicly known.[15] The linearity of the cost function is not necessary for our main results, but it simplifies the analysis and exposition. In Appendix B, we extend our results to a wider class of convex cost structures.

Payoffs. Each agent has unit demand for the items. If agent i gets a single item with probability $x_i \in [0, 1]$ when generating signal s_i, then his utility is [16]

$$u_i = x_i - \eta \cdot e(s_i, \theta_i).$$

The *matching efficiency* of an allocation $\boldsymbol{x}$ is measured by $\sum_i \theta_i \cdot x_i$. The matching efficiency is highest under perfect assortative matching, in which the highest k types each get one item. The principal's payoff from choosing an allocation $\boldsymbol{x}$ and utility profile $\boldsymbol{u} = (u_i)_{i=1}^n$ is

$$\sum_{i=1}^{n} \alpha \cdot \theta_i \cdot x_i + (1 - \alpha) \cdot u_i, \tag{1}$$

for some $\alpha \in [0, 1]$. To understand this objective function, one should view the principal as a benevolent social planner who trades off between matching efficiency and the agents' utilities. Matching efficiency provides a reduced-form approach to capturing the long-term benefit of allocating the best resources to the individuals who will make the best use of them. The idea is that, for example, a government subsidy program with a limited budget should give the most money to the firms or individuals with the most financial need; a college admissions program should match the brightest students to the best educational resources. However, when agents can expend wasteful effort to influence the principal's decisions, it is in society's interest to minimize such effort.

General mechanism design. We first define the broadest class of mechanisms we will consider for our setting. We restrict our attention to direct mechanisms, in which the principal can communicate with all the agents, and can make signal recommendations based on the communications, before the agents choose their signals. By the revelation principle, it is without loss to focus on direct mechanisms where first the agents

[14] We thank Carl-Christian Groh for suggesting these examples.

[15] It is without loss to assume that η is publicly known. If η were private, the principal could easily discover it by asking each agent to report η in addition to his type, and never allocating items to agents who reported differently from others. Truthfully reporting η would then constitute an equilibrium.

[16] Notice that the agent's utility does not satisfy single crossing in our setting. By contrast, the classical mechanism design and signaling game settings usually assume strict single crossing. Single crossing does not hold in the gaming literature (e.g., Frankel and Kartik, 2019), but these papers make different assumptions on the agent's utility.

report their private types to the principal, and then the principal recommends signals to the agents based on the aggregated reports. Formally, the timeline for a general mechanism is as follows:

(1) The principal commits to a signal recommendation policy $\tilde{s} : \prod_{i=1}^n \Theta_i \to \Delta(S)$ and allocation rule $p : \prod_{i=1}^n \Theta_i \times S \to X$.

(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 $s'_i \in \{\tilde{s}_i(\hat{\boldsymbol{\theta}}), 0\}$.[17]

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

For each agent i with private type θ_i, we define the *interim allocation* as

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

and the *interim utility* as

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

The class of general mechanisms defined here is slightly larger than the class of mechanisms considered in the classical setting (e.g. Myerson, 1981), where the allocation is determined by the reported type profile. In our setting, monetary transfers are not available, so the signal (or effort) recommendations are the principal's only means of managing the agents' incentives. In any mechanism in which the allocation is based on the reported type profile and not the agents' actual signals, the reports are purely cheap talk, and agents have strong incentives to misreport: since effort is unobservable, each agent can simply report the highest type possible, then save on effort costs by producing a signal no higher than his true type. Hence it is more natural for the allocation rule to depend on both the reported type profile and the actual signals.

We now formally define what it means for an interim allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ to be implementable by a general mechanism.

Definition 1. *An interim allocation–utility pair* $(\boldsymbol{Q}, \boldsymbol{U})$ *is* implementable by a general mechanism *if the following hold:*

(1) The pair $(\boldsymbol{Q}, \boldsymbol{U})$ *is induced by some signal recommendation policy* $\tilde{\mathbf{s}}$ *and some allocation rule* $\mathbf{p}$*:*

$$\begin{aligned} Q_i(\theta_i) &= \mathbf{E}_{\boldsymbol{\theta}_{-i}}\left[\mathbf{E}_{s_i \sim \tilde{s}_i(\theta_i, \boldsymbol{\theta}_{-i})}[p_i(\theta_i, \boldsymbol{\theta}_{-i}, s_i, \boldsymbol{s}_{-i}(\boldsymbol{\theta}_{-i}))]\right] , && \text{(consistency)} \\ U_i(\theta_i) &= Q_i(\theta_i) - \eta \cdot \mathbf{E}_{\boldsymbol{\theta}_{-i}}\left[\mathbf{E}_{s_i \sim \tilde{s}_i(\theta_i, \boldsymbol{\theta}_{-i})}[e(s_i, \theta_i)]\right] , \quad \forall i, \theta_i. \end{aligned}$$

[17] This is without loss of generality. The principal can partially enforce the recommendation by allocating no items to any agent who chooses a signal different from the recommended one. Hence each agent essentially has two choices: following the recommendation ($s'_i = \tilde{s}_i(\hat{\boldsymbol{\theta}})$) or opting out ($s'_i = 0$).

(2) Each agent has a weak incentive to truthfully report his own type and follow the signal recommendation:

$$U_i(\theta_i) \geq \mathbf{E}_{\boldsymbol{\theta}_{-i}}\Big[\mathbf{E}_{s_i \sim \tilde{\boldsymbol{s}}(\theta_i', \boldsymbol{\theta}_{-i})}\big[\max\{0, p_i(\theta_i', \boldsymbol{\theta}_{-i}, s_i, \tilde{s}_{-i}(\theta_i', \boldsymbol{\theta}_{-i})) - \eta \cdot e(s_i, \theta_i)\}\big]\Big],$$
$$\forall i, \theta_i, \theta_i'. \qquad \text{(incentives)}$$

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

In general mechanisms (as opposed to contests, which are defined below), the principal can communicate with all the agents, and the signal recommendation for one agent may depend on the reported types of the other agents.

Contest design. We now introduce a smaller class of mechanisms, which we refer to as contests. Here, the principal commits to a contest rule $\boldsymbol{x} : S \to X$, which maps each realized signal profile to a (randomized) allocation. After the contest rule is announced, each agent chooses his signal strategy. There is no communication between the principal and the agents before the agents make their choices. 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.

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 contest rule $\boldsymbol{x}(\boldsymbol{s}(\boldsymbol{\theta}))$. Formally, the timeline is as follows:

(1) The principal commits to a contest rule $\boldsymbol{x} : S \to X$.

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

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

A contest rule $\boldsymbol{x}$ defines a game in which the agents' equilibrium strategy profile is $\boldsymbol{s}^*$. Given an arbitrary strategy profile $\boldsymbol{s}'$, agent i's payoff is $v_i(\boldsymbol{s}', \theta_i) = x_i(\boldsymbol{s}') - \eta \cdot e_i(s_i', \theta_i)$. In equilibrium, agent i's expected utility is $u_i(s_i^*; \theta_i) = \mathbf{E}_{\boldsymbol{\theta}_{-i}}\big[v_i(s_i^*, \boldsymbol{s}_{-i}^*(\boldsymbol{\theta}_{-i}), \theta_i)\big]$.

Equilibrium analysis for contests is notoriously intractable (see for instance Olszewski and Siegel, 2016). Therefore, we instead analyze contests via mechanism design techniques. This is possible thanks to the useful observation that a contest can be viewed as a direct mechanism in which each agent reports his type to the principal, then receives a signal recommendation based *solely* on his own reported type. In the direct mechanism that implements the equilibrium $\boldsymbol{s}^*$ induced by the contest x, we can define the ex-post allocation for agent i as $q_i(\theta_i, \boldsymbol{\theta}_{-i}) = x_i(s_i^*(\theta_i), \boldsymbol{s}_{-i}^*(\boldsymbol{\theta}_{-i}))$, the interim allocation as $Q_i(\theta_i) = \mathbf{E}_{\theta_{-i}}[q_i(\theta_i, \theta_{-i})]$, and the interim utility as $U_i(\theta_i) = u_i(s_i^*(\theta_i); \theta_i)$. Let $\boldsymbol{Q} = (Q_i)_{i=1}^n$ denote the interim allocation rule and $\boldsymbol{U} = (U_i)_{i=1}^n$ the interim utility profile.

Definition 2. *An interim allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ is* implementable by a contest *if there exist a contest rule $\boldsymbol{x}$ and a signal strategy $\hat{\boldsymbol{s}}$ such that the following hold:*

(1) The pair $(\boldsymbol{Q}, \boldsymbol{U})$ is induced by a contest rule $\boldsymbol{x}$ and a collection of signal strategies $\hat{\boldsymbol{s}}$ given by $\hat{s}_i : \Theta_i \to S$ for each i:

$$Q_i(\theta_i) = \mathbf{E}_{\theta_{-i}}[x_i(\hat{s}_i(\theta_i), \hat{s}_{-i}(\theta_{-i}))], \qquad \text{(consistency)}$$
$$U_i(\theta_i) = Q_i(\theta_i) - \eta \cdot e(\hat{s}_i(\theta_i), \theta_i), \quad \forall i, \theta_i.$$

(2) The signal strategies $\hat{\boldsymbol{s}}$ form an equilibrium:

$$U_i(\theta_i) \geq \mathbf{E}_{\theta_{-i}}\big[x_i(s_i', \hat{s}_{-i}(\theta_{-i}))\big] - \eta \cdot e(s_i', \theta_i), \quad \forall i, s_i', \theta_i. \qquad \text{(incentives)}$$

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

In our definition, we do not require the contest rule $\boldsymbol{x}$ to be a mapping from signal rankings to allocations. In Section 5.1 we will show, by generalizing the commonly used notion of *strict ranking* to our notion of *coarse ranking*, that such a restriction would be redundant.

As stated earlier, the main difference between our definitions of implementability by a contest and implementability by a general mechanism is that in a contest, the recommended signal for agent i is a function only of agent i's *own* reported type, while in a general mechanism, it may be a function of the entire reported type *profile*. This difference in itself does not explain why we use the term “contest” for the more restricted class of mechanisms. We will explain this in Section 5.1, where we show that the requirement that an agent's recommended signal depend only on his own type allows us to construct the contest rule, a mapping from signal rankings to allocations, that implements the mechanism.

Interim approach. In the rest of the paper, instead of analyzing the set of ex-post allocation rules, i.e., mappings from the space of type profiles to the space of allocations, we analyze the set of interim allocation rules, i.e., mappings from the Cartesian product of each agent' individual type space to the space of allocations. Thus, instead of analyzing the interim allocation rule and signal recommendations, we analyze the interim allocation rule and interim utilities.

However, not all interim allocation rules are *feasible*. Specifically, let an interim allocation rule $\boldsymbol{Q} = (Q_i)_{i=1}^n$ be a profile of mappings $Q_i : \Theta_i \to [0,1]$. Let an ex-post allocation rule $\boldsymbol{q} = (q_i)_{i=1}^n$ be a profile of mappings $q_i : \prod_{i=1}^n \Theta_i \to X$. We say the interim allocation rule $\boldsymbol{Q}$ is *interim 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.

Given these technical considerations, our design problem is equivalent to a problem where the principal chooses a feasible interim allocation rule $\boldsymbol{Q}$ and interim utility profile $\boldsymbol{U}$ to maximize

$$\text{Obj}_\alpha(\boldsymbol{Q}, \boldsymbol{U}) = \mathbf{E}_{\boldsymbol{\theta}}\left[\alpha \cdot \sum_i \theta_i \cdot Q_i(\theta_i) + (1-\alpha) \cdot \sum_i U_i(\theta_i)\right]. \qquad (2)$$

4 Optimality of Contests

The main goal of this section is to show that contests are optimal among all mechanisms. In order to state our result, we need to establish several lemmas characterizing incentive compatibility constraints in both contests and general mechanisms.

Incentive compatibility. First we characterize the incentive compatibility conditions in any direct mechanism that implements a monotone allocation rule. In general, there exist non-monotone interim allocation rules that are implementable by contests and by general mechanisms (see Section 7.1 for an example). However, as we will show in Theorem 1, it is without loss of optimality to focus on monotone allocation rules.

Lemma 1. *An interim allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ with monotone $\boldsymbol{Q}$ is implementable by a contest if and only if $\boldsymbol{Q}$ is interim feasible, and for any agent i with type θ_i,*[18]

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

The idea behind Lemma 1 is as follows. For any interim allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ that is implementable by a contest, 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.

Similarly, we can derive a necessary condition for incentive compatibility in a general mechanism.

Lemma 2. *An interim allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ is implementable by a general mechanism only if $\boldsymbol{Q}$ is interim feasible, and for any agent i with type θ_i,*

(1) $U_i'(\theta_i) \in [0, \eta]$*;*

(2) $U_i(\theta_i) \leq Q_i(\theta_i)$*.*

The proofs of Lemmas 1 and 2 are provided in Appendix A.1. The characterization of incentive compatibility is more challenging for general mechanisms than for contests, because in a general mechanism, each agent's signal recommendation is essentially stochastic (since it may depend on the other agents' reported types). No existing technique is available to deal with stochastic mechanisms, since in the classical setting it is typically without loss to assume deterministic mechanisms.

[18] 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.

Payoff equivalence. From Lemma 1 we can establish Proposition 1, which says that for any implementable allocation–utility pair, 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.

Proposition 1. *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 contest. Moreover, for any interim allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ that is implementable by a contest, 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 1. 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 1, 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. In Appendix A.1, we provide a formal proof to circumvent this technicality.

It is immediate from Proposition 1 that the lowest type exerts zero effort in the optimal contest, 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.

Optimality of contests. We now state a lemma that will be useful in the proof of our main theorem.

Lemma 3. *Fix any monotone and interim feasible allocation rule $\boldsymbol{Q}$. There exists a unique interim utility profile $\hat{\boldsymbol{U}}$ with $\hat{U}_i(\underline{\theta}_i) = Q_i(\underline{\theta}_i)$ for all i such that $(\boldsymbol{Q}, \hat{\boldsymbol{U}})$ is implementable by a contest. Moreover, $(\boldsymbol{Q}, \hat{\boldsymbol{U}})$ achieves the highest utility for each agent among all general mechanisms that implement $\boldsymbol{Q}$.*

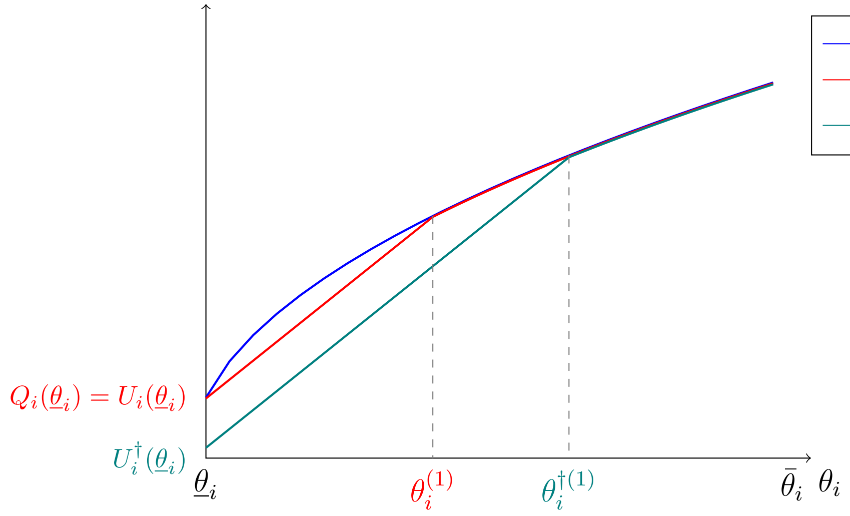
Proof. The first part is directly implied by Proposition 1.

From Lemma 1, we know that for any i, θ_i, if $\hat{U}_i(\theta_i) < Q_i(\theta_i)$, then $\hat{U}_i'(\theta_i) = \eta$. Therefore, if $\hat{U}_i'(\theta_i) < \eta$, then $\hat{U}_i(\theta_i) = Q_i(\theta_i)$. This implies that $\hat{U}_i(\theta_i)$ can be viewed as a function that pointwise maximizes the set of functions that has a derivative between 0 and η and bounded between 0 and $Q_i(\theta_i)$ for each θ_i; that is,

$$\hat{U}_i(\theta_i) = \max\{U_i(\theta_i) : U_i'(\theta_i) \in [0, \eta] \text{ and } 0 \leq U_i(\theta_i) \leq Q_i(\theta_i)\} \tag{3}$$

Combining this with Lemma 2, we know that for any $(\boldsymbol{Q}, \boldsymbol{U})$ that is implementable by a general mechanism, we must have $U_i'(\theta_i) \in [0, \eta]$ and $0 \leq U_i(\theta_i) \leq Q_i(\theta_i)$. Hence, $\hat{U}_i(\theta_i) \geq U_i(\theta_i)$ for any i and θ_i. □

Figure 1: Illustration of Proposition 1



†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 1, U_i is uniquely identified by the allocation rule and $U_i(\underline{\theta}_i)$.

Now we are ready to state our first main result.

Theorem 1 (optimality of contests)**.** *For any interim allocation–utility pair* $(\boldsymbol{Q}, \boldsymbol{U})$ *that is implementable by a general mechanism, there exists another interim allocation–utility pair* $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ *with monotone allocation* $\boldsymbol{Q}^\dagger$ *that is implementable by a contest and yields a weakly higher objective value for all* $\alpha \in [0,1]$, *where* α *is the welfare weight on* matching efficiency *defined in Eq.* (1).

Proof. **Step 1: Showing that monotone allocation rules increase matching efficiency.** Lemma 2 implies that $\boldsymbol{Q}$ is interim feasible and $U_i(\theta_i) \leq Q_i(\theta_i)$ for any agent i with type θ_i. If $\boldsymbol{Q}$ is monotone, set $\boldsymbol{Q}^\dagger = \boldsymbol{Q}$; then $\boldsymbol{Q}^\dagger$ is interim feasible. Otherwise, let $G_i(z) = \int_{\underline{\theta}_i}^{\bar{\theta}_i} \mathbf{1}\left[Q(\theta_i) \leq z\right] \mathrm{d}F_i(\theta_i)$ and let $Q_i^\dagger(\theta_i) = G_i^{-1}(F_i(\theta_i))$.[19] Essentially, $Q_i^\dagger$ is a rearrangement of the allocation Q_i such that $Q_i^\dagger$ is monotone, and the measure of the set of types with allocation at most z is the same for all z. Using results of Border (1991) and Che et al. (2013) (the latter restated below), it is easy to verify that $Q^\dagger$ is interim feasible.

Lemma 4 (Che et al., 2013)**.** *Given a set* $\boldsymbol{A} = \prod_{i=1}^n A_i \subset \prod_{i=1}^n \Theta_i$, *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.[20] *The interim allocation rule* $\boldsymbol{Q}$ *is interim feasible if and only if*

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

[19] Here, $\mathbf{1}[\cdot]$ denotes the indicator function.

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

Moreover, for monotone allocations in symmetric environments, (IF) *is equivalent to the following:*[21]

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

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.*

Moreover, as $Q_i^\dagger$ only shifts the allocation probability from low types to high types, for any agent i,

$$\mathbf{E}_{\theta_i}\left[\theta_i \cdot Q_i^\dagger(\theta_i)\right] \ge \mathbf{E}_{\theta_i}[\theta_i \cdot Q_i(\theta_i)]\,.$$

Step 2: Showing that $U_i(\theta_i) \le Q_i^\dagger(\theta_i)$ for all types θ_i. Suppose otherwise. Then there exists a type θ_i such that $U_i(\theta_i) > Q_i^\dagger(\theta_i)$. As $Q_i^\dagger$ is simply a rearrangement of the allocation Q_i, there exists a type $\theta_i' \ge \theta_i$ such that $Q_i(\theta_i') \le Q_i^\dagger(\theta_i)$. By incentive compatibility, $U_i(\theta_i) \le U_i(\theta_i')$. These three inequalities imply that

$$Q_i(\theta_i') \le Q_i^\dagger(\theta_i) < U_i(\theta_i) \le U_i(\theta_i').$$

This violates the requirement for $(\boldsymbol{Q}, \boldsymbol{U})$ to be implementable by a general mechanism, which is a contradiction.

Step 3: Optimality of contests. Given any interim allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U})$, Step 1 implies that there exists an interim feasible and monotone allocation $\boldsymbol{Q}^\dagger$ with weakly higher matching efficiency. Let $U^\dagger$ be the corresponding utility function given in Lemma 3. Then $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ is implementable by a contest. Moreover, combining Step 2 and the fact that $U_i'(\theta_i) \in [0, \eta]$ for all i and θ_i (from Lemma 2), we have $U_i^\dagger(\theta_i) \ge U_i(\theta_i)$ for all i and θ_i.

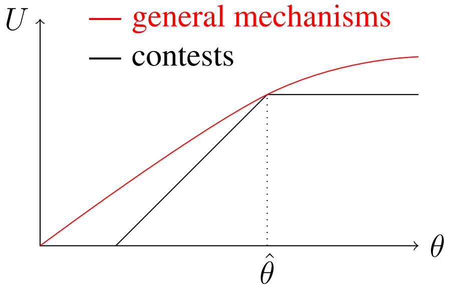
Since both the expected matching efficiency and the agents' expected utilities are weakly greater, the objective value for $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ is weakly greater than that of $(\boldsymbol{Q}, \boldsymbol{U})$. □

This result has two implications. First, for any non-monotone allocation rule that is implementable by a general mechanism, there exists a monotone modification that is both feasible and implementable. Second, any monotone allocation rule that is implementable by a general mechanism is also implementable by a contest that yields a weakly higher objective value for the principal. Together, these observations imply that it is without loss of optimality to restrict our attention to contests with monotone allocations.

The contest structure can be viewed as an all-pay mechanism, where each agent chooses some amount of effort to expend to compete for the items, independently of the other agents' effort choices. Our result indicates that this all-pay format is optimal for any choice of α in the objective function. At first glance this may seem surprising, as the goal is to minimize costly effort, while contests require every agent to pay the

[21] 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.

Figure 2: Intuition for Theorem 1



[†] The black (red) curve is the utility of type θ when he misreports his type as $\hat{\theta}$ in a contest (general mechanism). In a general mechanism, agents of every type have higher deviation utility, because each agent's signal recommendation reveals information about the other agents' types, based on which he may decide not to follow the recommendation. A general mechanism needs to deter such double deviation and hence is more "expensive" to implement.

cost of the effort he expends. The intuition is better seen in the discrete-type setting, as illustrated in Figure 2. The all-pay format minimizes the incentives for agents to deviate from the principal's recommendations. This alleviates the moral hazard issue, which reduces the expected amount of effort required from each agent to prove that he has the claimed type.

The linearity assumption is not crucial for Theorem 1 to hold. However, it ensures that local incentive compatibility is enough to guarantee global incentive compatibility, enabling us to characterize incentive compatibility in closed form, which greatly simplifies the analysis. In Appendix B, we show that contests remain optimal for a broad family of convex cost functions.

Greenwald et al. (2018) and Zhang (2023) show that the optimal mechanism can be implemented by a contest in the setting where effort is productive and the principal wants to maximize effort. Our results differ from theirs in at least two aspects. First, in our setting, effort is unproductive, and the principal aims to maximize not the agents' effort, but rather a weighted average of the matching efficiency and the agents' utilities. Second, the optimality of contests in Greenwald et al. (2018) and Zhang (2023) relies on payoff equivalence to hold for the class of general mechanisms, while in our model, payoff equivalence fails for all general mechanisms. For instance, the mechanism in Example 1 (Section 2) cannot be implemented by a contest.[22]

Strict suboptimality of VCG-format mechanism. In this section, we focus on symmetric environments, where $F_i = F_j$ for all $i, j \in [n]$. We show that even in symmetric environments, the VCG-format mechanism (defined below) is strictly suboptimal.[23] The proof of Proposition 2 is provided in Section 4.

[22] A related observation is made in Perez-Richet and Skreta (2023), Appendix C. There, however, the authors restrict their attention to score-based allocation rules that depend only on each agent's score. They argue that this is without loss of generality, because (1) their setting can be viewed as a single-agent problem (because it involves a continuum of agents), and (2) actions are *contractible* in their setting. In contrast, our model is a multi-agent problem, in which the principal may be able to benefit from communicating with the agents and thereby coordinating their actions. Therefore, our problem is more involved and is not implied by the result of Perez-Richet and Skreta (2023).

[23] When there is one item, the VCG-format mechanism coincides with the second-price-format mechanism described in Section 2.

Definition 3 (VCG-format mechanism for agents with unit demand)**.** *A mechanism is of* Vickrey–Clarke–Groves (VCG) format *if it admits the following form:*

- *Each agent reports his own type $\hat{\theta}_i$.*
- *For each reported type profile $\hat{\boldsymbol{\theta}}$, let $I^* = \{i : \hat{\theta}_i$ is among the k highest types$\}$ be the set of agents with the top k reported types. Let $s^*_{i^*} = \frac{1}{\eta} + \max_{i \notin I^*} \hat{\theta}_i$ be the signal that agent $i^* \in I^*$ has to produce in order to get one item.*
- *The principal recommends that agent i^* generate signal $s^*_{i^*}$ if $i^* \in I^*$, and signal 0 otherwise. Agent i^* receives an item if and only if the observed signal for i^* is at least $s^*_{i^*}$ and $i^* \in I^*$. Otherwise the principal keeps the item.*

Definition 4 (WTA contest for agents with unit demand)**.** *A mechanism is a* WTA contest *if it has the following form: each of the k agents who produce the highest k signals receives one item. If there are more than k such agents, the items are allocated randomly among them.*

Proposition 2. *For any $\alpha < 1$, any $n \geq 2$, any $k < n$, and any symmetric distribution with positive densities everywhere in the support, the VCG-format mechanism (for agents with unit demand) is strictly dominated by the WTA contest (for agents with unit demand).*

5 Optimal Contests

By Theorem 1, the principal's design problem can be simplified as that of choosing a monotone and feasible interim allocation rule and an interim utility profile such that the interim allocation–utility pair is implementable by a contest, which is equivalent to imposing the set of constraints in (IC). That is, the principal solves the following problem:

$$\begin{aligned} V_\alpha = \sup_{\boldsymbol{Q},\boldsymbol{U}} \quad & \mathrm{Obj}_\alpha(\boldsymbol{Q},\boldsymbol{U}) \\ \text{s.t.} \quad & \boldsymbol{Q} \text{ is monotone and interim feasible, and } (\boldsymbol{Q},\boldsymbol{U}) \text{ satisfies (IC).} \end{aligned} \tag{$\mathcal{P}_\alpha$}$$

From now on, when there is no risk of confusion, we use the terms "contest" and "direct mechanism that is implementable by a contest" interchangeably. Hence the task of finding the optimal mechanism is essentially the task of finding the optimal contest. Also, to simplify the analysis, in the rest of the paper we assume that the agents are ex-ante homogeneous.[24]

Assumption 1 (symmetric environment)**.** *The agents are ex-ante homogeneous: $\Theta_i = \Theta = [\underline{\theta}, \bar{\theta}]$, $F_i = F$, and $f_i = f$ for all i.*

[24] See Section 7.4 for a discussion of asymmetric environments.

Note that for Problem ($\mathcal{P}_\alpha$), although the objective function is linear, the (IC) constraints are not convex. That is, a convex combination of two allocation–utility pairs may violate the (IC) constraints. Nonetheless, as we show in the following lemma, the optimal contest for this non-convex optimization problem is always symmetric in symmetric environments.

Lemma 5. *Under Assumption 1, the optimal contest is symmetric for any* $\alpha \in [0, 1]$.

If we restrict our attention to symmetric contests, the problem of designing the optimal contest 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:

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

Che et al. (2013) provide a simplification of interim feasibility for symmetric environments.

Let (Q_α, U_α) be the optimal solution for Problem ($\mathcal{P}_\alpha$).[25] The following theorem implies that the optimal allocation partitions the type space into three types of intervals.

Theorem 2. *Under Assumption 1, for any* $\alpha \in (0, 1)$*, the optimal contest* (Q_α, U_α) *defines an interval partition* $\{(\underline{\theta}^{(j)}, \bar{\theta}^{(j)})\}_{j=1}^{\infty}$ *of the type space.*[26] *For any* $j \ge 1$*, the interval* $(\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$ *belongs to exactly one of the following three regions:*[27]

(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* $\theta \in (\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$.

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

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

[27] 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 contest

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

Each of these three regions is a union of (potentially countably many) intervals. If some of the highest k types among the n agents, say $0 < \ell \leq k$ of them, fall into the no-tension region, then these ℓ agents receive one item each and exert no effort. Similarly, if $\ell > 0$ of the highest k types fall into the efficient region, then each of them receives one item and exerts positive effort. However, if some of the highest $\ell > k$ types among the n agents fall into the same interval within the no-effort region, then none of these agents exerts any effort, and the items are allocated randomly among them, meaning the highest k types receive one item each with probability 1, but the higher type has higher probability of receiving one item. In Section 6.1, we provide an example where we characterize the optimal interim allocation rule (see Fig. 3) and illustrate the allocation of the item (see Fig. 4).

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.

In general, both the number of intervals in each region and the order of the intervals depend on the shape of the efficient allocation rule and the coefficient α. In the following sections, we derive a sharper characterization of the number and order of the intervals when the number of agents is sufficiently large.

5.1 Indirect Implementation: Coarse Ranking Contests

In the previous sections, we defined a contest rule as a mapping from signal profiles to allocation profiles. However, in some papers in the literature on contest design (e.g., Moldovanu and Sela, 2001; Lazear and Rosen, 1981; Skaperdas, 1996), this mapping must satisfy a subtler requirement: a contest is defined as allocating items based *solely* on the agents' rankings, instead of on the cardinal values of their signals.[28] In

[28] As defined by Skaperdas (1996), a Tullock contest allocates the items stochastically based on the contest success function, which depends on the cardinal values of the whole signal (performance) profile. However, Konrad et al. (2009) point out that the Tullock contest can alternatively be viewed as allocating the items based on the ranking of a noisy signal s, defined as $\log s = \log g(e) + \epsilon$, for some deterministic and strictly increasing function g and stochastic

this section, we propose the concept of *coarse ranking*, in which segments of signals are pooled and assigned the same rank. Correspondingly, we extend the usual notion of a contest (or contest rule) to the so-called *coarse ranking contest* (or *coarse ranking contest rule*), where the items are allocated to the k agents with the highest coarse rankings, with ties broken uniformly at random. This generalizes the idea of a contest as commonly used in the literature.

We first provide a formal definition of coarse ranking.

Definition 5 (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*

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

and the number of ties for agent i *is*

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

Here, for any signal s_i*,* j_{s_i} *is the index* j *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* $\bar{s}^{(j_{s_i})} = \underline{s}^{(j_{s_i})} = s_i$*. We also call the pair of functions* $(\mathbf{r}, \mathbf{z})$ *a* coarse ranking.[29]

When $\{(\underline{s}^{(j)}, \bar{s}^{(j)})\}_{j=1}^{\infty}$ *is empty, the induced pair* $(\mathbf{r}, \mathbf{z})$ *defines a* strict 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. When $\{(\underline{s}^{(j)}, \bar{s}^{(j)})\}_{j=1}^{\infty}$ 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 6 (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

noise term ϵ, given effort e.

[29] Intuitively, given signal profile $\boldsymbol{s}$, $r_i(\boldsymbol{s})$ is the number of agents ranked strictly above agent i, and $z_i(\boldsymbol{s})$ is the number of agents tied with agent i.

based on the strict ranking of their signals. Such rules include the three contests most commonly studied in the literature: (1) the all-pay contest, where the prize is allocated based on the ranking of the agents' effort; (2) the Lazear–Rosen contest, where the prize is allocated based on the ranking of a noisy signal s representing the agents' effort, which has the form $s = e + \epsilon$, where e is effort and ϵ is the stochastic noise; and (3) the Tullock contest, where the prize is allocated based on the ranking of a noisy signal s representing the agents' effort, described by $\log s = \log g(e) + \epsilon$, with g being some strictly increasing function and ϵ being the stochastic noise (Fu and Wu, 2019; Konrad et al., 2009).

The next proposition provides a consistency check for Definition 2: using results in Kleiner et al. (2021), we show that when the interim allocation rule is symmetric, the mapping from the signal space to the allocation space in Definition 2 does indeed, in a slightly broader sense, define a "contest" as that term is used in the economics literature. The proof is provided in Appendix A.2.

Proposition 3. *Any symmetric interim allocation–utility pair (Q, U) that is implementable by a contest has an indirect implementation that is a randomization over the coarse ranking contests.*

6 Large Contests

In many applications of interest, the number of agents participating in a contest is large. In this section we show that in such settings, the optimal contests exhibit particularly simple structures. To simplify the exposition, we make the following assumption (on top of Assumption 1) throughout the section. However, the main economic insights extend without this assumption.

Assumption 2 (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 $\theta \in [\underline{\theta}, \bar{\theta}]$.*

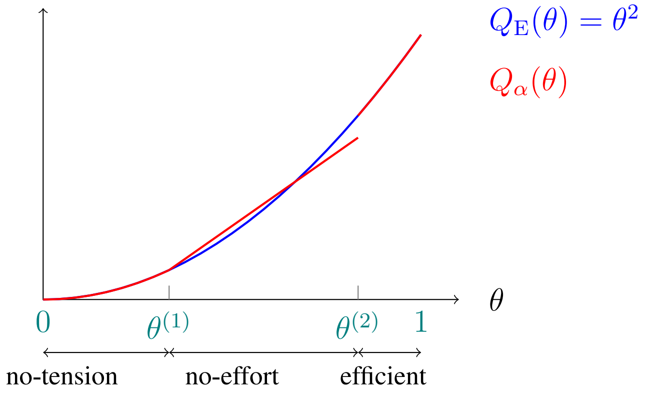
6.1 Scarce Resources

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 3 in the special case where $k = 1$ and the number of agents n is very large.[30] 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 6. *Let $k = 1$. Under Assumptions 1 and 2, 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.*

[30] 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 3: 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.

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

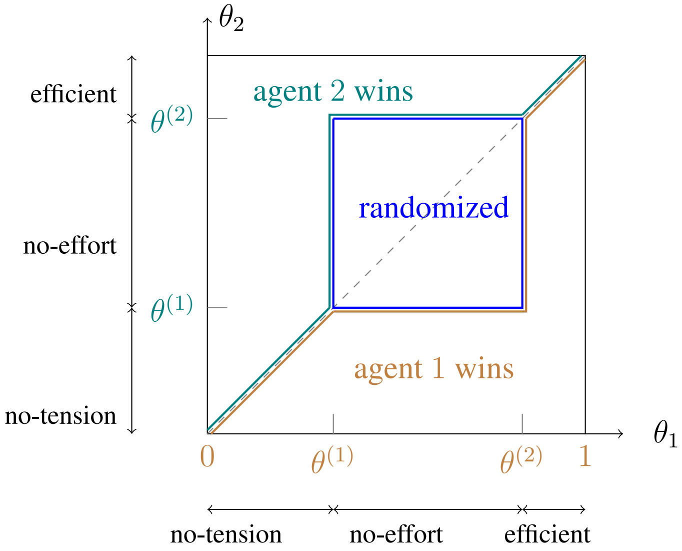
Proposition 4. *Suppose $Q_{\mathrm{E}}(\theta)$ is convex in θ. Under Assumption 1, for any $\alpha \in (0, 1)$, there exist cutoff types $\underline{\theta} \leq \theta^{(1)} \leq \theta^{(2)} \leq \bar{\theta}$ such that in the type space of each agent in the optimal contest Q_α, the interval $(\underline{\theta}, \theta^{(1)})$ is the no-tension region, $(\theta^{(1)}, \theta^{(2)})$ is the no-effort region, and $(\theta^{(2)}, \bar{\theta})$ is the efficient region.*

Figure 3 shows the optimal interim allocation rule arising from a convex efficient allocation rule in an example with two agents (as in Section 2). Figure 4 shows the contest rule that implements this optimal allocation rule. Intuitively, when the efficient allocation 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 is in sharp contrast to the classical auction design setting, where there is distortion for low types.

Convergence results. Using Lemma 6 and Proposition 4, 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.

Figure 4: 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.

Theorem 3 (convergence of contest format)**.** *Let $k = 1$. Under Assumptions 1 and 2, 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 4. 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 Definition 2 and Proposition 1, 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 Assumptions 1 and 2, 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.

Our non-convergence result also contrasts with the findings of Bulow and Klemperer (1996), who show that the principal can obtain a higher payoff by running the second-price auction with $n + 1$ agents, which is simple and implements the efficient allocation rule, than by running the optimal mechanism with n agents. In our setting, the WTA contest is the analogous simple mechanism that maximizes matching efficiency. However, it entails a loss in payoff, relative to the optimal mechanism, that cannot be offset by adding any number of agents.[31] The economic intuition for this discrepancy comes from the ambiguous effect of competition in our setting. In Bulow and Klemperer (1996), competition does not necessarily conflict with the principal's objective. In our setting, however, the principal must trade off between the two opposing interests of efficiency and the agents' utilities: competition promotes efficiency, but it also strongly incentivizes effort, which lowers the agents' utilities. The WTA contest maximizes competition, while the optimal contest creates the right level of competition to balance these two interests. This explains why the utility loss in the WTA contest cannot be offset by the increase in matching efficiency from adding any number of agents.

6.2 Large-Scale Economy

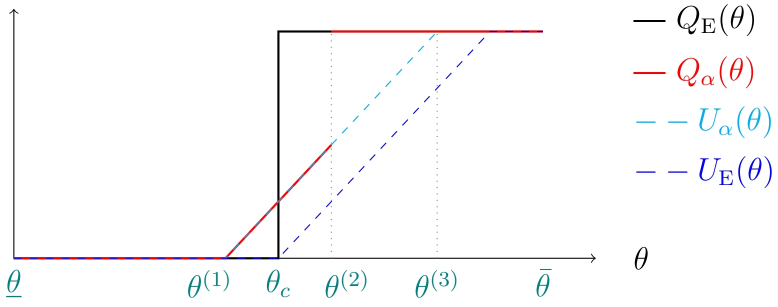
In applications such as college admissions and affordable housing programs, the resources to be allocated are not necessarily scarce. To model such situations, consider a setting with n agents and $0 < k < n$ items, and replicate both the agents and the items $z \in \mathbb{N}_+$ times. The parameter z captures the scale of the economy. As the scale z goes to infinity, the 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 Assumptions 1 and 2, 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 5. 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

[31] A caveat here is that in Bulow and Klemperer (1996) the principal maximizes revenue, while in our paper she maximizes welfare, which is defined as a weighted average of the matching efficiency and the agents' utilities.

Figure 5: 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.

$(\theta^{(1)}, \theta^{(2)})$. 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 for the benefit of 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 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 stress for all students. They also provide a theoretical analysis of when this is optimal. In our setting, however, we predict that it is optimal to randomize the allocation to types around the cutoff, rather than to low types.

7 Discussion

7.1 Implementation of Non-monotone Allocation Rules

In Theorem 1 we showed that it is without loss of optimality to focus on contests with monotone interim allocations. Here we show that there exist non-monotone interim allocations that can also be implemented as contests. The main intuition is that an agent could be indifferent between the expected allocations for a wide range of signal realizations, and hence it is possible for a high type to choose a low signal realization with low expected allocation.

Example 3. *Consider a simplified single-agent setting where the agent's type θ is drawn from a uniform distribution on $[0, 1]$, and the agent's ability is $\eta = 1$. Suppose the contest rule is $x(s) = \min\{1, s\}$. For any type $\theta \in [0, 1]$, the agent is indifferent between all signals in $[\theta, 1]$. One feasible equilibrium strategy for the agent is to choose signal $s(\theta) = 1 - \theta$ if $\theta \leq \frac{1}{2}$, and $s(\theta) = \theta$ otherwise. It is easy to verify that the induced interim allocation rule is $Q(\theta) = 1 - \theta$ if $\theta \leq \frac{1}{2}$, and $Q(\theta) = \theta$ otherwise. This interim allocation rule is strictly decreasing for types in $[0, \frac{1}{2}]$, and strictly increasing for types in $[\frac{1}{2}, 1]$.*

7.2 Cost for Downward Deviation

In this paper, we have assumed that an agent bears no cost for generating signals lower than his true type. This assumption has no bite in our model; all of our results can be directly generalized to a setting with positive costs for downward deviations. The main reason is that in the optimal contest, agents will have no incentive to generate signals lower than their true types in order to maximize their expected utilities. When positive costs are added for downward deviations, the incentive constraints remain slack and the optimal contests remain unchanged.

7.3 Concave Pareto Frontier and Linear Objective

In this paper, we have focused on the objective of maximizing a weighted average between the matching efficiency and the agents' expected utilities. An alternative is to consider the Pareto-optimal contest for these two objectives. It can then be shown that the Pareto frontier of the problem is concave, which means we can assume without loss of generality that the principal's objective function is a convex combination of the matching efficiency and the agents' expected utilities. Moreover, the concavity implies that there is no gain in randomizing over allocation rules.

Given any constant $U \geq 0$ as the lower bound of the agents' expected utilities, we define the Pareto frontier $E(U)$ as

$$\begin{aligned} E(U) = \quad & \sup_{\boldsymbol{Q}, \boldsymbol{U}} \mathbf{E}_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \theta_i \cdot Q_i(\theta_i) \right] \qquad \text{(PF)} \\ \text{s.t.} \quad & (\boldsymbol{Q}, \boldsymbol{U}) \text{ is implementable by a contest,} \\ & \mathbf{E}_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} U_i(\theta_i) \right] \geq U. \end{aligned}$$

Proposition 5. *The Pareto frontier $E(U)$ is concave.*

The proof of Proposition 5 can be found in Appendix A.5.

7.4 Asymmetric Environments

To simplify the exposition, we have stated most of the results in our paper for symmetric environments. However, our main insights (e.g., Theorems 2–5) extend to asymmetric environments, where agents are heterogeneous ex ante. Intuitively, even in an asymmetric environment, Boarder's feasibility constraint can be decomposed into n separate majorization constraints on the agents' interim allocations. Combining these with the incentive compatibility constraints for individual agents, we find that in the optimal contest for an asymmetric environment, the optimal interim allocation rule still features three distinct regions, with the partitions of the type space being different across agents.

However, in an asymmetric environment, the contest format implementing a given feasible interim allocation rule may be different from that in the symmetric case. For example, the efficient allocation rule can also be implemented in an asymmetric environment, but, in the case $k = 1$, it may not be implemented by a WTA contest as it would be in a symmetric environment.

Example 4. *Consider the setting with $n = 2$, $\Theta_1 = \Theta_2 = [0, 1]$, $\eta_1 = \eta_2 = 1$, and concave cumulative distribution functions (CDF) $F_1 \underset{FOSD}{\succsim} F_2$.*[32] *The efficient allocation is implementable by the following contest: $x_i(s_i, s_j) = 1$ if $F_i^{-1}(s_i) > F_j^{-1}(s_j)$, and $x_i(s_i, s_j) = 0$ otherwise.*[33] *Note that this is not a WTA contest.*

8 Conclusion

In this paper, we study the design of screening mechanisms to allocate limited resources to multiple agents based on manipulable signals. In the presence of competition and manipulation effects, we show that the welfare-maximizing mechanism must be a *contest*. This result is non-trivial because, ex ante, the principal might benefit from using a more sophisticated mechanism to coordinate agents' effort choices in order to eliminate unproductive effort. However, we show that while such mechanisms may provide better coordination, they also leak information, making double-deviation strategies feasible and appealing to agents. That is, they enable agents both to misreport their types and to exploit the principal's coordinating messages to update on whether it is profitable to participate. Contests rule out coordination but turn out to be welfare-maximizing.

Our paper opens the door to several possible future lines of research. First, it remains an open question under what conditions contests, or mechanisms without coordination, are optimal. Second, we have found it fruitful to study a contest design problem using mechanism design techniques. It may be possible to adopt a similar technical approach to explore related contest design questions. Third, our paper shows that a fully separating equilibrium is feasible even when the agents' preferences do not satisfy the single-crossing

[32] A function F_1 with support Θ is first-order stochastic dominant over (FOSD) another function F_2 with support Θ, denoted by $F_1 \underset{\text{FOSD}}{\succsim} F_2$, if and only if $F_1(\theta) < F_2(\theta)$ for all $\theta \in \Theta$.

[33] In this contest, each agent i's equilibrium signal choice is $\hat{s}_i(\theta_i) = F_i(\theta_i)$.

property. It would be interesting to study the existence of separating equilibria in other contexts where the single-crossing property is violated.

References

Askenazy, P., Breda, T., Moreau, F., and Pecheu, V. (March 2022). Do French companies under-report their workforce at 49 employees to get around the law? Technical Report policy brief no. 82, Institut des Politiques Publiques.

Ball, I. (2019). Scoring strategic agents. *arXiv preprint arXiv:1909.01888*.

Barut, Y. and Kovenock, D. (1998). The symmetric multiple prize all-pay auction with complete information. *European Journal of Political Economy*, 14(4):627–644.

Baye, M. R., Kovenock, D., and De Vries, C. G. (1993). Rigging the lobbying process: an application of the all-pay auction. *American Economic Review*, 83(1):289–294.

Ben-Porath, E., Dekel, E., and Lipman, B. (2023). Sequential mechanisms for evidence acquisition. *Working paper*.

Ben-Porath, E., Dekel, E., and Lipman, B. L. (2014). Optimal allocation with costly verification. *American Economic Review*, 104(12):3779–3813.

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

Bulow, J. and Klemperer, P. (1996). Auctions versus negotiations. *American Economic Review*, 86(1):180–194.

Chawla, S., Hartline, J. D., and Sivan, B. (2019). Optimal crowdsourcing contests. *Games and Economic Behavior*, 113:80–96.

Che, Y.-K. and Gale, I. L. (1998). Caps on political lobbying. *American Economic Review*, 88(3):643–651.

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

Clark, D. J. and Riis, C. (1998). Competition over more than one prize. *American Economic Review*, 88(1):276–289.

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

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

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.

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.

Greenwald, A., Oyakawa, T., and Syrgkanis, V. (2018). Simple vs optimal contests with convex costs. In *Proceedings of the 2018 World Wide Web Conference*, pages 1429–1438.

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.

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

Konrad, K. A. et al. (2009). *Strategy and Dynamics in Contests*. Oxford University Press.

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.

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

Moldovanu, B. and Sela, A. (2001). The optimal allocation of prizes in contests. *American Economic Review*, 91(3):542–558.

Moldovanu, B. and Sela, A. (2006). Contest architecture. *Journal of Economic Theory*, 126(1):70–96.

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

Mylovanov, T. and Zapechelnyuk, A. (2017). Optimal allocation with ex post verification and limited penalties. *American Economic Review*, 107(9):2666–94.

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*.

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

Siegel, R. (2009). All-pay contests. *Econometrica*, 77(1):71–92.

Siegel, R. (2010). Asymmetric contests with conditional investments. *American Economic Review*, 100(5):2230–60.

Siegel, R. (2014). Asymmetric contests with head starts and nonmonotonic costs. *American Economic Journal: Microeconomics*, 6(3):59–105.

Skaperdas, S. (1996). Contest success functions. *Economic Theory*, 7(2):283–290.

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

Zhang, M. (2023). Optimal contests with incomplete information and convex effort costs. *Theoretical Economics (to appear)*.

FOR ONLINE PUBLICATION

A Omitted Proofs

A.1 Proofs for Section 4

Proof of Lemma 1. We will prove each direction of the if-and-only-if condition separately.

Only if: If $(\boldsymbol{Q}, \boldsymbol{U})$ is implementable by a contest, then by Definition 2, there exist a signal recommendation policy $\hat{\boldsymbol{s}}$ and a contest 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 $q_i(\boldsymbol{\theta}) = x_i(\hat{s}_i(\theta_i), \hat{\boldsymbol{s}}_{-i}(\theta_{-i}))$ for all i and $\boldsymbol{\theta}$. Notice that for any signal recommendation policy $\hat{\boldsymbol{s}}$ and contest 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 $\hat{s}_i$.[34] Consider the contest 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

[34] 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.

recommendations. □

Proof of Lemma 2. We can view a general mechanism as offering a menu of randomized signal recommendations based on the reported type profiles. The partial derivative of agent i's utility with respect to his own type is between 0 and η. By the envelope theorem, the derivative of the interim utility is between 0 and η. Finally, in any general mechanism we have $U_i(\theta_i) \leq Q_i(\theta_i)$ for all θ_i and all i, because each agent's effort cost is non-negative. □

Proof of Proposition 1. For any agent i, given any monotone and interim feasible allocation $\boldsymbol{Q}$, and $\underline{u}_i \leq Q_i(\underline{\theta}_i)$, let

$$U_i(\theta_i) = \min\left\{\underline{u}_i + \eta(\theta_i - \underline{\theta}_i),\ \inf_{\theta_i' \leq \theta_i} Q_i(\theta_i') + \eta(\theta_i - \theta_i')\right\}. \tag{4}$$

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 $\theta_i' < \theta_i$. 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 contest.

Next we show that $\boldsymbol{U}$ is the unique utility profile such that $(\boldsymbol{Q}, \boldsymbol{U})$ is implementable by a contest, given utilities $\{\underline{u}_i\}_{i=1,\dots,n}$ for the lowest types. Suppose $\boldsymbol{U}^\dagger$ is a different utility profile such that $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ is implementable by a contest 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 contest.

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 contest, 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 (4).

Finally, for any $\boldsymbol{U}^\dagger$ such that $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ is implementable by a contest, if $U_i^\dagger(\underline{\theta}_i) < U_i(\underline{\theta}_i)$ for all i, then by (4) we must have $U_i^\dagger(\theta_i) \leq U_i(\theta_i)$ for all θ_i. □

Proof of Proposition 2. In a symmetric environment, let Q_{E} represent the interim allocation of a single agent for an efficient allocation rule, where the agents with the top k types each receive an item. Notice that both the WTA contest and the VCG-format mechanism implement the efficient allocation Q_{E}. Let the interim utility profiles under the WTA contest and the VCG-format mechanism be $\boldsymbol{U}^{\mathcal{C}}$ and $\boldsymbol{U}^{\mathcal{S}}$, respectively.

By Lemma 3, $U_i^{\mathcal{C}}(\theta_i) \geq U_i^{\mathcal{S}}(\theta_i)$ for all θ_i and i. It remains to show that $U_i^{\mathcal{C}}(\theta_i) > U_i^{\mathcal{S}}(\theta_i)$ for some θ_i and i. By direct calculation,

$$\begin{aligned} U_i^{\mathcal{S}}(\theta_i) &= \eta\left(\theta_i - \mathbf{E}\left[\theta_{(k+1)}|\theta_i - \frac{1}{\eta} < \theta_{(k+1)} < \theta_i\right]\right) \times \mathbf{Pr}\left[\theta_i - \frac{1}{\eta} < \theta_{(k+1)} < \theta_i\right] \\ &\qquad + \mathbf{Pr}\left[\theta_i \geq \frac{1}{\eta} + \theta_{(k+1)}\right] \\ &< \mathbf{Pr}\left[\theta_i \geq \theta_{(k+1)}\right] = Q_{\mathrm{E}}(\theta_i). \end{aligned}$$

We distinguish two cases based on the properties of the distribution.

Case 1: Suppose there exists $\theta_i \in (\underline{\theta}_i, \bar{\theta}_i)$ such that $Q_{\mathrm{E}}(\theta) < \eta(\theta_i - \underline{\theta}_i)$. In this case, according to the characterization in Lemma 1, there must also exist $\hat{\theta}_i \in (\underline{\theta}_i, \theta_i]$ such that $U_i^{\mathcal{C}}(\hat{\theta}_i) = Q_{\mathrm{E}}(\hat{\theta}_i)$. Moreover, $\hat{\theta}_i > \underline{\theta}_i$ implies that $U_i^{\mathcal{C}}(\theta_i) > U_i^{\mathcal{S}}(\theta_i)$. Since there exists a positive measure above $\hat{\theta}_i$, the expected welfare under the VCG-format mechanism is strictly lower.

Case 2: Suppose $Q_{\mathrm{E}}(\theta) \geq \eta(\theta_i - \underline{\theta}_i)$ for all $\theta_i \in [\underline{\theta}_i, \bar{\theta}_i]$. In this case, Lemma 1 implies that $U_i^{\mathcal{C}}(\theta_i)$ is linear with slope η for any $\theta \in [\underline{\theta}_i, \bar{\theta}_i]$, i.e., $U_i^{\mathcal{C}}(\theta_i) = \eta(\theta_i - \underline{\theta}_i)$. It remains to show that $U_i^{\mathcal{S}}(\theta_i)$ is not linear. Notice that

$$\begin{aligned} U_i^{\mathcal{S}}(\theta_i) &= \eta\left(\theta_i - \mathbf{E}\left[\theta_{(k+1)}|\theta_i - \frac{1}{\eta} < \theta_{(k+1)} < \theta_i\right]\right) \times \mathbf{Pr}\left[\theta_i - \frac{1}{\eta} < \theta_{(k+1)} < \theta_i\right] \\ &\qquad + \mathbf{Pr}\left[\theta_i \geq \frac{1}{\eta} + \theta_{(k+1)}\right] \\ &= \eta\left(\theta_i - \mathbf{E}\left[\theta_{(k+1)}|\theta_i - \frac{1}{\eta} < \theta_{(k+1)} < \theta_i\right]\right) \times \left[Q_{\mathrm{E}}(\theta_i) - Q_{\mathrm{E}}(\theta_i - \frac{1}{\eta})\right] + Q_{\mathrm{E}}(\theta_i - \frac{1}{\eta}) \\ &= \eta\left(\theta_i \times [Q_{\mathrm{E}}(\theta_i) - Q_{\mathrm{E}}(\theta_i - \frac{1}{\eta})] - \mathbf{E}\left[\theta_{(k+1)}\mathbf{1}\left[\theta_i - \frac{1}{\eta} < \theta_{(k+1)} < \theta_i\right]\right]\right) + Q_{\mathrm{E}}(\theta_i - \frac{1}{\eta}) \\ &= \eta\int_{\theta_i - \frac{1}{\eta}}^{\theta_i} Q_{\mathrm{E}}(\theta)d\theta, \end{aligned}$$

where the last equality is obtained by applying integration by parts, i.e.,

$$\mathbf{E}\left[\theta_{(k+1)}\mathbf{1}\left[\theta_i - \frac{1}{\eta} < \theta_{(k+1)} < \theta_i\right]\right] = \theta_i Q_{\mathrm{E}}(\theta_i) - (\theta_i - \frac{1}{\eta})Q_{\mathrm{E}}(\theta_i - \frac{1}{\eta}) - \int_{\theta_i - \frac{1}{\eta}}^{\theta_i} Q_{\mathrm{E}}(\theta)d\theta.$$

Note that since $Q_{\mathrm{E}}(\theta) \leq 1$ for $\theta_i \in [\underline{\theta}_i, \bar{\theta}_i]$ and $Q_{\mathrm{E}}(\theta) = 0$ for $\theta_i < \underline{\theta}_i$, in order for $U_i^{\mathcal{S}}(\theta_i)$ to attain the upper bound of $\eta(\theta_i - \underline{\theta}_i)$, we must have $Q_{\mathrm{E}}(\theta) = 1$ for all $\theta_i \in [\underline{\theta}_i, \underline{\theta}_i + \frac{1}{\eta}]$, and thus, by monotonicity, $Q_{\mathrm{E}}(\theta) = 1$ for all θ_i. However, this is not possible if the type distribution is continuous with positive density everywhere in its support. □

A.2 Proofs for Section 5

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 Lemma 5. 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 $U_i(\theta_i) \leq Q_i(\theta_i)$ 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 1. Therefore, $(\boldsymbol{Q}, \boldsymbol{U})$ is also feasible and hence is an optimal solution for Problem $(\mathcal{P}_\alpha)$. □

Proof of Theorem 2. By Lemma 1, 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 7. *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 8. *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 7. 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,[35]

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

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.

(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 $\kappa'(\theta) \neq 0$

[35] 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.

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 7 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 8. 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 ($\widehat{\text{IF}}$) 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 ($\widehat{\text{IF}}$) is violated at type $\bar{\theta}^{(j)} - \epsilon$, a contradiction.

- Interval j' belongs to Case 2. In this case, ($\widehat{\text{IF}}$) 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 ($\widehat{\text{IF}}$) to hold for type $\underline{\theta}^{(j')} + \epsilon$ given sufficiently small $\epsilon > 0$. However, this would imply that ($\widehat{\text{IF}}$) 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 $Q(\theta) = U(\theta) > Q_{\mathrm{E}}(\theta)$. 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 $\theta \in [\underline{\theta}^{(j)} - \epsilon, \bar{\theta}^{(j)} + \epsilon]$,

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

    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 $(\widehat{\text{IF}})$ 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. □

Proof of Proposition 3. For any symmetric interim allocation–utility pair (Q, U) that is implementable by a contest, 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 $\hat{s}$.[36] The distribution over types $F(\theta)$ induces a distribution over signals; call it $\hat{F}(s)$. Similarly, a feasibility constraint on the contest rule defined in the signal space may be induced by $(\widehat{\text{IF}})$. 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 contest 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 6. 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; hence the expected utility of each agent in the coarse ranking contest is still U. □

[36] Notice that such a mapping might not exist if (Q, U) is implementable by a general mechanism that is not a contest.

A.3 Proofs for Section 6.1

Proof of Lemma 6. 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 4. 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 contest.

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 contest 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 6, for sufficiently large n, the efficient allocation rule is convex. Therefore, the interim allocation rule of the optimal contest takes the form described in Proposition 4.

Let $Q_{\alpha,n}(\theta)$ and the $Q_{\mathrm{E},n}(\theta)$ be the optimal interim allocation rule and efficient allocation rule in a contest 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 9, whose proof is given later in this section. Lemma 9 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 9. *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 9, 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 9 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 contest rule converges to the optimal welfare when the number of agents is sufficiently large.

Lemma 10. *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 9. 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 2, 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

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

Proof of Lemma 10. 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 $n \geq N_1$, $n \cdot \mathbf{E}_{\theta \sim F}[\theta \cdot Q_n(\theta)] \geq \bar{\theta} - \epsilon$. □

A.4 Proof of Theorem 5

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.[37] 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 $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

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

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

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 11. *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

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

Similarly,

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

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

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

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 $\theta^{(3)} \ge \theta^{(2)}$ 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 11. It is sufficient to show that any contest $(\tilde{Q}_{\alpha,z}, \tilde{U}_{\alpha,z})$ such that $\tilde{U}_{\alpha,z}(\tilde{\theta}^{(1)}) \le Q_{\mathrm{E},z}(\tilde{\theta}^{(1)})$ 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:[38]

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

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

Consider a contest $(\hat{Q}_{\alpha,z}, \hat{U}_{\alpha,z})$ characterized by three cutoffs $\hat{\theta}^{(1)} < \hat{\theta}^{(2)} \le \hat{\theta}^{(3)}$ 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

[38] 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.

efficient region.

Step 1: In this step, we will show that if $\hat{\theta}^{(1)}$ is chosen so that $\theta_c - \frac{\underline{\beta}_1}{10\eta\cdot\bar{\beta}_1} \leq \hat{\theta}^{(1)}$,[39] 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

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

where the second inequality holds because the interim allocations are bounded within $[0, 1]$, and the last inequality holds by the continuity assumption (Assumption 2).

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 $\hat{U}_{\alpha,z}$

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

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. □

A.5 Proof of Proposition 5

Proof of Proposition 5. Consider any $U', U'' \geq 0$, $\lambda \in [0,1]$. Suppose $(\boldsymbol{Q}', \boldsymbol{U}')$ and $(\boldsymbol{Q}'', \boldsymbol{U}'')$ attain the values $E(U')$ and $E(U'')$ defined in Problem (PF).

Note that for any $\lambda \in [0,1]$, $\lambda\boldsymbol{Q}' + (1-\lambda)\boldsymbol{Q}''$ is still interim feasible. Although the allocation–utility pair $(\lambda\boldsymbol{Q}' + (1-\lambda)\boldsymbol{Q}'', \lambda\boldsymbol{U}' + (1-\lambda)\boldsymbol{U}'')$ may violate the IC constraints, by Theorem 1, there exists another utility profile $U^\dagger$ such that (1) $U_i^\dagger(\theta_i) \geq \lambda U_i'(\theta_i) + (1-\lambda)U_i''(\theta_i)$ for any agent i and any type θ_i, and (2) $(\lambda\boldsymbol{Q}' + (1-\lambda)\boldsymbol{Q}'', U^\dagger)$ is implementable by a contest. This implies that $\lambda\boldsymbol{Q}' + (1-\lambda)\boldsymbol{Q}''$ is a feasible solution to Problem (PF) when $U = \lambda U' + (1-\lambda)U''$. Therefore, $E(\lambda U' + (1-\lambda)U'') \geq \lambda E(U') + (1-\lambda)E(U'')$. □

B Nonlinear Costs

In the main text we have shown that when effort costs are linear, contests are optimal among general mechanisms. In this section we show that our result extends to convex cost functions.

B.1 Optimality of Contests among Monotone Allocations

Let $c_i(s_i|\theta_i)$ be the effort cost for agent i with type θ_i producing a signal s_i.

Assumption 3. *For any agent i, we have $c_i(s_i|\theta_i) = C_i((s_i - \theta_i)^+) = C_i(e_i)$ for all $e_i \geq 0$.*

Assumption 4. *For any agent i, we have $C_i''(e_i) \geq 0$ and $C_i'''(e_i) \leq 0$.*

Note that Assumption 4 is satisfied by a quadratic effort cost $C_i(e_i) = \frac{1}{2}e_i^2$.

Lemma 12. *Under Assumptions 3 and 4, for any agent i, any $\epsilon \geq 0$, any distribution G supported on $\mathbb{R}_+$, and any constant e_G such that $C_i(e_G) = \mathbf{E}_{e\sim G}[C_i(e)]$, we have $C_i(\epsilon + e_G) \geq \mathbf{E}_{e\sim G}[C_i(\epsilon + e)]$; that is, for any θ, $c_i(\theta + e_G + \epsilon|\theta) \geq \mathbf{E}_{e\sim G}[c_i(\theta + e + \epsilon|\theta)]$.*

Proof. Let $\Delta_i(\epsilon) \triangleq C_i(\epsilon + e_G) - \mathbf{E}_{e\sim G}[C_i(\epsilon + e)]$. Note that by the definition of e_G, we have $\Delta_i(0) = 0$. Therefore, to prove Lemma 12, it is sufficient to show that $\Delta_i'(\epsilon) \geq 0$ for any $\epsilon \geq 0$. Let the expected effort level be $\mu_G \triangleq \int e \, \mathrm{d}G(e)$; since the cost function C_i is convex, we have $e_G \geq \mu_G$. Therefore,

$$\Delta_i'(\epsilon) = C_i'(\epsilon + e_G) - \mathbf{E}_{e\sim G}\big[C_i'(\epsilon + e)\big] \geq C_i'(\epsilon + \mu_G) - \mathbf{E}_{e\sim G}\big[C_i'(\epsilon + e)\big] \geq 0,$$

where the inequalities hold because C_i' is increasing and concave for any i by Assumption 4. □

Intuitively, Lemma 12 states that when the stochastic effort recommendation drawn from G is replaced by a deterministic effort recommendation, given by its certainty-equivalent effort level e_G, each type's effort cost weakly increases if he misreports as a higher type. Lemma 12 lies at the heart of the proof that contests are optimal, and Assumption 4 is one sufficient assumption that guarantees this property.

To simplify the exposition in the later analysis, we introduce one more assumption and some notation.

Assumption 5. *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 $\hat{\theta}_i^{(0)} < \dots < \hat{\theta}_i^{(m)}$.*

For any agent i with type θ_i, let his expected utility from generating signal s_i and receiving allocation x_i be $\tilde{U}_i(\theta_i; x_i, s_i) \triangleq x_i - C_i((s_i - \theta_i)^+)$. For a stochastic mechanism, let D_i be the distribution over the allocation–signal pair, and define the expected utility of agent i with type θ_i as $\tilde{U}_i(\theta_i; D_i) \triangleq \mathbf{E}_{(x_i,s_i)\sim D_i}\Big[\tilde{U}_i(\theta_i; x_i, s_i)\Big]$. Notice that a stochastic mechanism is only implementable by a general mechanism that is not a contest. In contests, it is without loss to focus on deterministic signal recommendations.

Theorem 6. *Under Assumptions 3, 4, and 5, for any interim allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ where $\boldsymbol{Q}$ is monotone, if $(\boldsymbol{Q}, \boldsymbol{U})$ is implementable by a general mechanism, then there exists another utility profile $\boldsymbol{U}^\dagger$ such that $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ is implementable by a contest. Moreover, $U_i^\dagger(\theta_i) \geq U_i(\theta_i)$ for any agent i with type θ_i.*

Proof. For any agent i and any monotone interim allocation Q_i, we construct the utility function $U_i^\dagger$ by induction. First let $U_i^\dagger(\hat{\theta}_i^{(0)}) = Q_i(\hat{\theta}_i^{(0)})$. For any $k \geq 1$, let $s_i^{(k)}$ be the signal such that $U_i^\dagger(\hat{\theta}_i^{(k-1)}) = \tilde{U}_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k)}), s_i^{(k)})$. That is, the agent with type $\hat{\theta}_i^{(k-1)}$ is indifferent between reporting truthfully to receive his expected utility under $U_i^\dagger$ and deviating to the option $(s_i^{(k)}, Q_i(\hat{\theta}_i^{(k)}))$, i.e., generating signal $s_i^{(k)}$ and receiving allocation $Q_i(\hat{\theta}_i^{(k)})$.

By construction, the interim allocation and utility $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ coincide with the utility that agent i with type $\hat{\theta}_i^{(k)}$ would receive if he gets allocation $Q_i(\hat{\theta}_i^{(k)})$ as long as he generates signal $s_i^{(k)}$. It remains to show that this is an equilibrium of a contest, i.e., no agent and no type has strict incentive to deviate (**Step 1**), and that the corresponding interim utility $U_i^\dagger(\hat{\theta}_i^{(k)})$ is weakly higher than $U_i(\hat{\theta}_i^{(k)})$ (**Step 2**):

- **Step 1**: *Incentive compatibility.* By construction, the local incentive compatibility is guaranteed, i.e., for each agent and each type, there is no strict incentive to deviate to an adjacent type. Thus, we only need to make sure there is global incentive compatibility. This is guaranteed if the agent's utility satisfies the standard single-crossing property in the type–allocation space. For a convex cost function, this property is satisfied for menu options with deterministic signal recommendations.

- **Step 2**: *Higher utilities.* We prove this by induction. First note that $U_i^\dagger(\hat{\theta}_i^{(0)}) = Q_i(\hat{\theta}_i^{(0)}) \geq U_i(\hat{\theta}_i^{(0)})$. For any $k \geq 1$, let $\tilde{s}_i^{(k)}$ be the signal such that $U_i(\hat{\theta}_i^{(k)}) = \tilde{U}_i(\hat{\theta}_i^{(k)}; Q_i(\hat{\theta}_i^{(k)}), \tilde{s}_i^{(k)})$. Let $u_{k-1,k}$ be the expected utility of the agent with type $\hat{\theta}_i^{(k-1)}$ if he misreports as $\hat{\theta}_i^{(k)}$. Notice that $\tilde{s}_i^{(k)} - \hat{\theta}_i^{(k)}$ is the equivalent deterministic effort recommendation, while the effort recommendation for reported type $\hat{\theta}_i^{(k)}$ is potentially stochastic. Thus Lemma 12 implies that $u_{k-1,k} \geq \tilde{U}_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k)}), \tilde{s}_i^{(k)})$. Recall that by construction, $\tilde{U}_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k)}), s_i^{(k)}) = U_i(\hat{\theta}_i^{(k-1)})$, and by incentive compatibility, $U_i(\hat{\theta}_i^{(k-1)}) \geq u_{k-1,k}$. Combining the two, we have that $s_i^{(k)} \leq \tilde{s}_i^{(k)}$. Therefore,

$$U_i^\dagger(\hat{\theta}_i^{(k)}) = \tilde{U}_i(\hat{\theta}_i^{(k)}; Q_i(\hat{\theta}_i^{(k)}), s_i^{(k)}) \geq \tilde{U}_i(\hat{\theta}_i^{(k)}; Q_i(\hat{\theta}_i^{(k)}), \tilde{s}_i^{(k)}) = U_i(\hat{\theta}_i^{(k)}). \qquad \square$$

Theorem 6 implies that contests are optimal among general mechanisms with monotone interim allocations. The monotonicity constraint is a reasonable assumption, because in practice it may be perceived as unfair to provide higher allocations to lower types.[40] 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.

B.2 Optimality of Contests among General Allocations

In Section 7.1 we showed that there exist non-monotone interim allocation rules that are implementable by general mechanisms. This is because the agent's utility does not satisfy the single-crossing property in general mechanisms. However, Theorem 1 showed that such non-monotone interim allocation rules are not optimal when costs are linear. In Theorem 7, we generalize this idea by showing that under an additional assumption, which we call "no concave crossing," non-monotone interim allocation rules are not optimal.

Assumption 6 (no concave crossing). *For any agent i, any x_i, s_i, and any D_i, if there exist $\hat{\theta}_i < \hat{\theta}_i'$ such that $\tilde{U}_i(\hat{\theta}_i; x_i, s_i) \leq \tilde{U}_i(\hat{\theta}_i; D_i)$ and $\tilde{U}_i(\hat{\theta}_i'; x_i, s_i) \leq \tilde{U}_i(\hat{\theta}_i'; D_i)$, then we have $\tilde{U}_i(\theta_i; x_i, s_i) \leq \tilde{U}_i(\theta_i; D_i)$ for any $\theta_i \in [\hat{\theta}_i, \hat{\theta}_i']$.*

[40] Monotonicity constraints are imposed similarly in the application of insurance contracts owing to practical concerns (Gershkov et al., 2022).

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.

Theorem 7. *Under Assumptions 3, 4, 5, and 6, for any interim allocation–utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ that is implementable by a general mechanism, there exists $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ with monotone $\boldsymbol{Q}^\dagger$ that is implementable by a contest and yields a weakly higher objective value.*

Proof. As in the proof of Theorem 1, let $\boldsymbol{Q}^\dagger$ be a monotone rearrangement of $\boldsymbol{Q}$ that is feasible and weakly improves matching efficiency. We construct $\boldsymbol{U}^\dagger$ as in Theorem 6 so that $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ is implementable by a contest. It remains to show that $\boldsymbol{U}^\dagger$ weakly improves the agents' utilities for all types under Assumption 6.

Since $\boldsymbol{Q}^\dagger$ is a monotone rearrangement of $\boldsymbol{Q}$, for any θ_i in the support of F_i, there exists another type $\theta_i' \geq \theta_i$ in the support such that $Q_i(\theta_i') \leq Q_i^\dagger(\theta_i)$. Therefore, for any agent i, since U_i is non-decreasing and no greater than the allocation Q_i, we have

$$U_i^\dagger(\hat{\theta}_i^{(0)}) = Q_i^\dagger(\hat{\theta}_i^{(0)}) \geq \min_{\theta_i} Q_i(\theta_i) \geq U_i(\hat{\theta}_i^{(0)}).$$

For any $k \geq 1$, if $Q_i^\dagger(\hat{\theta}_i^{(k)}) \geq Q_i(\hat{\theta}_i^{(k)})$, then similarly to Theorem 6, we have $U_i^\dagger(\hat{\theta}_i^{(k)}) \geq U_i(\hat{\theta}_i^{(k)})$ and we are done.

If instead $Q_i^\dagger(\hat{\theta}_i^{(k)}) < Q_i(\hat{\theta}_i^{(k)})$, then there exists $k' > k$ such that $Q_i^\dagger(\hat{\theta}_i^{(k)}) \geq Q_i(\hat{\theta}_i^{(k')})$. Let s_i^k and $\tilde{s}_i^{(k')}$ be the signals such that

$$\begin{aligned} U_i^\dagger(\hat{\theta}_i^{(k-1)}) &= \tilde{U}_i(\hat{\theta}_i^{(k-1)}; Q_i^\dagger(\hat{\theta}_i^{(k)}), s_i^{(k)}), \\ U_i^\dagger(\hat{\theta}_i^{(k-1)}) &= \tilde{U}_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i^{(k')}). \end{aligned}$$

Note that $Q_i^\dagger(\hat{\theta}_i^{(k)}) \geq Q_i(\hat{\theta}_i^{(k')})$ implies $s_i^{(k)} \geq \tilde{s}_i^{(k')}$.

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

- $\tilde{U}_i(\hat{\theta}_i^{(k')}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i^{(k')}) \geq \tilde{U}_i(\hat{\theta}_i^{(k')}; D_i^{(k)})$. This is because

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

  where the first inequality is implied by Lemma 12 and the second is implied by incentive compatibility.

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

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

> The first inequality holds by the induction assumption. The second inequality holds because $\tilde{U}_i(\hat{\theta}_i^{(k-1)}; D_i^{(k)})$ 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 6 that

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

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

$$U_i^\dagger(\hat{\theta}_i^{(k)}) = \tilde{U}_i(\hat{\theta}_i^{(k)}; Q_i^\dagger(\hat{\theta}_i^{(k)}), s_i^{(k)}) \geq \tilde{U}_i(\hat{\theta}_i^{(k)}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i^{(k')}) \geq \tilde{U}_i(\hat{\theta}_i^{(k)}; D_i^{(k)}) = U_i(\hat{\theta}_i^{(k)}). \qquad \square$$

There are two caveats in the results of Theorems 6 and 7. First, our construction only works for distributions with finite support. We have extended the argument to continuous distributions in the setting with linear costs (cf. Theorem 1). It is not clear whether a similarly general argument exists for convex costs. Second, our no-concave-crossing assumption (Assumption 6) is only sufficient, not necessary, for showing that contests are optimal among non-monotone mechanisms. Moreover, this assumption is hard to interpret in practice. We conjecture that there exist weaker and more easily interpreted conditions for ensuring the optimality of contests. We leave these as interesting open questions for future work.