Testing under Strategic Manipulation: Mechanism Design for Human and AI Institutions

Xiaoyun Qiu[1]*, **Liren Shan**[2]*
[1] Dartmouth College
[2] Toyota Technological Institute at Chicago
xiaoyun.qiu@Dartmouth.edu, lirenshan@ttic.edu

Abstract

We study how the design of testing institutions, encompassing both the tests themselves and the procedures used to administer them, shapes selection outcomes in environments with multiple criteria and strategic agents. We model the testing agency as either a set of independent bureaucracies (each test administered separately) or a joint bureaucracy (where test order and personalization can be coordinated). Our mechanism design analysis shows that under a joint bureaucracy, fixed-order sequential mechanisms with stringent tests are optimal for maximizing the probability mass of qualified candidates selected. Furthermore, we demonstrate that personalizing tests through upfront communication, now increasingly feasible via AI and automation, can select all qualified candidates. Finally, we compare institutional settings and quantify the value of controlling test order, showing that the benefit depends critically on the distribution of testees and the stringency of optimal tests. Our results contribute to the design of robust, efficient, and fair testing systems in both human and AI-mediated environments.

1 Introduction

Testing plays a central role in decision-making, from regulatory agencies evaluating financial institutions or firms to algorithms screening job applicants or filtering online content. Whether implemented by human bureaucracies or artificial intelligence systems, testing aims to identify "qualified" entities under uncertainty. But tests are rarely perfect: they often rely on unverifiable inputs, and testees may manipulate their attributes at a cost. Recent research in economics and AI has highlighted how strategic manipulation, even subtle, can distort the goals of testing institutions (Hardt et al. 2016; Perez-Richet and Skreta 2022; Cohen et al. 2023).

In this paper, we ask: How should we design the institution that administers tests, especially when multiple criteria are involved? In particular, tests could be administered by a joint bureaucracy or separately by independent bureaucracies. Given the coordination power of a joint bureaucracy, what is the optimal organization of a joint bureaucracy? In particular, should tests be administered in a fixed order or randomized? Should they be uniform across agents, or personalized based on upfront communication, for instance, through AI-powered tools like chatbots or adaptive algorithms? We develop a mechanism design framework to address these questions, focusing on settings where agents can strategically manipulate their performance across multiple dimensions.

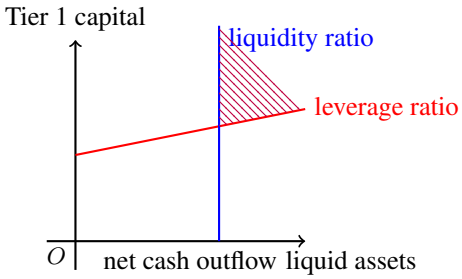


Figure 1: Regulatory requirements on a commercial bank

We begin with an example from the financial regulation on commercial banks. After the global financial crisis, central banks introduced multiple regulatory criteria to stabilize the banking system, including the leverage ratio (Tier 1 capital to total assets) and the liquidity ratio (liquid assets to net cash outflows over a 30-day period). Suppose that the minimum requirement is 3% for the leverage ratio and 1 for the liquidity ratio. Then the requirement on the leverage ratio is: Tier 1 capital $\geq 3\% \times$ liquid assets $+ 3\% \times$ other assets, and that on the liquidity ratio is: liquid assets $\geq$ net cash outflow. These requirements can be visualized in the space of Tier 1 capital and liquid assets (Fig. 1), where passing both thresholds places a bank in the regulatory approval region.

In Europe, banks report these ratios at different frequencies: quarterly for the leverage ratio and monthly for the liquidity ratio. This mismatch has led to documented cases of "window dressing", temporarily reducing liquid assets before the quarter-end to inflate the leverage ratio, then replenishing them afterward (Bassi et al. 2024; Egelhof, Martin, and Zinsmeister 2024). In response, the European Central Bank (ECB) raised regulatory requirements on selected banks.[1] While these behaviors are legal, they reflect strategic

*These authors contributed equally.

[1] In particular, a P2R leverage ratio add-on was applied to six

manipulation in response to predictable, rigid testing procedures.

More broadly, this example illustrates a core challenge: the ECB seeks to pass qualified banks and fail unqualified ones (*selection accuracy*), but is limited to using (1) tests: rules for evaluating performance, and (2) a testing procedure: how and when tests are administered. In practice, tests are often applied in a fixed sequence, but alternative designs are possible. For instance, under *independent bureaucracies*, each test is administered by a separate testing party, and the testee chooses the order. Under a *joint bureaucracy*, the testing party can coordinate tests, fix or randomize the order, or even personalize tests for each agent. We analyze this broader design space through a mechanism design lens. Our framework encompasses fixed-order, random-order, disclosure decision, and personalized sequential testing procedures. Personalization is becoming increasingly feasible via AI systems. It allows the testing party to tailor tests based on cheap-talk communication, enabling greater flexibility and cost-efficiency.

While motivated by financial regulation, our results apply directly to modern AI systems that serve as testing agents. From automated hiring tools and credit scoring models to content moderation and personalized education platforms, these systems evaluate agents based on multiple criteria, often using unverifiable information. As AI systems grow more interactive and adaptive, strategic manipulation, test sequencing, and institutional structure become central design concerns. Our work provides a general framework for designing such testing systems, whether human, algorithmic, or hybrid.

Our main findings are as follows. First, fixed-order sequential mechanisms with stringent tests are optimal for maximizing the probability mass of qualified candidates selected (*selection efficiency*, Theorem 1). Under the same tests, fixing the test order leads to more agents passing both tests. However, it requires more stringent tests to exclude unqualified agents. We show that the choice of tests and the testing procedure cannot be separated: the set of feasible tests under a random-order procedure is typically larger, but we construct a fixed-order mechanism that dominates any random-order one. This involves a novel two-step constructive argument, detailed in Section 3. Interestingly, the optimal tests may aggregate attributes differently from the true criteria, leading to *non-parallel tests*.

Second, personalizing tests can improve selection efficiency without sacrificing accuracy (Theorem 2). We construct two fixed-order procedures with different tests. Each excludes unqualified testees, and together they admit all qualified ones. With upfront communication, for instance, via chatbots, the testing party can assign testees to the appropriate procedure without running two rounds of testing. This enables cost-efficient implementation without sacrificing accuracy.

Third, comparing institutional designs, we analyze the value of controlling test order. Any equilibrium under independent bureaucracies (where testees choose the order) can be implemented under a joint bureaucracy. But giving testees control over sequencing introduces a tradeoff: it reduces the cost for some qualified testees but may force the testing party to raise thresholds to block manipulation by unqualified ones. Whether this tradeoff is beneficial depends on the distribution of types and the stringency of optimal tests.

Our work is closely related to Cohen et al. (2023), who compare simultaneous and fixed-order testing mechanisms. We build on their insights but expand the scope in two key directions: first, by considering a broader class of sequential mechanisms, including randomized and personalized procedures; and second, by explicitly modeling the institutional organization of the testing agency, whether tests are coordinated or decentralized. These extensions are particularly relevant in AI-mediated environments, where adaptivity and personalization are technologically feasible.

To summarize, we study how the design of tests and testing procedures affects selection under strategic behavior. Our results apply to a wide range of settings, from banking oversight to automated decision-making, and offer new insights into how institutional design interacts with information and manipulation in complex evaluation systems.

2 Model

We study the design of testing institutions that evaluate whether an agent (henceforth he) possesses desirable attributes. The agent's *true* attributes (type) are denoted by $\boldsymbol{x}^0 \in X = \mathbb{R}^d$, for any $d \geq 2$, and are privately known to him. The testing parties only observe the distribution: $\boldsymbol{x}^0 \sim \boldsymbol{\mathcal{F}}$. The agent can strategically modify his attributes to a different vector $\boldsymbol{x}$ at a cost $C(\boldsymbol{x}^0, \boldsymbol{x})$. In large parts of the paper, we will assume that $C(\boldsymbol{x}^0, \boldsymbol{x}) = \eta \cdot \|\boldsymbol{x} - \boldsymbol{x}^0\|_2$, where $\|\cdot\|_2$ is the the Euclidean norm and $\eta = 1$ without loss of generality.[2] The true attributes $\boldsymbol{x}^0$ remain unchanged; only the altered attributes $\boldsymbol{x}$ are used in testing.

There are two testing parties, $i \in A, B$, each with a criterion defined by a linear constraint $h_i = \{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{w}_i \cdot \boldsymbol{z} \geq 0\}$. An agent is said to be *qualified* for party i if his true attributes satisfy $\boldsymbol{x}^0 \in h_i$. Let θ denote the angle between the separating hyperplanes of the two criteria.

Each party i selects a test $\widetilde{h}_i$, also defined by a linear constraint. A test determines whether the agent's attributes (possibly manipulated) at the time of the test belong to the halfspace $\widetilde{h}_i$. That is, the tests apply to current altered attributes, not the true ones. We analyze two institutional settings:

Independent bureaucracies. Each party independently selects a test $\widetilde{h}_i$. The agent observes the tests and chooses the order in which to take them.[3] Suppose the agent chooses $t_1 \in \{A, B\}$ to take first, and t_2 as the second. He can change his attributes from $\boldsymbol{x}^0$ to $\boldsymbol{x}^1$ before taking the first test, and from $\boldsymbol{x}^1$ to $\boldsymbol{x}^2$ before taking the second test.

The total cost is denoted by $c(\boldsymbol{x}^0, \boldsymbol{x}^1, \boldsymbol{x}^2)$. In most of the paper we assume additivity: $c(\boldsymbol{x}^0, \boldsymbol{x}^1, \boldsymbol{x}^2) = C(\boldsymbol{x}^0, \boldsymbol{x}^1) +$

banks (European Central Bank December 19 2023).

[2] The scaling constant η can be normalized to 1.

[3] Because the testing parties operate independently, it is natural to assume they cannot coordinate on the testing order.

$C(\boldsymbol{x}^1, \boldsymbol{x}^2)$. The agent wants to pass the tests and receives a payoff of 1 upon passing. His utility is

$$\mathbf{1}\left[\boldsymbol{x}^1 \in \widetilde{h}_{t_1} \cap \boldsymbol{x}^2 \in \widetilde{h}_{t_2}\right] - c(\boldsymbol{x}^0, \boldsymbol{x}^1, \boldsymbol{x}^2). \tag{1}$$

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

Each party i aims to maximize the probability that a qualified agent (i.e., one with $\boldsymbol{x}^0 \in h_i$) passes the test, subject to the constraint that no unqualified agent is ever selected:

$$\begin{array}{ll} \max & \mathbf{P}[\text{selecting qualified agent}] \\ s.t. & \text{no unqualified agent is selected;} \end{array} \tag{$\mathcal{P}_I$}$$

This lexicographic preference reflects settings where false positives are significantly more costly than false negatives. Such settings are common in regulation, hiring, and safety-critical applications.

Joint Bureaucracy. Now suppose the two parties act as a single principal. The principal's criterion is the intersection $H = h_A \cap h_B$, illustrated as the shaded region in Fig. 1. The principal chooses two tests $(\widetilde{h}_A, \widetilde{h}_B)$ and implements a *sequential mechanism*, deciding the order in which the agent is tested.[4]

Let $t_1 \in \{A, B\}$ denote the first test, chosen by the principal with probability q. When the test order is deterministic ($q \in \{0, 1\}$), the sequential mechanism is described by the tuple $s = (\widetilde{h}_A, \widetilde{h}_B, q)$, meaning that test $\widetilde{h}_A$ is applied as first with probability q, and test $\widetilde{h}_B$ as first with the complementary probability.[5] When the order is stochastic ($q \in (0, 1)$), the sequential mechanism is described by $s = (\widetilde{h}_A, \widetilde{h}_B, q, D)$. Here, $D \in \{\widetilde{h}_1, \varnothing\}$ indicates whether the identity of the first test is revealed to the agent after it is applied, and $\widetilde{h}_1 \in \{\widetilde{h}_A, \widetilde{h}_B\}$ denotes the first test performed.

The agent can again change his attributes twice: (1) from $\boldsymbol{x}^0$ to $\boldsymbol{x}^1$ before the first test, and (2) from $\boldsymbol{x}^1$ to $\boldsymbol{x}^2$ before the second test. When $q \in (0, 1)$, the agent only knows that $\widetilde{h}_A$ will come first with probability q when choosing $\boldsymbol{x}^1$. If $D = \widetilde{h}_1$, he learns the first test realized before choosing $\boldsymbol{x}^2$; otherwise, he proceeds under uncertainty.

The agent's expected utility under a sequential mechanism is

$$\mathbf{E}\left\{\mathbf{1}\left[\boldsymbol{x}^1 \in \widetilde{h}_1 \cap \boldsymbol{x}^2 \in \widetilde{h}_2\right] - c(\boldsymbol{x}^0, \boldsymbol{x}^1, \boldsymbol{x}^2)\right\},$$

where the expectation is taken over both the randomness in test order and the agent's decision.

The principal's objective is $\mathcal{P}_I$, now interpreted as maximizing the probability that a qualified agent in H passes both tests, subject to the constraint that no unqualified agent ever does.

We will now recap the timeline of the game induced by any mechanism: (1) The principal commits to a sequential mechanism s. (2) The agent first chooses $\boldsymbol{x}^1$. (3) A random device determines the test order, and test $\widetilde{h}_1$ is applied. (4) If $D = \widetilde{h}_1$, the identity of the first test is revealed. (5) The agent chooses $\boldsymbol{x}^2$. (6) Test $\widetilde{h}_2$ is applied. (7) The agent passes the tests if both tests are passed: $\boldsymbol{x}^1 \in \widetilde{h}_1$ and $\boldsymbol{x}^2 \in \widetilde{h}_2$, and the game ends.[6]

3 Optimal Organization of A Joint Bureaucracy

First, we show that the optimal sequential mechanism offers the two tests in a known, fixed order.

Theorem 1. *For any distribution $\mathcal{F}$, the optimal sequential mechanism uses:*

1. Stringent *tests $\widetilde{h}_A$ and $\widetilde{h}_B$, such that $\widetilde{h}_A \cap \widetilde{h}_B \subset h_A \cap h_B$;*
2. *A fixed-order procedure.*

We define a test as *stringent* if it is stricter than the true criterion. When there is only one criterion, stringency corresponds to raising the threshold, as discussed by Perez-Richet and Skreta (2022). However, with multiple criteria, stringency involves the intersection of tests being strictly contained within the original qualified region: $\widetilde{h}_A \cap \widetilde{h}_B \subset h_A \cap h_B$.[7]

In addition to selecting the tests, the choice of testing procedure has a surprising and nontrivial impact on outcomes. To formally compare procedures, we define one procedure to be more stringent than another if, under the same tests, it selects a smaller set of types (or selects with lower probability). This definition allows us to rank procedures by stringency.

We establish the following hierarchy of procedures: First, *random-order without disclosure* is more stringent than *fixed-order*: Under the same tests, the set of types who can pass both tests with cost no larger than one is larger under the fixed-order procedure (see Fig. 2).

Second, *random-order with disclosure* is also more stringent than fixed-order, although the relation is more subtle. While the set inclusion relation does not hold (Fig. 3), the probability measure of types who pass both tests with cost no greater than one is smaller under a random-order procedure with disclosure than under the fixed-order procedure. To show this, consider an arbitrary $(h_A, h_B, q, \widetilde{h}_1)$ for $q \in (0, 1)$. We introduce a mixed mechanism, which is a convex combination of two fixed-order mechanisms: with probability q, offer $(h_A, h_B, 1)$; with probability $1 - q$, offer $(h_A, h_B, 0)$. It dominates the random-order one in terms of selection efficiency by applying a similar set inclusion argument. The construction of the mixed mechanism determines that it is dominated by one of the fixed-order mechanisms.

Third, random-order without disclosure is more stringent than random-order with disclosure. For instance, in Fig. 3,

[4] The principal can replicate the independent bureaucracy outcome by allowing the agent to choose the order, so we focus on settings where the principal uses this control strategically.

[5] Note that $(\widetilde{h}_A, \widetilde{h}_B, q) = (\widetilde{h}_B, \widetilde{h}_A, 1 - q)$.

[6] Of course, if $\boldsymbol{x}^1 \notin \widetilde{h}_1$ and the mechanism discloses the first test, the agent fails the first test, and the game could end at this point.

[7] One might conjecture that stringency requires each test to be a stricter version of the original (i.e., $\widetilde{h}_i \subset h_i$ for $i \in \{A, B\}$). However, as shown in Fig. 5b, non-parallel tests can outperform parallel ones under certain distributions.

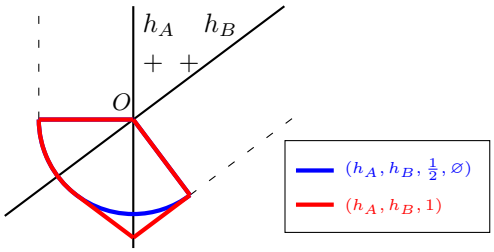


Figure 2: Random-order without disclosure is a more stringent procedure than fixed-order. The area in blue and in red are types with profitable strategies to pass both tests under a random-order procedure without disclosure $(h_A, h_B, \frac{1}{2}, \varnothing)$, and that under a fixed-order procedure $(h_A, h_B, 1)$, respectively.

type C can exploit the disclosure if the first test happens to be h_A, in which case he can pass both tests at cost 1. Without disclosure, C's expected utility from using the same strategy is negative, and C would not be selected.

Since both the stringency of the tests and the procedure affect outcomes, the key design question becomes: Which combination is optimal? While fixed-order procedures can help qualified agents get selected more easily, they also lower the barrier for unqualified types. The latter restricts the set of feasible test pairs under fixed-order procedures compared to random-order ones. Ex ante, it's unclear which configuration is optimal.

Our main result resolves this tension. We show that the optimal mechanism combines the least stringent testing procedure, fixed-order, with stringent tests. We establish this via two constructive arguments that handle random-order mechanisms with and without disclosure: Lemma 1 shows that any feasible random-order mechanism with disclosure can be weakly dominated by one of two fixed-order procedures using the same tests. Lemma 2 addresses the more challenging case of random-order mechanisms without disclosure, where feasible tests can be non-parallel. We construct two fixed-order mechanisms using one original and one adjusted test. Although proving feasibility is delicate, we again show that one of the two dominates in terms of selection efficiency. Together, these results imply that a fixed-order procedure using well-chosen stringent tests is optimal, even in the face of strategic manipulation.

Lemma 1. *Suppose the random-order mechanism with disclosure $(\widetilde{h}_A, \widetilde{h}_B, q, \widetilde{h}_1)$ is feasible. Then the two fixed-order mechanisms $(\widetilde{h}_A, \widetilde{h}_B, 1)$ and $(\widetilde{h}_A, \widetilde{h}_B, 0)$ are feasible and one of them is (weakly) better than the random-order mechanism with disclosure.*

We claim that the set of tests feasible for random-order mechanisms with disclosure is the same as that for fixed-order mechanisms. The challenge of this proof comes from that fixing the tests, neither of the two fixed-order mechanisms is directly comparable to any random-order mechanisms with disclosure (Fig. 3). Our proof strategy is to leverage a mixed mechanism: Given any random-order mechanism with disclosure $(\widetilde{h}_A, \widetilde{h}_B, q, \widetilde{h}_1)$, consider

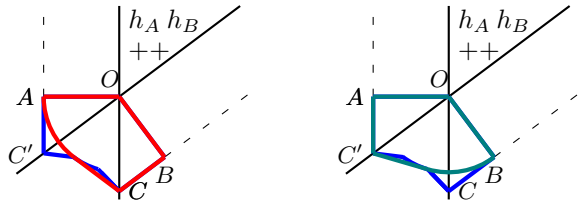


Figure 3: Random-order procedure with disclosure vs fixed-order procedures. The area in blue, red, and green are types with profitable strategies to pass both tests under a random-order procedure without disclosure $(h_A, h_B, \frac{1}{2}, \varnothing)$, and that under fixed-order procedures $(h_A, h_B, 1)$ and $(h_A, h_B, 0)$, respectively.

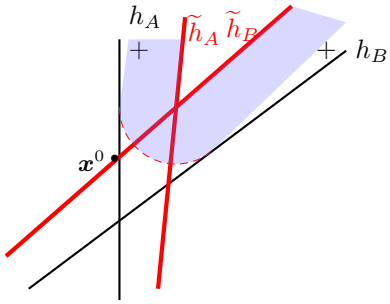


Figure 4: Feasible non-parallel tests for random-order mechanism without disclosure

a mixed mechanism that offers the fixed-order mechanism $(\widetilde{h}_A, \widetilde{h}_B, 1)$ with probability q and $(\widetilde{h}_A, \widetilde{h}_B, 0)$ with probability $1-q$. Applying a set inclusion argument, we show that the mixed mechanism dominates the random-order mechanism with disclosure. Then, applying a probability argument, we show that one of the fixed-order mechanisms dominates the mixed mechanism.

Lemma 2. *Fix any feasible random-order mechanism without disclosure $(\widetilde{h}_A, \widetilde{h}_B, q, \varnothing)$. Then there exist two feasible fixed-order mechanisms: $(\widetilde{h}_A, h_B^+, 1)$ and $(h_A^+, \widetilde{h}_B, 0)$. Moreover, one of them is (weakly) better than the random-order mechanism without disclosure.*

As we have eluded earlier, the set of tests feasible for random-order mechanisms without disclosure is usually larger. Let us use Fig. 4 to illustrate. Under the random order mechanism without disclosure $(\widetilde{h}_A, \widetilde{h}_B, \frac{1}{2}, \varnothing)$, every selected agent uses a one-step strategy to pass both tests together. Hence the set of attributes being selected under this mechanism is the blue region and $\widetilde{h}_A \cap \widetilde{h}_B$. Since the blue region is contained in the qualified region, this mechanism is feasible. However, since $\widetilde{h}_i$ is not parallel to $h_i, i \in \{A, B\}$, the corresponding fixed-order mechanisms are not feasible. To see this, consider the unqualified attributes $\boldsymbol{x}^0 \in \widetilde{h}_B$ that are very close to h_A but do not satisfy h_A. Because the tests are non-parallel, we can always find such $\boldsymbol{x}^0$ so that its cost to pass $\widetilde{h}_A$ is less than one. This implies that the fixed-order mechanism with $\widetilde{h}_B$ as the first test is not feasible. Similarly,

we can argue that the fixed-order mechanism with $\widetilde{h}_A$ as the first test is not feasible.

Given this challenge, we construct two fixed-order mechanisms, each uses one original test ($\widetilde{h}_A$ or $\widetilde{h}_B$), one constructed test, and a careful choice of test order. Showing that both constructed fixed-order mechanisms are feasible (without selecting any unqualified agent) is technically challenging. However, after this step, we show dominance with the help of a mixed mechanism as discussed above. All omitted proofs are in Appendix C.

4 Personalized Mechanisms: Joint Bureaucracies with Communication

We now consider the optimal design of a joint bureaucracy when the principal can personalize testing procedures based on agent-specific information. Such personalization becomes feasible when agents and the principal can engage in cheap-talk communication: costless messages exchanged before test selection.

Advances in automation and AI make such interactions increasingly practical. Digital platforms can collect structured inputs (e.g., preferred formats or test orderings), and conversational agents, such as chatbots or automated advisors, can tailor bureaucratic processes dynamically.

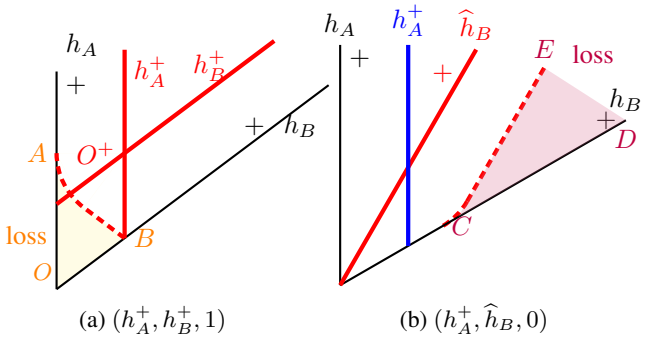
Formally, let (M, s) denote a sequential mechanism, where M is the message space and $s : M \to \mathcal{T}$ maps messages to sequential testing procedures. A direct mechanism sets $M = X$, and the mechanism becomes a menu of testing procedures: $< \widetilde{h}_A(m), \widetilde{h}_B(m), q(m), D(m); m \in M >$.

Given such a mechanism, each agent selects a message $m \in M$ and a strategy to manipulate their attributes before taking each test to maximize their utility under the assigned procedure $s(m)$. The following result shows that cheap-talk personalization enables a principal to select all qualified types, achieving the first-best outcome.

Theorem 2. *When up-front communication is available, there exists a feasible fixed-order sequential mechanism that selects* all *and* only *the qualified types for any distribution* $\mathcal{F}$.

The proof is constructive. Let $h_i^+, i \in \{A, B\}$ denote the test obtained by shifting h_i along its normal vector $\boldsymbol{w}_i$ by a distance of 1 (Fig. 5a), and let O^+ be the intersection of their boundaries. Define a rotated test $\widehat{h}_B$ by rotating h_B^+ around O^+ until its boundary bisects the angle between h_A^+ and h_B^+. This rotation ensures the new classifier shares a boundary point O with h_A and h_B, the true criteria. Now compare two fixed-order procedures: $(h_A^+, h_B^+, 1)$, which applies stringent parallel tests h_A^+ before h_B^+; and $(h_A^+, \widehat{h}_B, 0)$, which applies one rotated, non-parallel test $\widehat{h}_B$ before the stringent parallel test h_A^+. The first procedure misses the triangle-like region OAB. This region, however, is covered by the second procedure. Offering a menu of the two procedures allows the principal to cover the full qualified region without selecting any unqualified types.

An alternative to the cheap-talk approach is to run both procedures in sequence and accept any agent who passes either one. However, this doubles the testing workload. By contrast, via cheap-talk communication, the appropriate test can be assigned ex ante. This maintains full selection efficiency while halving testing costs. As the advancement of automated tools makes implementing personalized tests technologically feasible, our finding underscores the potential of AI tools to enhance the efficiency and scalability of bureaucratic processes in evaluation, certification, and regulation.



(a) $(h_A^+, h_B^+, 1)$ (b) $(h_A^+, \widehat{h}_B, 0)$

Figure 5: A menu of procedures

5 Independent Bureaucracies vs. Joint Bureaucracy: The Value of Test Ordering

We now analyze equilibrium outcomes under *independent bureaucracies*, a setting where each testing agency (A and B) independently selects its own test. The agent strategically chooses how to present themselves in each test to maximize their chance of selection.

Unlike the joint bureaucracy setting, where the principal designs a unified sequential mechanism, independent bureaucracies lack coordination. Nevertheless, any equilibrium under independent bureaucracies can be implemented by a joint bureaucracy that delegates test order to the agent. This equivalence enables us to isolate the design value of controlling the *test order* itself.

Let $h_A^\dagger, h_B^\dagger$ denote the equilibrium tests under independent bureaucracies, and let $S^\dagger$ be the probability mass of qualified agents who pass both. Likewise, let h_A^*, h_B^* be the tests in the optimal sequential mechanism under the joint bureaucracy, and let S^* be the probability mass of qualified agents selected.

Proposition 1. *If* $h_A^*, h_B^* \in \{h_A^+, h_B^+\}$*, then there exists an equilibrium under independent bureaucracies such that* $S^\dagger > S^*$.

If either $h_i^* \notin \{h_A^+, h_B^+\}$ *or* $h_i^\dagger \notin \{h_A^+, h_B^+\}$ *for some* $i \in \{A, B\}$*, the comparison between* $S^\dagger$ *and* S^* *depends on the distribution* $\mathcal{F}$.

Strategic Behavior and the Cost of Controlling Test Order. To prove the first part, we characterize the agent's best response under any pair of tests. Suppose the mechanism uses test h_A first, followed by h_B. Any agent whose true attributes satisfy both tests passes without manipulation. But

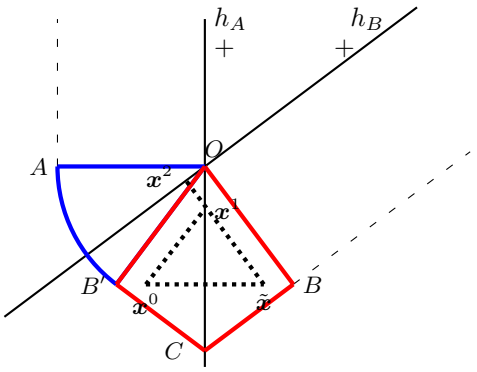


Figure 6: Agent's best response in $(h_A, h_B, 1)$

an unqualified agent has multiple strategic options: (1) *one-step* manipulation: directly adopt attributes in the intersection of tests. (2) *Two-step* manipulation: pass one test with a manipulated profile, then shift again to pass the second.

A powerful geometric insight (adapted from Cohen et al. 2023) is that the agent's optimal two-step manipulation can be visualized by reflecting their true type $\boldsymbol{x}^0$ across the first test boundary. The minimal-cost path aligns when this reflected point $\widetilde{\boldsymbol{x}}$, the first manipulated point $\boldsymbol{x}^1$, and the second manipulated point $\boldsymbol{x}^2$ are co-linear (Fig. 6). This mirrors Fermat's principle in optics, where a light beam reflects to minimize path length.

As shown in Cohen et al. (2023), for agents in the region $OBCB'$, the two-step strategy is strictly cheaper than the one-step move to O, due to the triangle inequality. The challenge in our problem is that when the testing order is unknown (in random-order mechanisms) or up to the choice of the agent (under independent bureaucracies), a third option can also be optimal: The agent can adopt attributes $\boldsymbol{x}^1 \in h_B$ for the first test, and after passing the first test, adopt yet other attributes $\boldsymbol{x}^2 \in h_A$ (right panel in Fig. 3). Our proof idea is to first characterize the set of types with a cost no larger than one under each option, then the union of the three sets contains all types that will be selected.

Crucially, in independent bureaucracies, agents are free to choose the test order, enabling not only the A-to-B path, but also the reverse. Hence, if h_A^+, h_B^+ are the equilibrium tests, the set of agents selected in equilibrium includes: (1) those in $h_A^+ \cap h_B^+$; (2) those who can reach the intersection $h_A^+ \cap h_B^+$ at cost no larger than one; (3) those whose cost for the two-step strategy: $\boldsymbol{x}^1 \in h_A^+$ and $\boldsymbol{x}^2 \in h_B^+$ is no larger than one; and (4) those whose cost for the two-step strategy: $\boldsymbol{x}^1 \in h_B^+$ and $\boldsymbol{x}^2 \in h_A^+$ is no larger than one.

In contrast, the optimal mechanism under joint bureaucracy fixes a test order. So only one manipulation path is available, excluding some agents (either (3) or (4)) selected under independent bureaucracies. This flexibility advantage is the benefit of independent bureaucracies.

To complete the proof, consider the best equilibrium using $h_A^\dagger, h_B^\dagger$ under independent bureaucracies. We argue that if h_A^+, h_B^+ are the optimal tests under the joint bureaucracy, then the best equilibrium under independent bureaucracies (h_A^+, h_B^+ may not be the equilibrium tests) must be weakly better than the optimal sequential mechanism using h_A^+, h_B^+. This is because the optimal *joint mechanism with agent-chosen test order* uses the best equilibrium tests $h_A^\dagger, h_B^\dagger$. Since tests h_A^+, h_B^+ are also feasible under the joint mechanism with agent-chosen test order, the optimal tests must select more potential qualified agents.

The Benefit of Controlling Test Order. The second part of Proposition 1 highlights the downside of losing control over the test sequence. When the tests in the optimal sequential mechanism under joint bureaucracy are not h_A^+, h_B^+, it implies that either they are not very stringent, or only a specific order is feasible. For example, in Fig. 5b, the optimal mechanism uses a non-parallel test $\widehat{h}_B$, which excludes all unqualified types when paired with h_A^+ as the second test. But in independent bureaucracies, this test cannot be safely used: unqualified agents could game the order and pass both tests with low effort, making $\widehat{h}_B$ not feasible.

Hence, control over test order allows the joint bureaucracy to employ a larger set of tests, improving selection efficiency. But this comes at the cost of the agents, which potentially reduces access for some qualified types. Proposition 1 suggests that when the optimal tests under the joint bureaucracy are h_A^+, h_B^+, there is a negative value of controlling the order of tests. However, when one optimal test is not h_A^+ or h_B^+, then the value of controlling the order of tests depends on the distribution.

Implications for AI and Automated Institutions. Independent bureaucracies mirror decentralized platforms, where each subsystem makes decisions based on its own algorithm or classifier. Joint bureaucracies correspond to centralized systems that coordinate components and fix decision pathways. Modern AI tools can help resolve this tradeoff: When a system can infer the best test order for each type (possibly via cheap-talk), it restores flexibility while maintaining centralized control.

6 Discussions

Investment. In AI-driven institutions, such as hiring platforms, educational testing, or online certification systems, designers increasingly aim to promote real skill acquisition rather than superficial optimization. In this subsection, we switch gear and consider the *investment* setting. The investment setting captures this modern shift: we don't just want people to "game" the algorithm; we want them to improve in meaningful ways.

As before, the agent starts with $\boldsymbol{x}^0$ and can invest effort to attain $\boldsymbol{x}$ at a cost $C(\boldsymbol{x}^0, \boldsymbol{x})$. The key difference is that we assume that an agent's new attributes $\boldsymbol{x}$ become their true attributes in the investment setting. While the agent's strategic incentive remains unchanged, the designer's goal shifts: instead of screening manipulators, the principal now wants to encourage agents to become qualified.

This subtle but crucial change in institutional objective calls for different design. Under a mild geometric condition, we show that a random-order sequential mechanism without disclosure achieves the first best in the investment setting.

Theorem 3. *If $\theta \geq 30°$, for any distribution $\mathcal{F}$, then the random order mechanism without disclosure $(h_A, h_B, \frac{1}{2}, \varnothing)$ achieves the first best.*

The key insight lies in how agents respond to uncertainty in test order. When the angle θ between test boundaries is sufficiently large, every agent who ultimately gets accepted under this mechanism finds it optimal to invest directly in a one-step strategy, choosing attributes that pass both tests simultaneously (Lemma 11 in Appendix E). This property ensures selection efficiency. This mechanism highlights the benefit of coordination: It leverages randomness and non-disclosure to induce genuine improvement. All omitted proofs are provided in Appendix E.

More tests do not help. In the manipulation setting, adding more tests only makes a procedure more stringent and it is counter-productive. Suppose the principal offers test h_1 and h_2 repeatedly. For simplicity, suppose the first test is h_1, the second is h_2 and the third is h_1. For those types that find it profitable to pass both tests together under a fixed-order procedure with two tests, adding an extra test does not change their incentives. For those types that find it profitable to pass one test at a time, adding a third test only makes it more costly. Hence, the benefit of the fixed-order procedure is diminished with more tests.

7 Relation to the Literature

We contribute to the literature on designing algorithms that interact with strategic agents, stemming from Brückner and Scheffer (2009); Hardt et al. (2016). Prior studies in this area primarily examine the impact of strategic behavior in response to a classification algorithm (i.e., a linear test). We are most related to Cohen et al. (2023). They studied how the agent can exploit the sequential ordering of tests in fixed-order sequential mechanisms to achieve a favorable outcome at a limited cost. The main differences are twofold. First, building on Cohen et al. (2023), we conduct a comprehensive mechanism design analysis that encompasses random sequential mechanisms and personalized procedures, which are central to three key results in our paper (Theorem 1, 2, and 3).[8] Second, we explicitly model the institutional organization of the testing agency, incorporating scenarios where tests are coordinated or decentralized. We discuss additional related work in Appendix A.

We also contribute to a line of work that considers multi-round decision-making and screening processes, where decisions are made sequentially and each stage can influence subsequent outcomes (Bower et al. 2017; Dwork and Ilvento 2019; Dwork, Ilvento, and Jagadeesan 2020). However, these works focus primarily on fairness rather than the strategic behavior of agents. Harris, Heidari, and Wu (2021) study a multi-round testing model involving strategic agents. The principal aims to maximize the investments of agents, where the agent's utility is the sum of scores across all tests. This additive formulation enables each stage to be analyzed independently. In contrast, the agent's utility in our setting depends on the final outcome, i.e., whether they pass all tests. Thus, the agent's incentives are influenced by the entire structure of the testing pipeline.

Conceptually, our paper is closely related to an Economics literature on test design where the agent can take hidden action to affect the outcome of the test. Perez-Richet and Skreta (2022) study a setting where the agent's type is one-dimensional and the principal can use only one test, while in our setting, the agent's attributes are high-dimensional and the principal uses two tests to select the agent. The multi-dimensionality introduces a new tradeoff on setting tests and determining a testing procedure. We also study a setting where the agent invests, which is not studied by them. Deb and Stewart (2018) study the design of adaptive testing, where the choice of the next test depends on the history of previous tests and their outcomes. In their model, the agent has a discrete type and selects independent effort levels in $[0, 1]$ for each test. They characterize the optimal mechanism in terms of test informativeness. Frankel and Kartik (2022); Ball (2025) study settings in which the principal aims to infer the agent's quality (natural type) based on their manipulation effort to minimize the quadratic loss. While the quality in their settings is similar to the original attributes in our manipulation setting, the objectives of the principal differ significantly. Moreover, they focus on studying a linear scoring rule under different commitment settings (i.e., whether the principal commits to the rule before the agent acts), rather than on the sequencing of multiple tests.

8 Conclusion

Testing institutions, whether human or algorithmic, increasingly face the challenge of evaluating agents across multiple criteria using imperfect, and often unverifiable, information. In such environments, strategic manipulation is a predictable response to rigid and transparent testing procedures. We develop a mechanism design framework to study how the sequencing, coordination, and personalization of tests can mitigate manipulation and improve selection. We show that fixed-order mechanisms with carefully chosen tests can outperform randomized procedures, and that personalizing tests via cheap-talk communication can achieve selection efficiency. We show that institutional structure matters. While independent bureaucracies leave the flexibility to agents, joint bureaucracies enable better control over manipulation by coordinating test orders.

Although motivated by financial regulation, our framework applies broadly, from automated hiring to algorithmic credit scoring. We highlight that the effectiveness of testing institutions depends not only on what is tested but also on how testing is organized and administered. Careful institutional design, leveraging both economic insights and modern AI capabilities, can help align incentives and build more robust evaluation systems across domains.

[8] It is worth noting that Theorem 4.4 of Cohen et al. (2023) implicitly assumed that optimal tests are shifted parallel tests. In contrast, we identify scenarios where the optimal tests are not shifted parallel tests.

Acknowledgments

We are grateful to Asher Wolinsky and Wojciech Olszewski for their valuable insights and constructive feedback. We also benefited greatly from the advice and suggestions of Bruno Strulovici. We acknowledge Abhishek Sarkar for inspiring conversations throughout the development of this project, as well as Junko Oguri and Matthew Thomas for their domain expertise on the applications.

Additionally, we thank Nemanja Antic, Ian Ball, Eddie Dekel, David Dranove, Piotr Dworczak, Yingni Guo, Joshua Mollner, Marciano Siniscalchi, Alison Zhao, Saeed Sharifi Malvajerdi, and Kevin Stangl for their helpful suggestions, comments, and discussions. Any remaining errors are our own.

References

Ahmadi, S.; Beyhaghi, H.; Blum, A.; and Naggita, K. 2022. On classification of strategic agents who can both game and improve. arXiv:2203.00124.

Ball, I. 2025. Scoring strategic agents. *American Economic Journal: Microeconomics*, 17(1): 97–129.

Bassi, C.; Behn, M.; Grill, M.; and Waibel, M. 2024. Window dressing of regulatory metrics: evidence from repo markets. *Journal of Financial Intermediation*, 58: 101086.

Bower, A.; Kitchen, S. N.; Niss, L.; Strauss, M. J.; Vargas, A.; and Venkatasubramanian, S. 2017. Fair pipelines. *arXiv preprint arXiv:1707.00391*.

Brückner, M.; and Scheffer, T. 2009. Nash equilibria of static prediction games. *Advances in neural information processing systems*, 22.

Cohen, L.; Sharifi-Malvajerdi, S.; Stangl, K.; Vakilian, A.; and Ziani, J. 2023. Sequential strategic screening. In *International Conference on Machine Learning*, 6279–6295. PMLR.

Deb, R.; and Stewart, C. 2018. Optimal adaptive testing: Informativeness and incentives. *Theoretical Economics*, 13(3): 1233–1274.

Dwork, C.; and Ilvento, C. 2019. Fairness Under Composition. In *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*.

Dwork, C.; Ilvento, C.; and Jagadeesan, M. 2020. Individual Fairness in Pipelines. In *1st Symposium on Foundations of Responsible Computing*.

Egelhof, J.; Martin, A.; and Zinsmeister, N. 2024. Regulatory incentives and quarter-end dynamics in The repo market. https://libertystreeteconomics.newyorkfed.org/2017/08/regulatory-incentives-and-quarter-end-dynamics-in-the-repo-market/.

European Central Bank. December 19 2023. Speech by Andrea Enria, Chair of the Supervisory Board of the ECB, at the press conference on the 2023 SREP results and the supervisory priorities for 2024-26. https://www.bankingsupervision.europa.eu/press/speeches/date/2023/html/ssm.sp231219~421bbae836.en.html.

Frankel, A.; and Kartik, N. 2022. Improving information from manipulable data. *Journal of the European Economic Association*, 20(1): 79–115.

Haghtalab, N.; Immorlica, N.; Lucier, B.; and Wang, J. Z. 2020. Maximizing Welfare with Incentive-Aware Evaluation Mechanisms. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 160–166. International Joint Conferences on Artificial Intelligence Organization. Main track.

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*, 111–122.

Harris, K.; Heidari, H.; and Wu, S. Z. 2021. Stateful strategic regression. *Advances in Neural Information Processing Systems*, 34: 28728–28741.

Kleinberg, J.; and Raghavan, M. 2020. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4): 1–23.

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