Screening with Disadvantaged Agents

Hedyeh Beyhaghi[1]
Carnegie Mellon University, Pittsburgh, PA, USA

Modibo K. Camara
University of Chicago, IL, USA

Jason Hartline
Northwestern University, Evanston, IL, USA

Aleck Johnsen[1]
Geminus Research, Cambridge, MA, USA

Sheng Long
Northwestern University, Evanston, IL, USA

Abstract

Motivated by school admissions, this paper studies screening in a population with both advantaged and disadvantaged agents. A school is interested in admitting the most skilled students, but relies on imperfect test scores that reflect both skill and effort. Students are limited by a budget on effort, with disadvantaged students having tighter budgets. This raises a challenge for the principal: among agents with similar test scores, it is difficult to distinguish between students with high skills and students with large budgets.

Our main result is an optimal stochastic mechanism that maximizes the gains achieved from admitting "high-skill" students minus the costs incurred from admitting "low-skill" students when considering two skill types and n budget types. Our mechanism makes it possible to give higher probability of admission to a high-skill student than to a low-skill, even when the low-skill student can potentially get higher test-score due to a higher budget. Further, we extend our admission problem to a setting in which students uniformly receive an exogenous subsidy to increase their budget for effort. This extension can only help the school's admission objective and we show that the optimal mechanism with exogenous subsidies has the same characterization as optimal mechanisms for the original problem.

2012 ACM Subject Classification Applied computing → Economics; Theory of computation → Algorithmic mechanism design

Keywords and phrases screening, strategic classification, budgeted mechanism design, fairness, effort-incentives, subsidies, school admission

Digital Object Identifier 10.4230/LIPIcs.FORC.2023.6

Related Version *Full Version*: `https://arxiv.org/abs/2305.18068`

Funding *Jason Hartline*: Supported in part by NSF CCF 1934931.
Sheng Long: Supported in part by NSF CCF 1934931.

1 Introduction

Screening is a problem in which a *principal* desires to select only a qualified sub-population of *agents* who exceed an appropriate threshold applied to the agents' *private types*.

Many real-world problems may be interpreted as special cases of screening problems, including some well-studied problems in standard frameworks (for example within auction design, how to give away an item to an agent who values it the most). As further examples: school admissions, hiring employees, selecting romantic partners, identifying winners of

[1] Corresponding authors

prestigious awards, qualifying applicants for government-issued licenses, assigning school grades at any level of evaluation (from homework grades to testing grades to overall-course grades), drug-testing, tournament-qualifying, ..., all of these and many more scenarios may be modeled as problems of screening.

The challenge of screening is that the principal has only indirect access to the agents' private types, and critically, the agents are either unwilling to reveal their types or are incentivized to take actions that make it difficult for the principal to infer their types. Since these agents are *strategic*, their private information is only fully or partially elicited by offering appropriate incentives.

This paper considers a screening model of *school admissions* where some students may be disadvantaged relative to others. The school seeks the most skilled students but only has access to an imperfect measure of skill, via test scores. Relative to their inherent skill, disadvantaged students may perform worse on tests because they have less time to prepare (e.g., due to work obligations or childcare). Advantaged students may perform better on tests relative to their skill because they have access to additional resources (e.g., a private tutor or test prep). We model this heterogeneity by assuming that applicants are distinguished both in their skill as well as their *budget* (i.e., how much time and resources they are able to put towards the test). Students with high skill and high budget are able to excel, provided that they are willing to put in the effort. However, students with similarly high skill may test poorly if their budgets are too low.

More precisely, we study a mechanism design problem for screening of budgeted agents. A principal is interested in admitting only an agent with high skill-type above a given threshold. The agent can only reveal private skill-level to the principal indirectly, by combining it with an amount of effort into a publicly displayable signal of quality. However, the agent is limited by a budget on effort, which induces a key difficulty for the principal: amongst agent types exhibiting similar-quality signals, how to distinguish between talented agents with high-skill-low-budget types and endowed agents with low-skill-high-budget types, while contending with agents' incentive-compatability constraints.

A key observation from the model is that it may be beneficial to admit students with average test scores with nonzero probability, while at the same time always admitting students with the highest test scores. By not guaranteeing admission for students with average test scores, we limit the incentive for those students to put in effort. High-skill agents (regardless of their budget) find effort less costly than low-skill students; therefore, as we decrease the probability of admission, the low-skill students will reduce their effort more sharply than high-skill students. Loosely speaking, if we lower the probability of admission enough for students with average test scores, the equilibrium level of effort will drop until the high-skill disadvantaged students' budget constraint is no longer binding. This allows the school to screen efficiently, at the cost of admitting high-skill students at a lower rate.

As a result, these randomized admission policies make it possible to implement a counter-intuitive outcome. A student with high skill but low maximum test score (due to limited budget) can receive strictly larger allocation than a student with low skill but high maximum test score. The latter student is able to achieve scores that are strictly higher than the former student can achieve, but the benefit of obtaining those scores (some probability of admission) is not worth the effort for a low-skill student.

Our main result formalizes this intuition. It gives (1) a characterization of the structure of the optimal mechanism for a (one-agent) setting with 2 skill types and n budget types, and (2) a polynomial-time algorithm to find it. An interpretation of our main result is that high-skill agent types may be shown *preference* over high-budget types despite the difference in the types' *exogenous* resources. Thus, our setting effectively studies the possibilities and limits of improved-welfare of allocation to effort-budgeted agents.

The paper ends with an introductory study of an extended setting which introduces *uniform, exogenous, unconditional subsidies* to relax the agents' budget constraints.[2] Intuitively, the goal is to modify the environment of the admissions problem (as screening) to further increase the balance of allocation in favor of high-skilled types. Subsidies are a potent intervention because high-skill, budget-constrained agents are best able to use additional effort to increase their highly-valued allocations. We show that the setting with subsidies has optimal mechanisms with the same characterization as the original screening problem.

Related Works

Previous literature has varied its modeling of this central challenge of screening. [27] models agents as having private abilities (types) that the market doesn't observe, and agents with higher abilities have economic incentives to be identified. [26] studied the role of interest rates as a screening device, and showed that returns are not necessarily monotone with respect to interest rates – a result that holds in equilibrium whenever borrowers *strategically* react to the interest-rate mechanism.

In addition to the economics literature on screening, this work contributes to ongoing research on strategic classification, mechanism design with budgeted agents, and fairness.

There is a well-developed literature on mechanism design where agents face budget constraints. Earlier work focused on the case where budgets were public knowledge (e.g., [20, 21]). More recent work, like ours, focuses on the case where the agents' budgets are their private knowledge (e.g., [24, 11, 9]). Typically, budgets are monetary: they represent upper bounds on how much each agent can transfer to the principal. In contrast, we consider budgets on effort: upper bounds on how much effort the agent can put into its task.

In recent years, there has been a lot of interest in strategic classification problems, where a principal is trying to classify agents on the basis of observed scores and agents are able to manipulate (or "game") the scores to influence the principal's actions [13, 8, 15, 23, 1, 6, 10, 5, 18, 14, 3, 28, 22, 12, 4, 2]. Our model can be considered a strategic classification problem where the school attempts to classify students into "admit" or "not admit", but students are strategic in how much effort they put in. The closest to our work are [15] and [5]. [5] show the power of randomization when agents are able to manipulate their scores, while [15] study a similar problem where disadvantaged students find it more difficult to manipulate their scores.

In most models of strategic classification, agents obscure their true type at a cost. As a result, costly effort makes scores less informative. In contrast, in our model of screening, even high skill students need to put in effort in order to achieve a high score (albeit less effort than low-skill students). If no students put in effort, they will all achieve a score of zero, and the school will not be able to distinguish high-skill from low-skill students. As a result, costly effort is necessary for scores to be informative in our model. We must balance the benefits of costly effort in screening with the challenges of costly effort in strategic classification.

Finally, our work relates to a growing literature on fairness in mechanism design and algorithms. Much of this literature is concerned with fair treatment of different subgroups (e.g., based on demographic variables like race or gender), and various different definitions of fairness have been proposed and criticized (see e.g., [7, 19]). Some of this work, like ours, has been explicitly applied to school admissions (e.g., [17]). In line with the fairness literature,

[2] Subsidies, measured in units of effort, can for example be monetary transfers from third-parties that increase an agent's effort-budget by freeing up time by reducing other paid work or by buying services.

we consider the implications of a biased test (where high budget students may perform better, regardless of their skill) for admissions. Unlike race and gender, the subgroups we are interested in (students with a particular budget) are not publicly observable. Like [25] and [16], we explicitly consider how economic incentives interact with policies designed to correct for sources of unfairness.

2 Setting and Fundamental Structures

A principal P considers admitting an agent $A = (s, b)$ with private types as skill s and budget b (budget on *effort*, see below). The agent's skill and budget are treated as independent, positive Bayesian variables drawn respectively from known distributions S with support $\mathbf{S} \in \mathbb{R}_+$ and B with support $\mathbf{B} \in [0, 1]$, i.e., $s \sim S$ and $b \sim B$. The principal only wants to admit the agent in the case that the agent's skill is above a threshold $\tau \in \mathbb{R}_+$ (which we implicitly treat as the principal's fixed type). In summary, the principal's problem is an admission game $\mathcal{G} = (S, B, \tau)$.

An agent of any skill will want to be admitted and thus, the principal must design a test which uses incentives to elicit information from the agent. The principal will ask the agent to commit to a *private* level of effort e which (a) is constrained by individual budget b, and (b) induces a *public*, deterministic signal of quality $q = s \cdot e$. Note that quality is a multiplicative function of effort, rather than additive. This captures two features of our motivating example of school admissions: (i) even high-skill agents that put in no effort will obtain a low score, but (ii) high-skill agents require less effort to achieve a given score than low-skill agents.[3]

The principal's problem will be to design an admission allocation rule $y : \mathbb{R}_+ \to [0, 1]$ which maps quality q to a stochastic allocation x of admitting the agent. Practically, the principal's challenge is to optimally discriminate against resource-rich agent types with quality resulting from large effort-budgets, in favor of agents with quality resulting from high skill. (Note, any "reasonable" rule will inherently admit all high-skill-high-budget agents.)

The agent's utility is defined to be $-\infty$ if effort exceeds budget, and otherwise is defined to be the probability of allocation minus effort:

$$u_A(e, x) = x - e \tag{1}$$

which implicitly sets the agent's value of being admitted to 1. Utility can be equivalently written as a function of the allocation rule y and either effort or quality:

$$u_A(y, e) = y(s \cdot e) - e \qquad \text{or} \qquad u_A(y, q) = y(q) - q/s \tag{2}$$

(Further, we may drop the input y where its assignment is clear from context.) The agent perceives the allocation rule y as a menu (for which the domain is quality space), albeit top-truncated at the agent's maximum quality set by $q^\dagger = s \cdot b$. This perspective induces for the agent an optimal utility function u_A^* and an allocation rule x in skill space (which overloads notation):

$$u_A^*(y, s) := \max_{e \in [0,b]} y(s \cdot e) - e = \max_{q \in [0, q^\dagger]} y(q) - q/s \tag{3}$$

$$x = x(y, s) := y(s \cdot \left[\operatorname*{argmax}_{e \in [0,b]} y(s \cdot e) - e \right]) \tag{4}$$

[3] In contrast, suppose quality were an additive function $q = s + e$ of skill and effort. Then property (ii) would hold, but not property (i).

For a given agent A, the principal's utility from admitting A is $u_P(A \mid \text{admitted}) = s - \tau$. Thus, our principal's mechanism design problem is to maximize $u_P(\mathcal{G}, y)$ which is the expected utility from an admitted agent's skill versus the threshold, weighted by allocation probability:

$$\max_y u_P(\mathcal{G}, y) := \max_y \mathbf{E}_{A\sim(S\times B)}\left[x(y,s)\cdot(s-\tau)\right] \tag{5}$$

Threshold Mechanisms

A natural mechanism to consider is a *threshold mechanism* with the threshold set in quality space.

▶ **Definition 1.** *A* (deterministic) threshold mechanism $y^{q'}$ *sets a quality threshold* $q' \in \mathbb{R}_+$ *and admits an agent if and only if the agent exhibits public quality* $q \geq q'$.

The intuition for a threshold mechanism is that an agent who is able to exhibit the threshold quality with effort less than budget will put in the (minimal amount of) effort necessary to be admitted with probability 1; versus, an agent with maximum-quality $q^\dagger$ less than the threshold will put in zero effort and get passed. Recall that the agent's skill and budget are independent in our setting. The role of thresholds generally is to conditionally allocate agents in decreasing order of skill:

▶ **Fact 2.** *Given a population of agents as* $S \times B$*, consider the subset* $\mathbf{B}_{\bar{b}}$ *of agent types which conditionally have a specific budget* $\bar{b}$*. For a threshold mechanism with any* $q' > 0$*, the subset of* $\mathbf{B}_{\bar{b}}$ *of agent skill-types which are admitted is upward-closed.*

Fact 2 implies that threshold mechanisms are sufficient for the special case in which there is only one budget type (with the proof of Proposition 3 in Appendix A.1):

▶ **Proposition 3.** *Assume that an agent has constant budget* $\bar{b}$ *on effort, i.e., the distribution* B *is a singular point mass. The threshold mechanism* $y^{q'}$ *with* $q' = \tau \cdot \bar{b}$ *is optimal.*

Intuitively, Proposition 3 holds because single-budget is a simple setting in which quality-thresholds directly implement skill-thresholds, in particular for the principal's threshold τ.

To outline this section: Section 2.1 shows that threshold mechanisms are not optimal for arbitrary distributions B and thus, we will need more-complicated mechanism forms. Section 2.2 quantifies agent feasibility to achieve a given quality-allocation pair and, given an allocation rule y, discusses implications of feasibility for optimal design. Section 2.3 gives geometric interpretation of agent types (s, b) and their demand under an allocation rule y (for input as quality q).

2.1 Generalization of Threshold Mechanisms to Lottery Menus

This section states that deterministic threshold mechanisms are *not optimal* in general (when the distribution over budgets has multiple support). Consequently, we need to generalize the class of mechanisms being considered. This section gives the sufficient extension to *lottery menus* (Definition 6 below).

Insufficiency of deterministic thresholds is stated simply:

▶ **Proposition 4.** *For admission games* $\mathcal{G}$ *in which the set of budgets is multiple, i.e.* $|B| > 1$*, (deterministic) threshold mechanisms are not optimal generally.*

The proof is by counter-example – we give the details and analysis of Example 28 in Appendix A.2 where we conclude that all deterministic threshold mechanisms are dominated by stochastic allocation $x = (1 - \epsilon_L/1 + \epsilon_L - \epsilon_x)$ with "small" ϵ_x for agents exhibiting at least a minimum quality.

Although we must now consider allocation rules y more generally than threshold mechanisms, without loss of generality, we may assume monotonicity of y:

▶ **Lemma 5** (Monotonicity). *For every admission game $\mathcal{G}$, there exists an optimal allocation rule that is weakly monotone increasing.*

Proof. For every allocation rule $\tilde{y}$ that is strictly decreasing somewhere on its domain, the principal gets the same utility from the "ratcheted" allocation rule $\bar{y}(\tilde{y})$ which increases the allocation in every decreasing region of $\tilde{y}$ to be equal to the left end point of the region, i.e., flat on the region. (The resulting $\bar{y}(\tilde{y})$ is weakly monotone increasing.)

Principal utility is the same for $\bar{y}$ and $\tilde{y}$ because every agent $A = (s, b)$ gets the same allocation: all qualities q where $\bar{y}(q) \neq \tilde{y}(q)$ are ignored because they are dominated for both functions by the "ratchet point"-quality (a weakly larger allocation requiring strictly less effort is preferred). ◀

Thus, in order to identify the optimal mechanism, we propose lottery menus:[4]

▶ **Definition 6** (Menu). *A* lottery menu *mechanism is a (weakly) monotone allocation rule y with* menu options $(q, x = y(q))$*, where x is the allocation probability for an agent exhibiting quality q.*

2.2 Leveraging Agent Feasibility to Improve Screening

This section formalizes the feasibility for an agent to choose a given menu option. Subsequently, this section explains how lottery menus effectively leverage feasibility to promote the principal's objective: decreasing allocation necessarily discriminates in favor of higher-skill agents (summarized below as Proposition 9; note, we can already observe this effect working in Example 28).

Feasibility is due to (a) the budget constraint, and (b) a non-negative utility requirement:

▶ **Definition 7.** *Menu option (q, x) is* feasible *for agent $A = (s, b)$ if:*

1. (affordability) *minimal effort $e^* = q/s$ (to achieve quality q) is at most b, i.e., $e^* \leq b$; and*
2. (rationality) *(q, x) induces non-negative utility for A, i.e., $u_A(e^*, x) = x - e^* \geq 0$.*

▶ **Fact 8.** *Menu option (q, x) is feasible for agent $A = (s, b)$ if and only if $q/s \leq \min\{b, x\}$. Upon choosing this option, A achieves utility $u_A(q) = x - q/s$.*

Fact 8 implies that we can use stochastic (partial) allocation to improve the principal's expected utility by discouraging a low-skill agent from applying. Consider two agents described qualitatively as: high-skill-low-budget (A_H) and low-skill-high-budget (A_L), where we naturally prefer to admit A_H. Intuitively, we decrease x for a fixed $\bar{q}$, we get the following effects: (a) for larger b, the upper bound on $\bar{q}/s$ is set by x "sooner" (as it decreases, rather than by budget); and (b) rationality is violated for A_L before it is violated for A_H. Both effects (a) and (b) threaten A_L's utility. We state this formally as a *ceteris paribus* result, where dependence on feasibility is clear in the proof:

[4] If we consider admitting multiple agents drawn independently from $S \times B$ and our utility is (independently) additive across decisions, it may be possible to negatively correlate admission decisions to target the total number of admits. For example, if our setting is discrete and we choose an allocation rule y, if k_q agents apply with the same quality q, we may decide to run a *lottery* which admits exactly $1/y(q)$ of the agents uniformly at random.

▶ **Proposition 9** (The Lotteries-in-Screening Proposition). *For a fixed quality $\bar{q}$, decreasing the allocation $y(\bar{q})$ when an agent exhibits quality $\bar{q}$ increases the lower bound on the skill of agents who feasibly choose $(\bar{q}, y(\bar{q}))$.*

Proof. The agent's utility is the difference between allocation and effort: $y(\bar{q}) - e$. Utility is 0 for a marginally-skilled agent with skill s^* who must put in effort $e^* = y(\bar{q})$ to achieve quality $\bar{q}$. We also have the abstract definition: $q = s \cdot e$. Substituting from the definition, we have $y(\bar{q}) = \bar{q}/s^*$. The quality $\bar{q}$ is fixed, thus decreasing the left-hand side requires increasing the skill threshold s^*. ◀

2.3 Geometric Interpretation

This section introduces geometric interpretations of the problem (that will be useful for our analysis of optimal mechanisms). The first of these visualizations is graphical representation of an agent's feasible allocations. Regions of feasibility map directly onto a graph of an allocation rule y which has quality space as its domain and allocation as its output. As exhibited in Figure 1(Top) which gives two graphic examples of these regions, we have the following geometric observations:

▶ **Fact 10.** *An agent A with skill s (ignoring budget and affordability):*

- *is partially identified by a ray out of the origin with slope $1/s$; this ray necessarily lower-bounds A's feasible region because this is the* zero-utility line*, i.e., points (q, x) on this line result in A achieving utility of 0;*
- *who chooses a menu option (q, x) – independent of being rational or not – will get utility equal to the vertical difference between the chosen allocation x and the height $q \cdot (1/s)$ of the zero-utility line at q (which directly interprets from definitions: $u_A(q) = x - q/s$).*

From the points of Fact 10, agent types partitioned by skill $s_i \in \mathbf{S}$ are identified with their respective zero-utility lines. We illustrate this in Figure 1(Bottom) by expanding its (Top)graphics to show a setting with two skill types: low skill s_L and high skill s_H. Within this context, we give formal definitions:

▶ **Definition 11.** *The* low-skill agents' line *is their zero-utility line with slope $1/s_L$ on the (quality, allocation) graph for (budget-unconstrained) low-skill agents. Similarly, the* high-skill agents' line *is their zero-utilily line with slope $1/s_H$. Generally, we refer to zero-utility lines as* skill lines*.*

3 The Optimal Mechanism for 2-skill, Discrete-budget Types

This section solves the discrete-type setting for a principal with skill threshold τ and a stochastic agent $A = (s, b)$ with type-space defined by two skill-types with $s_L < \tau < s_H$ and n budget-types with $0 < b_1 < b_2 < \ldots < b_n$. Due to the discrete type-space, the optimal mechanism may not be unique. Theorem 13 is sufficient to identify an optimal mechanism, which is a *slanted-stair function*:

▶ **Definition 12.** *A* slanted-stair function $f : \mathbb{R}_+ \to [0, 1]$ *(as an allocation rule) has $f(0) = 0$; and is a weakly increasing function that begins as a sequence of line segments that all have the same (constant), positive derivative. Each line segment has open lower bound and closed upper bound. (The function's output must reach 1 and is identically 1 for larger inputs.)*

We refer to the line segments as slanted-steps*. We refer to the (necessarily positive) vertical gaps between slanted-steps as* jumps*.*

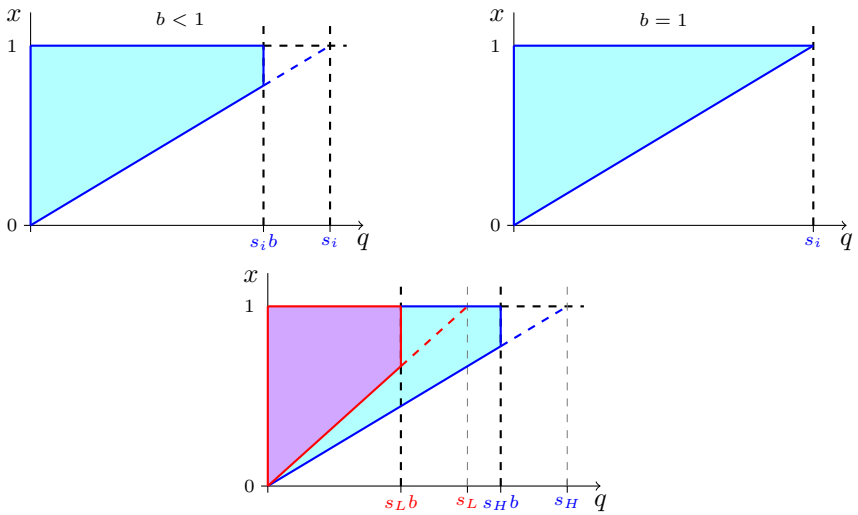

Figure 1 (Top) A menu option is a point (q, x) with coordinates respectively from quality space $\mathbb{R}_+$ and allocation space $[0, 1]$. The blue regions are feasible for agent $A = (s_i, b)$, i.e., A can select these menu options (when they exist) and achieve non-negative utility. The regions' lower-bound line has slope $1/s_i$. (Bottom) The red region is feasible for an agent $A_L = (s_L, b)$. The blue region (which entirely encompasses red) is feasible for agent $A_H = (s_H, b)$. Regarding discussion of Proposition 9 in Section 2.2, observe how for fixed quality set by $q = s_L \cdot b$, it is possible to use decreased allocation awarded to a fixed quality (at/below the vertical boundary between red and blue regions), in order to exclusively admit a high-skill agent.

For a set of types T, let $\Delta(T)$ be the probability simplex over the elements of T. Before giving our main result, we state an interesting observation: there will be nothing in the proof of Theorem 13 that requires the independence of S and B. Thus to state a stronger main result, we define a *correlated* admission game by $\mathcal{H} = (S, \mathcal{B}^S, \tau)$ where $\mathcal{B}^S$ is a set of conditional budget-distributions: one budget-distribution corresponding to each skill-type with positive support in S.[5]

Agents in a correlated game have the same description as in the original, independent game. By contrast, the principal's objective must be updated to reflect the correlation:

$$\max_y u_P(\mathcal{H}, y) := \max_y \mathbf{E}_{A \sim (S,\ \mathcal{B}^S)} [x(y, s) \cdot (s - \tau)] \tag{6}$$

▶ **Theorem 13** (Main Result). *Given a correlated admission game* $\mathcal{H} = (S = \Delta(\{s_L, s_H\}), \mathcal{B}^S = \{B_{s_L} = \Delta_{s_L}(\{b_1, \ldots, b_n\}), B_{s_H} = \Delta_{s_H}(\{b_1, \ldots, b_n\})\}, \tau)$ *with* $0 < s_L < \tau < s_H$ *and* $0 < b_1 < b_2 < \ldots < b_n$. *There exists an optimal mechanism* y^* *for the correlated admission game* $\mathcal{H}$ *that is a slanted-stair function* f *with constant slope equal to* $1/s_L$ *and at most one jump, and with:*

1. *the region of the first slanted-step characterized by: equality to the* low-skill agents' zero-utility line*;*
2. *the quality-index at which* f *jumps* q^{jump} *(if it exists) characterized by: occurring either at quality* $q_0 = 0$*, or occurring at some maximum-possible quality exhibited by some low-skilled agent* $A_{L,i} = (s_L, b_i)$*, i.e., at some* $q^{\text{jump}} = q^{\dagger}_{L,i} = s_L \cdot b_i$*;*

[5] Assuming discrete budget-distributions, while elements of $\mathcal{B}^S$ may have distinct support, it is without loss of generality to assume that they all have common, enumerated support $b_1, \ldots, b_n$ because any locally-unused budget type b_i can be locally assigned probability 0.

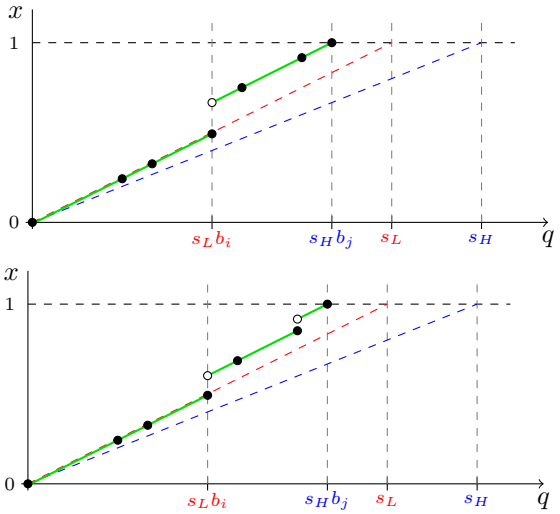

Figure 2 (Top) A one-jump, slanted-stair allocation curve y (solid green) with $q^{\text{jump}} = s_L \cdot b_i$ and $q^{x=1} = s_H \cdot b_j$. The black dots are an example of discrete menu options. Recall that agent utility is interpretable as the vertical difference between allocation and (zero-utility) skill line. Any low-skilled agent $A_L = (s_L, b_k)$ with $q^\dagger_{L,k} = s_L \cdot b_k \leq s_L \cdot b_i$ will choose menu option $(0, 0)$ (per the tie-breaking rule, see Definition 18). Any low-skilled agent with $q^\dagger_{L,k} = s_L \cdot b_k > s_L \cdot b_i$ will choose $(s_L b_i + \epsilon, y(s_L b_i + \epsilon))$ with $\epsilon \to 0$. Each high-skilled agent $A_H = (s_H, b_k)$ with $k < j$ will achieve maximum quality $q^\dagger_{H,k} = s_H \cdot b_k < s_H b_j$; and those with $k \geq j$ will achieve quality $s_H b_j$ (and are allocated with probability 1). (Bottom) A two-jump, slanted-stair allocation curve y (solid green).

3. *the region of the second slanted-step (if it exists) characterized by: the quality at which f intersects the allocation-of-1 horizontal line is the maximum-possible quality exhibited by some high-skilled agent $A_{H,j} = (s_H, b_j)$, i.e., at some $q^{x=1} = q^\dagger_{H,j} = s_H \cdot b_j$.*

(Note, optimal assignment of mechanism parameters and the given characterizations of Theorem 13 are sufficient to identify the height of the vertical jump, starting from the low-skill agents' line.)

The proof of Theorem 13 depends on a sequence of lemmas which we state at the end of this section. The proofs of Theorem 13 and its supporting lemmas appear in the main version of the paper. Graphically, the optimal menu (which may be discrete, corresponding to our discrete setting) will qualitatively have the single-jump structure of Figure 2(Top) with menu options on only two line segments (as two slanted-steps). Multi-jump structures are precluded, such as the three-slanted-steps in Figure 2(Bottom).

The statement of Theorem 13 induces the following corollary regarding the polynomial running time of a brute-force algorithm that searches over the possible combinations of jump-points and jump-heights, which is sufficient to find the optimal algorithm of the statement's setting.

▶ **Corollary 14** (Running Time). *Given a correlated admission game $\mathcal{H} = (S = \Delta(\{s_L, s_H\}), \mathcal{B}^S = \{B_{s_L} = \Delta_{s_L}(\{b_1, \ldots, b_n\}), B_{s_H} = \Delta_{s_H}(\{b_1, \ldots, b_n\})\}, \tau)$ with $0 < s_L < \tau < s_H$ and $0 < b_1 < b_2 < \ldots < b_n$ and – per Theorem 13 – the sufficient, discrete search space for an optimal algorithm.*

The optimal mechanism may be identified by a brute-force search over the $O(n^2)$ unknown combinations of parameters of the optimal characterization (Theorem 13). The time to evaluate each allocation rule (resulting from a combination of parameters) also runs in polynomial time.

3.1 Discussion of Optimal Characterization in Theorem 13

Having a characterization of optimal mechanisms, we would like to understand qualitatively their performance. We will discuss two dimensions of efficacy: (1) mechanism performance, of course, as the originally-defined objective; and (2) *fairness*, which informally is a measurement of how well outcomes-per-agent-type conform to some definition of what outcomes the agents *arguably deserve*, specifically compared to other agents' type-outcome pairs.

Regarding mechanism performance, we know that the single-jump, slanted-stair characterization of Theorem 13 improves on the (deterministic) threshold mechanisms of Definition 1 which are not generally optimal (by Proposition 4), except for games with convenient distributions S and B (e.g., Proposition 3). On the other hand, optimal mechanism performance still falls short of the *offline optimal* benchmark which has full information and which is generally unachievable; rather, we may use it as a reference mechanism to which we compare performance:

▶ **Definition 15.** *Given a stochastic agent $A = (s, b)$, the* offline optimal mechanism *for a principal requiring skill-threshold τ – which is assumed to know the realized skill type of the agent as $\hat{s} \sim S$ – admits the agent if $\hat{s} > \tau$ and only if $\hat{s} \geq \tau$; and this admission decision is independent of the agent's realized budget type $\hat{b} \sim B$.*

The offline optimal mechanism is unconditionally optimal, as it fully allocates every agent with skill above the threshold and fully rejects every agent with skill below it. In order to increase the performance of mechanisms beyond what is possible from Theorem 13 – i.e., from standard mechanism design subject to agents' incentive compatibility constraints – in Section 4 we consider a modified admission problem in which the agent may have exogenous access to a *subsidy*.

Regarding fairness, we first must consider the philosophical concept of what comparisons between distinct agents' type-outcome pairs may arise as fair or as unfair within the parameters of our model (Section 2). Loosely summarizing: our agents independently have higher or lower skills and higher or lower budgets; and by best-responding to a given allocation rule based on skill and budget, agents are consequently admitted with larger or smaller probability. Reasonably, agent "skill" is positively correlated with student value and agent budget is independent, so we posit that higher skill types are *more-deserving* of being admitted than lower types, independent of budget. Moreover, the degree of worthiness should increase with increasing *cardinal* difference in skill types.

Thus, we consider the following concept of fairness: regardless of budget, larger (admission) allocations given to lower-skilled agents are comparatively judged to be unfair outcomes as the higher-skilled type is more-deserving; and the larger the skill-difference, the larger the unfairness. Furthermore the strict contrapositive also holds: comparatively larger allocations given to higher-skilled agents are more fair. However, the choice of function used to measure technically the unfairness of an allocation rule remains debatable.

From the following intuition, the mechanism design problem of our admission-game model should be positively aligned with objectives resulting from our concept of fairness. First, recall the principal's utility from admitting an agent A, which is $u_P(A \mid \text{admitted}) = s - \tau$. Given this utility function, the principal has a precise, cardinal utility measure over admitting agent skill-types, which has both order and cardinality aligned with fairness as desired, regardless of the technical fairness measure. I.e., the principal is incentivized to choose an allocation rule that increases fairness. In at least one sense, this is strictly true, which moreover motivates the principal's objective function itself (see equation (6)) as a formal example of fairness measure:

▶ **Fact 16.** *Where incentive compatibility permits, the principal is incentivized to inherently prefer that between two agent types with different skill levels, the agent type with higher skill will receive the larger allocation probability.*

▶ **Corollary 17.** *Given a correlated admission game* $\mathcal{H} = (S = \Delta(\{s_L, s_H\}), \mathcal{B}^S = \{B_{s_L}, B_{s_H}\}, \tau)$ *with* $0 < s_L < \tau < s_H$ *and* $0 < b_1 < b_2 < \ldots < b_n$. *For the fair mechanism design problem which maximizes the fairness measure set equal to the principal's utility function, the optimal mechanism and characterization of optimal mechanisms are determined identically to Theorem 13.*

Second, the offline optimal mechanism can illustrate the alignment between the principal's mechanism design incentives and fairness. On one hand, offline optimal represents perfect – albeit generally unachievable – performance for the mechanism. On the other hand, by giving allocation 1 to an upward-closed set of skill-types above τ, allocation 0 to a downward-closed set of skill-types below τ, and any constant allocation to skill-type exactly τ, the allocation is arguably fair because no rejected skill-type can protest for increased allocation on the basis that it is strictly more-deserving than any admitted skill-type. Thus, the offline optimal mechanism as ideal-objective further aligns the principal and fairness.

For purposes of space, we defer discussion of a third intuitive perspective supporting the alignment of optimal mechanisms and fairness to Appendix A.3.

3.2 The Proof of Theorem 13

We need one more critical detail to set up the proof of Theorem 13. Depending on allocation rule y, an agent A may be indifferent between a set of quality-allocation menu options that are optimal for A. To address this, we define our tie-breaking rule:

▶ **Definition 18.** *When an agent's set of optimal menu options is multiple, the* tie-breaking rule *is: all agents choose* the smallest menu option of the set*. (Note, "smallest" is the same in either dimension of quality or allocation.)*

This tie-breaking rule is material for our results: it is sufficient to break ties optimally in favor of the principal's objective.[6] Recalling that utility is equal to the vertical difference between the allocation and the height of the zero-utility line (Fact 10), the key effect of tie-breaking is observed in Figure 2: within a region of a single slanted-step, *low-skill agents are indifferent everywhere and choose the minimal allocation at the left endpoint of the region.* This tie-breaking rule applies for all result statements and proofs in this paper.

As an overview, the proof of Theorem 13 proceeds as a search for the optimal mechanism. This search is organized as a sequence of reductions of the search space: it starts with an allocation rule that is monotone (Lemma 5 on page 6) but is otherwise arbitrary; and then with each successive lemma, we prove that it is sufficient to restrict attention to a smaller set of allocation rules. Lemma 23 is the last reduction in the sequence and states that the optimal mechanism must be a slanted-stair function (Definition 12) with at most one jump. The final proof of Theorem 13 starts from the statement of Lemma 23 and proves the additional details in its own statement.

[6] This tie-breaking rule is justified similarly to tie-breaking in other areas of mechanism design, e.g., in auctions with a revenue objective in which agents with value equal to price are assumed to buy, in favor of the designer's objective. Intuitively, the justification is that small perturbations to the design can achieve the same outcome within arbitrary (lossy) required precision; so instead, we simplify the analysis by allowing ties and breaking them favorably, rather than accounting for a notation-heavy perturbation.

All of the following lemmas assume the same setting as the statement of Theorem 13, which is: given a correlated admission game $\mathcal{H} = (S = \Delta(\{s_L, s_H\}), \mathcal{B}^S = \{B_{s_L} = \Delta_{s_L}(\{b_1, \ldots, b_n\}), B_{s_H} = \Delta_{s_H}(\{b_1, \ldots, b_n\})\}, \tau)$ with $0 < s_L < \tau < s_H$ and $0 < b_1 < b_2 < \ldots < b_n$.

With this overview in place, the sequence of reductions of the search space is:

▶ **Lemma 19** (Lower bound). *An optimal allocation rule is never under the low-skill agents' line.*

▶ **Lemma 20** (Strong monotonicity). *There exists an optimal allocation rule y^* that everywhere has a derivative lower bound set by $1/s_L$ (the slope of the low-skill agents' line).*

▶ **Lemma 21** (Constant allocation slope). *There exists an optimal allocation rule y^* that is a slanted-stair function, i.e., it everywhere has constant derivative equal to $1/s_L$ (the slope of a low-skill agents' line), allowing for arbitrary, discretely-indexed, positive, vertical jumps.*

▶ **Lemma 22** (A corner-case exclusion). *There exists an optimal allocation rule y^* for which the optimal menu option of the agent type with smallest maximum-quality gives 0-allocation. (This agent is $A = (s_L, b_1)$ with $q^\dagger_{L,1} = s_L \cdot b_1$.)*

▶ **Lemma 23** (Sufficiency of at-most one jump). *There exists an optimal allocation rule y^* that is a slanted-stair function with at most one jump; furthermore, if there is a jump in a given y^*, then its allocation in the region of the first slanted-step must be equal to the low-skilled agents' line.*

The proofs for each lemma in this sequence appear in the full version of the paper.

4 Mechanisms for Agents with Subsidized Effort

This section considers agent subsidies directly in *effort-space*. A budget on effort implies a time-constraint. Effort-subsidies are an intervention that increases the agent's effort-budget by freeing up an agent's time spent on other obligatory activities. Technically, we consider subsidies as uniform, additive increases to agents' budget constraints. These subsidies are offered *unconditionally*: agents may spend the time on an outside-option (leisure) activity; or they may invest the time in effort, which they experience as *costly* (i.e., as the opportunity cost of the forfeited leisure time). E.g., subsidies may be provided by performing time-costly tasks for agents' benefit (like uniformly offering free postal pickup/delivery) – freed from the burden of the task, agents enjoy leisure or spend their time exerting effort in our model.

The main goal of this section is to solve for the characterization of the optimal mechanism of the (modified) admission game which has expanded setting parameters that make it possible to consider a combined-question of screening and *design of unconditional subsidies*. Corollary 24 states that its characterization is the same as Theorem 13. We also show that this subsidies setting can only help the principal's objective (in Proposition 26).

4.1 The Setting with Subsidies

We add the following elements to the correlated setting of Theorem 13 (based on Section 2).

The mechanism designer may a priori offer to the agent $A = (s, b)$ an *effort-budget subsidy* d from a non-negative range, i.e., the subsidy is $d \in [D_-, D_+]$. The agent accepts the whole subsidy unconditionally and the agent's new budget is $b + d$.

It is not possible to restrict access to the subsidy to sub-classes of agent-types: not to high-skill agents and not to disadvantaged agents. The constant subsidy amount is necessarily available to each type indiscriminately because the realizations of an agent's skill/budget types are unknown at the time of the offer, i.e., at the time of subsidized-mechanism design. While we can not use uniform subsidies to discriminate directly, we will be able to improve the *principal's objective* using the following observation: given an optimal single-jump, slanted-stair allocation (as characterized by Theorem 13), note that the budget constraint binds for *all* high-skill agents receiving allocation less than 1 and they would benefit from relaxing the budget constraint; but for almost all low-skill agents, the budget constraint is not binding because their utility is constant on each slanted-step. This first-order-condition analysis suggests that high-skill agents will voluntarily convert unconditional subsidies to effort and increased allocation, whereas low-skill agents will not.

The subsidy (to increase effort-budget) is exogenous as if enacted and paid by an unrelated third party at no cost to the mechanism. E.g., in an admission problem, the school may be a city's unique, public, magnet high school. The subsidy may be paid uniformly to each eligible applicant by a citywide scholarship program which is separate from the school's admissions office but which has the money to provide the subsidy (up to D_+ per student) and *must support* a citywide goal of maximizing utility from specifically the magnet school's admissions policies. In this case, the magnet school admissions office (as our model's principal) optimizes $d \in [D_-, D_+]$ and the scholarship program must approve it.

For this Section 4, the updated correlated admission game with subsidies is given by $\mathcal{D} = (S, \mathcal{B}^S, \tau, D_-, D_+)$. For a given subsidy $d > 0$, agent $A = (s, b)$ has maximum quality $q^\ddagger = s \cdot (b + d)$, which is larger than the maximum quality without the subsidy ($q^\dagger = s \cdot b$). The agent's updated optimal utility function v_A^* and updated optimal allocation rule w in skill space – subject to allocation rule y – are:

$$v_A^*(y, s, d) := \max_{e \in [0, b+d]} y(s \cdot e) - e = \max_{q \in [0, q^\ddagger]} y(q) - q/s \tag{7}$$

$$x = w(y, s, d) := y(s \cdot \left\lceil \operatorname*{argmax}_{e \in [0, b+d]} y(s \cdot e) - e \right\rceil) \tag{8}$$

In equation (8), note that because the subsidy is unconditional, the agent pays the full cost of effort e, including the (opportunity) cost of effort above the original budget b.

For a given agent A, the principal's utility from admitting A remains the function $u_P(A \mid \text{admitted}) = s - \tau$. Thus, the principal's updated mechanism design problem is to maximize $v_P(\mathcal{D}, y, d)$ which is the expected utility from an admitted agent's skill versus the principal's threshold τ, weighted by allocation probability according to w (which accounts for the subsidy):

$$\max_{y,\ d \in [D_-, D_+]} v_P(\mathcal{D}, y, d) := \max_y \mathbf{E}_{A \sim (S,\ \mathcal{B}^S)} \left[w(y, s, d) \cdot (s - \tau) \right] \tag{9}$$

4.2 Results with Subsidies

The main result of this section is: the optimal mechanism when agents have access to unconditional subsidies has the same characterization as the original game, as described in Theorem 13. Moreover, we are immediately ready to state and prove it as a corollary:

▶ **Corollary 24.** *Given a correlated admission game with subsidies* $\mathcal{D} = (S = \Delta(\{s_L, s_H\}), \mathcal{B}^S = \{B_{s_L} = \Delta_{s_L}(\{b_1, \ldots, b_n\}), B_{s_H} = \Delta_{s_H}(\{b_1, \ldots, b_n\})\}, \tau, D_-, D_+)$ *with* $0 < s_L < \tau < s_H$ *and* $0 < b_1 < b_2 < \ldots < b_n$. *The structure of the optimal mechanism has the same characterization as the standard game, as given in Theorem 13.*

Proof. As part of identifying the optimal mechanism – according to equation (9) – the designer selects an optimal assignment to the subsidy variable $d \in [D_-, D_+]$.

Consider an optimal assignment d^* (any element of the argmax is fine). The optimal mechanism associated with d^* must be the same as the optimal mechanism for an alternative game $\mathcal{D}'$ which sets parameters $S, \mathcal{B}^S, \tau$ to be the same as $\mathcal{D}$, but which assigns the endpoints of allowable subsidies to both be d^*, i.e., $\mathcal{D}'$ has $D_- = D_+ = d^*$.

This corollary then follows directly from Lemma 25(2) below. ◀

While Corollary 24 is sufficient to give us characterization, it does not give us an algorithm to find the optimal mechanism because it uses theoretical existence of the optimal subsidy d^* without identifying it.

The following observations regarding correlated admission games with subsidies are straightforward. Omitted proofs in this section appear in the full version of the paper.

▶ **Lemma 25.** *A correlated admission game with subsidies is* $\mathcal{D} = (S = \Delta(\{s_L, s_H\}), \mathcal{B}^S = \{B_{s_L} = \Delta_{s_L}(\{b_1, \ldots, b_n\}), B_{s_H} = \Delta_{s_H}(\{b_1, \ldots, b_n\})\}, \tau, D_-, D_+)$. *Consider arbitrary* $\mathcal{D}$*, i.e., consider its inputs as variables.*

1. *Without loss of generality, we may reduce* $\mathcal{D}$ *to a correlated game* $\mathcal{D}'$ *which has* $D'_- = 0$.
2. *A game* $\mathcal{D}$ *fixing an exact subsidy by setting* $D_- = D_+$ *is equivalently described by a game* $\mathcal{H}_{\mathcal{D}}$ *and thus is characterized by the statement of Theorem 13.*
3. *If* $D_- = 0$*, then expanding the original correlated admission game* $\mathcal{H} = (S, \mathcal{B}^S, \tau)$ *to consider admissions with subsidies – formulated as the updated game* $\mathcal{D}$ *– can only increase the utility of the principal.*
4. *Given* $S, \mathcal{B}^S, \tau$*, there exists a minimal subsidy upper bound* D_+^m *such that for all* $D_+ \geq D_+^m$*, the optimal mechanism achieves the offline optimal performance (see Definition 15), i.e., it is able to perfectly discriminate between high-skill and low-skill agent types regardless of their budgets.*

Lemma 25(3) is fairly obvious: if $D_- = 0$, then the principal has the option of "free disposal" of the subsidy-variable and can do no worse than the game without subsidies. The more interesting statement is that the principal's objective can only improve for $D_- > 0$ generally:

▶ **Proposition 26.** *For arbitrary* $D_- \geq 0$*, expanding the original correlated admission game* $\mathcal{H} = (S, \mathcal{B}^S, \tau)$ *to consider admissions with subsidies can only increase the utility of the principal.*

In the proof of Proposition 26, we consider specifically the subsidy $d = D_- > 0$ and (deterministically) transform the optimal allocation rule without subsidies into a new allocation rule with weakly larger performance given the uniform, unconditional agent's budget-subsidy D_-. In particular in comparison to the optimal allocation without subsidies, the new allocation gives all low-skill agent-types weakly smaller allocation, and gives all high-skill agent-types weakly larger allocation.

This new allocation rule is not necessarily optimal for its (subsidized) setting, but by dominating the original setting, its existence proves that the principal's objective can only improve. On the other hand, the new allocation rule may harm the agents' utilities (for any agent type, except low-skill-low-budget agents who already get 0-allocation before subsidies and who still get 0). While this assessment is not a final judgment (because the new allocation is not necessarily optimal), it is consistent with observations in [15] which showed that subsidies for disadvantaged agents might harm their utilities.

References

1. Saba Ahmadi, Hedyeh Beyhaghi, Avrim Blum, and Keziah Naggita. The strategic perceptron. In Péter Biró, Shuchi Chawla, and Federico Echenique, editors, *EC '21: The 22nd ACM Conference on Economics and Computation, Budapest, Hungary, July 18-23, 2021*, pages 6–25. ACM, 2021. doi:10.1145/3465456.3467629.
2. Saba Ahmadi, Hedyeh Beyhaghi, Avrim Blum, and Keziah Naggita. On classification of strategic agents who can both game and improve. In L. Elisa Celis, editor, *3rd Symposium on Foundations of Responsible Computing, FORC 2022, June 6-8, 2022, Cambridge, MA, USA*, volume 218 of *LIPIcs*, pages 3:1–3:22. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022. doi:10.4230/LIPIcs.FORC.2022.3.
3. Tal Alon, Magdalen Dobson, Ariel Procaccia, Inbal Talgam-Cohen, and Jamie Tucker-Foltz. Multiagent evaluation mechanisms. *In Proceedings of the AAAI Conference on Artificial Intelligence*, 34(02):1774–1781, April 2020. doi:10.1609/aaai.v34i02.5543.
4. Yahav Bechavod, Katrina Ligett, Zhiwei Steven Wu, and Juba Ziani. Causal feature discovery through strategic modification. *ArXiv*, abs/2002.07024, 2020. URL: https://arxiv.org/abs/2002.07024.
5. Mark Braverman and Sumegha Garg. The role of randomness and noise in strategic classification. In Aaron Roth, editor, *1st Symposium on Foundations of Responsible Computing, FORC 2020, June 1-3, 2020, Harvard University, Cambridge, MA, USA (virtual conference)*, volume 156 of *LIPIcs*, pages 9:1–9:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPIcs.FORC.2020.9.
6. Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 547–555, New York, NY, USA, 2011. Association for Computing Machinery. doi:10.1145/2020408.2020495.
7. Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018. arXiv:1808.00023.
8. Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, EC '18, pages 55–70, New York, NY, USA, 2018. Association for Computing Machinery. doi:10.1145/3219166.3219193.
9. Yiding Feng, Jason D. Hartline, and Yingkai Li. Simple mechanisms for non-linear agents. In Nikhil Bansal and Viswanath Nagarajan, editors, *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023, Florence, Italy, January 22-25, 2023*, pages 3802–3816. SIAM, 2023. doi:10.1137/1.9781611977554.ch148.
10. Alex M. Frankel and Navin Kartik. Improving information from manipulable data. *arXiv: Theoretical Economics*, June 2019. doi:10.1093/jeea/jvab017.
11. Jason Gaitonde, Yingkai Li, Bar Light, Brendan Lucier, and Aleksandrs Slivkins. Budget pacing in repeated auctions: Regret and efficiency without convergence. In Yael Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference, ITCS 2023, January 10-13, 2023, MIT, Cambridge, Massachusetts, USA*, volume 251 of *LIPIcs*, pages 52:1–52:1. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023. doi:10.4230/LIPIcs.ITCS.2023.52.
12. Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Z. Wang. Maximizing welfare with incentive-aware evaluation mechanisms. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 160–166. International Joint Conferences on Artificial Intelligence Organization, July 2020. Main track. doi:10.24963/ijcai.2020/23.
13. Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, pages 111–122, New York, NY, USA, 2016. Association for Computing Machinery. doi:10.1145/2840728.2840730.

14 Keegan Harris, Hoda Heidari, and Zhiwei Steven Wu. Stateful strategic regression. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 28728–28741, 2021. URL: https://proceedings.neurips.cc/paper/2021/hash/f1404c2624fa7f2507ba04fd9dfc5fb1-Abstract.html.

15 Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In danah boyd and Jamie H. Morgenstern, editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 259–268. ACM, 2019. doi:10.1145/3287560.3287597.

16 Christopher Jung, Sampath Kannan, Changhwa Lee, Mallesh M. Pai, Aaron Roth, and Rakesh Vohra. Fair prediction with endogenous behavior. In Péter Biró, Jason D. Hartline, Michael Ostrovsky, and Ariel D. Procaccia, editors, *EC '20: The 21st ACM Conference on Economics and Computation, Virtual Event, Hungary, July 13-17, 2020*, pages 677–678. ACM, 2020. doi:10.1145/3391403.3399473.

17 Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. *AEA Papers and Proceedings*, 108:22–27, May 2018. doi:10.1257/pandp.20181018.

18 Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, pages 825–844, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3328526.3329584.

19 Jon M. Kleinberg. Inherent trade-offs in algorithmic fairness. In Konstantinos Psounis, Aditya Akella, and Adam Wierman, editors, *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS 2018, Irvine, CA, USA, June 18-22, 2018*, page 40. ACM, 2018. doi:10.1145/3219617.3219634.

20 Jean-Jacques Laffont and Jacques Robert. Optimal auction with financially constrained buyers. *Economics Letters*, 52(2):181–186, 1996. doi:10.1016/S0165-1765(96)00849-X.

21 Eric S. Maskin. Auctions, development, and privatization: Efficient auctions with liquidity-constrained buyers. *European Economic Review*, 44(4):667–681, 2000. doi:10.1016/S0014-2921(00)00057-X.

22 John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6917–6926. PMLR, 2020. URL: http://proceedings.mlr.press/v119/miller20b.html.

23 Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 230–239, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3287560.3287576.

24 Mallesh M. Pai and Rakesh Vohra. Optimal auctions with financially constrained buyers. *J. Econ. Theory*, 150:383–425, 2014. doi:10.1016/j.jet.2013.09.015.

25 Ashesh Rambachan, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. An economic perspective on algorithmic fairness. *AEA Papers and Proceedings*, 110:91–95, May 2020. doi:10.1257/pandp.20201036.

26 Joseph Stiglitz and Andrew Weiss. Credit rationing in markets with imperfect information. *American Economic Review*, 71(3):393–410, 1981. URL: https://EconPapers.repec.org/RePEc:aea:aecrev:v:71:y:1981:i:3:p:393-410.

27 Joseph E Stiglitz. The Theory of "Screening," Education, and the Distribution of Income. *American Economic Review*, 65(3):283–300, June 1975. URL: https://ideas.repec.org/a/aea/aecrev/v65y1975i3p283-300.html.

28 Shenke Xiao, Zihe Wang, Mengjing Chen, Pingzhong Tang, and Xiwang Yang. Optimal common contract with heterogeneous agents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7309–7316, April 2020. doi:10.1609/aaai.v34i05.6224.

A Deferred Proofs of Propositions and Lemmas

A.1 Proof that a Threshold Mechanism is Optimal for Single-budget

▶ **Proposition 3.** *Assume that an agent has constant budget $\bar{b}$ on effort, i.e., the distribution B is a singular point mass. The threshold mechanism $y^{q'}$ with $q' = \tau \cdot \bar{b}$ is optimal.*

Proof. The optimality of $y^{q'}$ in fact follows from the stronger statement in Lemma 27 (below). ◀

The offline optimal mechanism (Definition 15) is generally unachievable. Despite that caveat, it is possible to achieve the offline optimal mechanism for the special case of singular budgets, as subsequently stated in Lemma 27. Recall the intuition given in the main body of the paper: "Proposition 3 holds because single-budget is a simple setting in which quality-thresholds directly implement skill-thresholds, in particular for the principal's threshold τ."

▶ **Lemma 27.** *Assume that an agent has constant budget $\bar{b}$ on effort, i.e., the distribution B is a singleton point mass.* Without directly observing the realization of the agent's skill $\hat{s} \sim S$, *the threshold mechanism $y^{q'}$ with $q' = \tau \cdot \bar{b}$ is* offline optimal.

Proof. We will show that $y^{q'}$ is offline optimal by showing that it gives allocation 1 to every (randomized) agent skill-type which gives positive utility to the principal, and gives allocation 0 to every agent skill-type which gives negative utility to the principal, thus pointwise-maximizing the principal's utility function.

For agent $A = (s, \bar{b})$, the minimum effort required to reach threshold q' is $e' = {}^{q'}/_{s} = ({}^{\tau}/_{s})\,\bar{b}$. Then $e' \leq \bar{b}$ is affordable (and rational) for A if and only if ${}^{\tau}/_{s} \leq 1$, an inequality which itself is true if and only the principal's utility $s - \tau \geq 0$ (from admitting A; see page 5). By setting the quality threshold to be the maximum achievable by the skill level τ (which corresponds to 0-utility for skill-type τ), the mechanism allocates to exactly the upward closed set of all agent types from which it receives positive utility (Fact 2), and no others. ◀

A.2 Example of Insufficiency of Deterministic Mechanisms

The following Example 28 provides the proof-by-counterexample for Proposition 4.

▶ **Example 28.** Admission game admission game $\mathcal{G} = (S, B, \tau)$ is defined as follows.

Agent A has discrete skill space with two types (i.e, $|S| = 2$) with low skill $s_L = 1 + \epsilon_L$ (for $\epsilon_L \to 0$) and high skill $s_H = 2$. Agent A has discrete budget space with two types ($|B| = 2$) with low budget $b_L = {}^1/_2$ and high budget $b_H = 1$. The distributions S and B have positive mass on each element of their respective supports but otherwise we leave them indeterminate. The principal P's skill threshold to measure utility is $\tau = {}^3/_2$.

The following analysis will show that for the setting of Example 28, all deterministic threshold mechanisms are dominated by stochastic allocation $x = ({}^{1-\epsilon_L}/_{1+\epsilon_L-\epsilon_x})$ for agents exhibiting quality at least 1.

As a starting point, consider the deterministic threshold mechanism $y^{q'}$ which picks $q' = 1$. The following gives initial analysis of an agent with high skill type s_H:

- minimum effort to achieve q' is: $e_H = {}^1/_2$;
- utility from achieving threshold q' is: $1 - {}^1/_2 = {}^1/_2$;
- an agent of type (s_H, b_H) will put in effort to be admitted (given $q' = 1$ and furthermore, whenever $q' < 2$);

- an agent of type (s_H, b_L) will also put in effort to be admitted, but critically, can not put in effort to be admitted if q' is increased above 1 by any $\epsilon_q > 0$ because this agent-type (s_H, b_L) is bounded by maximum quality $q^\dagger = s_H \cdot b_L = 2 \cdot 1/2 = 1$.

Alternatively, the following gives initial analysis of an agent with low skill type s_L:

- minimum effort to achieve q' is: $e_L = 1/1 + \epsilon_L$;
- utility from achieving threshold q' is: $1 - 1/1 + \epsilon_L = \epsilon_L/1 + \epsilon_L$;
- an agent of type (s_L, b_H) will put in effort to be admitted (given $q' = 1$);
- an agent of type (s_L, b_L) will put in 0 effort (because maximum quality is less than q').

Offline-optimal (Definition 15) allocates all agents with skill $(s_H, \cdot)$ and rejects all agents $(s_L, \cdot)$. The current quality threshold under consideration $q' = 1$ is the largest threshold that will admit types (s_H, b_L). Let $\pi_{a,T}$ be the probability corresponding to arbitrary agent type-attribute $a \in \{s, b\}$ and tier $T \in \{L, H\}$. The performance of every threshold mechanism fails to approach the performance of offline optimal (we write "$\gg$" to indicate that the gap is bounded away from 0):

- thresholds $q'_+ > q' = 1$ will not admit types (s_H, b_L) and thus will additively underperform offline optimal by at least:

$$\pi_{s,H} \cdot \pi_{b,L} \cdot (s_H - \tau) = \pi_{s,H} \cdot \pi_{b,L} \cdot (1/2) \gg 0$$

- thresholds $q'_- \leq q' = 1$ will admit types (s_L, b_H) and thus will additively underperform offline optimal by at least:

$$\pi_{s,L} \cdot \pi_{b,H} \cdot (\tau - s_L) = \pi_{s,L} \cdot \pi_{b,H} \cdot (1/2 - \epsilon_L) \gg 0$$

However, if we maintain $q' = 1$ and rather *decrease the probability of allocation* from 1 to $(1 - \epsilon_L/1 + \epsilon_L - \epsilon_x)$ for $\epsilon_x \to 0$, then all high types still strictly put in effort and will be admitted (with near-certainty), but the low types now strictly prefer to put in 0 effort.

Formally, for (single-menu-option) allocation $x = y(1) = (1 - \epsilon_L/1 + \epsilon_L - \epsilon_x)$, the utility calculations are (assuming minimum effort to be admitted, ignoring affordability due to budget):

- agents with high skill type s_H have utility: $u_{(H,\cdot)} = y(1) - e_H = 1/2 - \epsilon_L/1 + \epsilon_L - \epsilon_x > 0$;
- agents with low skill type s_L have utility: $u_{(L,\cdot)} = y(1) - e_L = -\epsilon_x < 0$.

Considering, $\epsilon_L \to 0$ and $\epsilon_x \to 0$, the admission-rate of high-skill agents approaches 1 and thus the expected performance of this mechanism becomes arbitrarily close to the performance of offline optimal. Therefore, it strictly improves on the best of any deterministic threshold mechanism (which can't approach performance of offline optimal by the analysis above).

This completes the counterexample to illustrate that deterministic mechanisms are not sufficient.

A.3 A Comparison of Slanted-Stair Allocation to Deterministic Threshold

Section 3.1 gives discussion of the optimal characterization of mechanisms in Theorem 13. For purposes of space, we complete here the discussion of alignment between optimal mechanisms and *fairness*. To summarize the initial discussion in the main body, this alignment first is observed intuitively from the structure of the principal's utility from admitting an agent A, which is $u_P(A \mid \text{admitted}) = s - \tau$. Second, the offline optimal allocation is the "perfect" mechanism performance and is also arguably an ideal allocation in terms of fairness. We now give an additional intuitive perspective supporting this alignment.

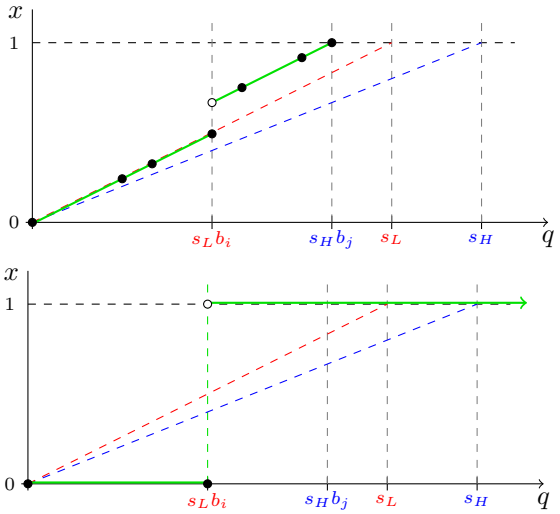

Figure 3 (Top) Illustration of a one-jump, slanted-stair allocation curve y^* (solid green), which is assumed to be optimal for its game parameters (for analysis purposes). The black dots are an example of discrete menu options. The single jump occurs at $q^{\text{jump}} = s_L b_i$. Regarding discussion in Appendix A.3: the first, left-most region is "below the jump;" the second, middle region is "above the jump but not fully allocated;" and the third, right-most region is "full allocation." (Bottom) The (strictly-greater-than) threshold mechanism with threshold set equal to the jump-point in (Top), i.e., with $q' = q^{\text{jump}} = s_L b_i$.

Third – analyzing qualitatively for both mechanism performance and fairness – we can make a comparison between (a) an optimal single-jump-at-q^{jump}, slanted-stair allocation rule y^* of Theorem 13; and (b) the specific – albeit modified – threshold mechanism that jumps from allocation 0 to 1 at the same quality q^{jump}. For convenience, we copy Figure 2(Top) into Figure 3.

The modification is that the threshold mechanism in this section will require for admission that an agent's exhibited quality be *strictly greater* than the threshold. This organizes the closed-versus-open endpoints of the threshold-step in a way that allows for a more-direct comparison to slanted-stair functions. This is illustrated in Figure 3(Bottom).

Graphically, the optimal mechanism y^* (which may be a discrete menu, corresponding to our discrete setting) will qualitatively have the single-jump structure of Figure 3(Top). Using agent skill/budget-indexing of Figure 3 (i.e., notation), the general structure of y^* has three regions:

1. the left-most region is "below the jump" defined by qualities $q \in [0, q^{\text{jump}} = s_L b_j]$;
2. the middle region is "above the jump but not full allocation" defined by qualities $q \in (q^{\text{jump}} = s_L b_j, s_H b_L)$;
3. the right-most region is "full allocation" defined by the quality $q = s_H b_L$ (and all larger qualities, though rational agents never choose these larger levels, which require exerting superfluous effort to achieve, without an increase in allocation).

The allocation rule y^* is optimal for the standard principal-objective, so it obviously dominates the threshold mechanism with its quality-space threshold set to be $q' = q^{\text{jump}} = s_L b_j$. In the following discussion, agents are considered to be "in" the region which contains their optimally-chosen quality for the given mechanism (subject to tie-breaking). We qualitatively analyze the same comparison for fairness:

1. in the left-most region, low-skill agents receive 0-allocation according to both y^* and the threshold mechanism; by contrast, high-skill agents receive 0-allocation according to the threshold mechanism, but positive allocation according to y^* (the solid green line in Figure 3(Top); we suggest in this first region – regardless of the choice of fairness measure – that the fairness of y^* dominates the fairness of the threshold mechanism;
2. in the middle region, *all* low-skill agents receive allocation $y^*(q)$ for $q \to (q^{\text{jump}})^+$ (from above) according to y^* (by tie-breaking), which for *all* low-skill agents increases to full-allocation of 1 according to the threshold mechanism; whereas each high-skill agent (s_H, b_j is exhibiting its respective maximum quality $q^\dagger_{H,j} = s_H \cdot b_j$ and receives allocation $y^*(q^\dagger_{H,j})$ according to y^* which increases to full-allocation of 1 according to the threshold mechanism;
   in this second region, we can not make a dominance argument because it partially depends on the unknown densities of agent-types represented in this region and it also depends on the technical measure of fairness; however, ignoring expectation and proportional density and instead simply comparing agents one-to-one, we do observe that low-skill types receive the larger benefit (increase in allocation) if we start with y^* as our default mechanism and consider changing to the threshold mechanism; furthermore, the threshold mechanism abolishes the (properly oriented) cardinal difference between low-skill and high-skill agents by instead awarding them an "arguably unfair" constant allocation (of 1);
3. in the right-most region, all skill-types in all mechanisms receive the same allocation of 1; thus in this third region, the mechanism y^* and the threshold mechanism are equally fair (or equally unfair).

Intuitively, the preceding comparison between the optimal mechanism y^* and the threshold mechanism – which specifically have jumps at the same quality-index q^{jump} – suggests that (single-jump) slanted-stair mechanisms are indeed more fair. In fact, we have already stated a strict dominance relationship for an obvious, special-case choice of the technical fairness measure.

▶ **Corollary 17.** *Given a correlated admission game* $\mathcal{H} = (S = \Delta(\{s_L, s_H\}),$ $\mathcal{B}^S = \{B_{s_L}, B_{s_H}\}, \tau)$ *with* $0 < s_L < \tau < s_H$ *and* $0 < b_1 < b_2 < \ldots < b_n$. *For the fair mechanism design problem which maximizes the fairness measure set equal to the principal's utility function, the optimal mechanism and characterization of optimal mechanisms are determined identically to Theorem 13.*

Recall, the principal is naturally aligned with fairness. Then if we assign the fairness measure to be equal to the utility function of the principal, the analysis of the optimal mechanism for fairness gives the identical result as Theorem 13.