Inputs or Outputs: What to Test and How to Test

Matteo Camboni[*] Christoph Carnehl[†]

November 21, 2025

We study optimal test design in settings where the testing variable is itself a choice. Agents with heterogeneous productivity invest inputs (such as money or effort) to increase outputs (such as product quality or human capital) that they sell in a competitive market. The market cares only about outputs and receives credible information solely through the ratings assigned by a public test. Aiming to maximize expected output, the test designer may base ratings on inputs, outputs, or any combination of the two. Although both the market and the designer ultimately care only about outputs, output-only tests are always dominated because they allow high-productivity agents to "coast on their talent" and pass with minimal input. By contrast, input-only tests best incentivize input investments across all types and are optimal if the designer can coordinate the market and agents on her preferred equilibrium. Yet input tests are fragile: because they provide no guarantee on output, which still depends on type, they are vulnerable to no-investment equilibria. To balance robustness to adverse equilibria with input incentive provision, the designer adopts tests that optimally combine input and output components. For pass-fail tests, the optimal design takes the form of a *step test* with one input threshold and two output thresholds: agents pass either by meeting the higher output bar or by satisfying the minimum output requirement along with the input threshold.

[*] Camboni: University of Wisconsin-Madison; email: camboni@wisc.edu

[†] Carnehl: Bocconi University and IGIER; email: christoph.carnehl@unibocconi.it

1 Introduction

Certification systems are central to modern economic activity. They influence market transactions, hiring and promotion, and the conduct of firms across industries. Although certifications disclose performance information, their impact extends well beyond disclosure: by shaping market beliefs, they also shape the incentives and effort choices of those being evaluated. As a result, designing a certification system is fundamentally an incentive-design problem.

A defining feature of certification systems is that designers must decide which aspects to evaluate and communicate through ratings. A certification can focus on outputs, the attributes ultimately valued by the market (e.g., product quality), or it can focus on inputs, meaning procedures, practices, or investments that are expected to improve performance and that the market does not value for their own sake but only insofar as they raise output. While the literature has focused on how a given performance measure should be mapped into ratings, it has largely overlooked how inputs and outputs should be combined into an effective performance measure in the first place. Our contribution is to place this choice at the center by jointly endogenizing the scoring rule that combines inputs and outputs into a performance measure and the rating rule that maps this measure into observable ratings.

Accounting for the scoring design matters both theoretically and in practice. Certification authorities often have broad discretion over what to test, and this choice directly shapes incentives, the informativeness of ratings, and ultimately which equilibria may ultimately arise. ENERGY STAR focuses entirely on outputs by certifying appliances whose measured energy use falls below category specific thresholds. ISO 9001 takes the opposite approach by certifying documented quality management processes such as standardized procedures, training sessions, and internal audits rather than the realized quality of a firm's products. Even organic certifications, rather than relying on output measures, such as random testing of farmers' produce, choose instead to focus on inputs. They verify production practices such as restrictions on synthetic pesticides and land use methods, but they do not directly test properties of the final product.[1] Most systems instead adopt hybrid designs. For example, LEED certification for buildings awards points both for meeting specific performance benchmarks and for adopting verifiable design and construction practices.

This raises a fundamental question: How should a certification system jointly design its scoring and rating rules to best induce the performance valued by the market and the designer? In a setting with heterogeneous agents, our analysis shows that the answer hinges on the designer's concerns about adverse equilibrium outcomes. Although only output is ultimately valued, a certification test should focus solely on inputs if the designer expects the most favorable equilibrium to follow any certification design. In contrast, a designer should combine input and output components if she is concerned about adverse equilibria.

Formally, we consider a model in which agents can invest inputs (such as money or effort) to improve outputs (such as product quality or human capital) at heterogeneous input-to-output conversion rates (productivity types). These outputs are then supplied in a competitive market that values the output itself but cannot observe it directly. Credible communication occurs only through the certification offered by a designer who aims to maximize total output and can commit to a testing technology that specifies (i) the scoring rule, which determines how inputs and outputs are combined into a performance measure, and (ii) the rating rule, which maps this measure into observable ratings. After observing the test design and the ratings received by agents, the competitive market offers payments equal to each agent's conditional expected output.

We start by considering simple pass–fail tests that focus on a single dimension: inputs or outputs. Intuitively, a pure output test is the most informative because it directly evaluates the only outcome the market cares about. However, such a test allows high-productivity agents to "coast on their productivity," passing the test with lower input investment than their lower-productivity peers. This is particularly costly for the designer because input investments by high-productivity agents generate the largest gains in output. A pure input test avoids this problem by inducing the

[1] An organic zucchini grown in Chicago may still be less healthy than a non-organic zucchini produced under better growing conditions in southern Illinois.

same input investment from all passing agents, regardless of their type, thereby eliminating the possibility to scale down investments for high productivity types. For this reason, an input test outperforms any other pass–fail test when the market is sufficiently optimistic about passing agents. Yet without any guarantee on output, pure input tests expose the designer to adverse equilibria in which the market holds more pessimistic beliefs, assigns little value to certification, and discourages agents from investing any input at all.

These observations highlight the strengths and weaknesses of each test dimension. Output-based tests discipline market beliefs but provide only weak incentives for high-productivity agents. Input-based tests offer symmetric incentives for all agents but their effectiveness is severely constrained under designer-adversarial equilibrium selection. How should the test design combine input and output components optimally, so as to provide strong investment incentives while remaining robust to equilibrium selection?

Our main result shows that the optimal pass–fail test combines input and output components in a flexible yet simple manner. The optimal design specifies two output thresholds and one input threshold. An agent passes if she produces output above the higher output threshold or if she satisfies a minimum output requirement and also meets an input requirement. It is worth to note that this characterization holds for general convex cost functions and arbitrary distributions over productivity types not of minor technical assumptions.

The logic is as follows. The minimum output threshold guarantees a floor on the output of passing agents and therefore a floor on the market value of certification. To rule out zero-investment equilibria, the test must ensure that at least the highest productivity type is willing to invest enough to pass even when the benefit from certification is at this lower bound. This requirement imposes an upper bound on the input the highest type can be asked to invest and therefore on the output it can be asked to generate. Since no agent can be asked to meet a higher passing output than the highest type, the designer sets this upper bound as the maximum output threshold, above which agents pass the test directly regardless of their input. Together, these minimum and maximum output thresholds ensure that the test is robust to adversarial equilibrium selection.

To maximize input investment by intermediate types, the designer then introduces an input threshold for agents who do not meet the maximum output threshold. In the unique equilibrium induced by this test, high productivity types pass by reaching the maximum output threshold, earning positive rents because the equilibrium value of certification exceeds the test's minimum output requirement. Intermediate types instead pass by meeting both the minimum output requirement and the input threshold, and are left indifferent between passing and failing. These two paths to passing raise input investment up to the incentive limit while preserving robustness to designer-adversarial equilibrium selection. Finally, note that the minimum input requirement may discourage the lowest productivity types from making any investment.

Beyond the baseline environment, we show that our insights extend along two important dimensions. First, our results are not tied to a specific production technology. The same step-test structure arises under any production function that is increasing in input and weakly supermodular in input and type, including additive specifications. Second, allowing for arbitrarily rich rating schemes does not alter the core logic of optimal test design. With multiple ratings, an optimistic designer continues to rely on pure input tests, while the pessimistic designer must still employ composite tests that combine input and output components to discipline market beliefs and provide high types with sufficient investment incentives.

Taken together, our results highlight that the choice of what to test is central to the design of effective certification systems. Selecting the dimensions of performance that enter the test, rather than only deciding how to map a fixed performance measure into ratings, shapes incentives, determines which equilibria arise, and ultimately governs both the informational content and welfare consequences of certification.

1.1 Literature review

Our paper primarily contributes to the literature on the design of tests, grades, and allocation rules, while sharing features with the signaling literature.

Within the growing literature on allocation design, researchers have examined how a principal optimally designs allocation rules based on an exogenously given output measure that heterogeneous agents may affect (at a cost) in their attempt to obtain the allocated good. While most studies consider settings in which the principal derives no direct benefit from this output investment—treating it merely as a means for agents to misrepresent their type (e.g., Perez-Richet and Skreta (2022, 2023))—we focus on a setting where such investment is productive, and the principal aims to maximize it. In this respect, our paper is related to Augias and Perez-Richet (2023), which establishes a regularity condition on the distribution of agents' types that renders a deterministic pass-fail test optimal. In the context of certification, our paper is related to Xiao (2025), who studies optimal certification design under output testing and also applies optimal control methods to identify the optimal test. Our paper departs from this literature by considering settings where the value of being allocated the good (i.e., passing the test) is endogenous rather than exogenously given and, most importantly, by allowing the principal to choose on which variable to design the allocation rule.

Within the theoretical literature on education, a strand pioneered by Becker and Rosen (1992) and Costrell (1994) analyzes how different grading rules incentivize students' effort and academic achievement. Dubey and Geanakoplos (2010) demonstrates that coarse grading can incentivize status-motivated students to work harder, while Popov and Bernhardt (2013) indicates that increased demand for skilled jobs may lead to grade inflation, and Boleslavsky and Cotton (2015) further argues that the mechanism driving grade inflation may also result in higher investments in school quality when competition rises. Finally, Bizzotto and Vigier (2021) shows that sorting students into schools based on ability (stratification) and using lenient grading at top-tier schools maximizes student effort. We contribute to this literature by endogenizing the mapping from inputs/outputs to scores, thereby enabling the joint design of scoring and grading rules.

Our model also relates to the extensive literature on signaling, in that agents can exert costly efforts to influence their payoffs through outsiders' beliefs (see, for example, Spence, 1978; Daley and Green, 2014; Wolinsky, 1993; Frankel and Kartik, 2019; Ball, 2019; Frankel and Kartik, 2022). However, we differ from this literature by assuming that the costly action is productive rather than wasteful and that a designer maps a combination of inputs and outputs into a public signal about agents.

Finally, our paper relates to the growing literature in mechanism design departing from partial implementation and considering a more cautious designer who is concerned about its minimum equilibrium payoff (see, for example, Ma, 1988; Bergemann and Morris, 2009; Dworczak and Pavan, 2022; Kapon, 2023; Halac, Lipnowski, and Rappoport, 2024; Mishra, Patil, and Pavan, 2025).

2 Model

We consider a model where heterogeneous agents invest costly inputs to improve output quality and can inform the market about their investments and performance only through certification provided by an intermediary. This framework captures settings ranging from firms that invest resources to improve product quality but must rely on certification (e.g., ISO, LEED, ENERGY STAR) to convince the market, to students who invest input to acquire human capital but can credibly demonstrate achievements only through exam results or degrees.

Agents We consider a unit mass of agents whose productivity types are independently and identically distributed according to an atomless, continuously differentiable cumulative distribution function F with density f and support $[\underline{\lambda}, \overline{\lambda}] \subseteq \mathbb{R}_{++}$. Each agent privately knows her productivity type λ. Without loss of generality, we assume that agents with the same type behave identically, so that we can index them by their type λ. An agent of type λ chooses an input $e \in \mathbb{R}_+$ to generate output $\pi(e, \lambda) = b + \lambda \cdot e$, where $b \geq 0$ denotes the baseline output in the absence of investment.[2] The cost $c(e)$ of choosing input $e \geq 0$ is independent of an agent's type with $c : \mathbb{R}_+ \to \mathbb{R}_+$ nonnegative, increasing, and strictly convex, satisfying $c(0) = 0$ and $c'(0) = 0$.

[2] In Section 5.2, we discuss how our results extend to general (weakly) supermodular production functions.

Certification design Agents can credibly communicate with the market only through a certification test offered by a designer whose objective is to maximize the expected output generated by the agents. Formally, a test $T \in \mathcal{T}$ is a right-continuous surjective function $T : \mathbb{R}_+^2 \to G_T$ that maps each displayed input–output pair (e^d, π^d) to a public rating $g = T(e^d, \pi^d) \in G_T$. While agents cannot fabricate evidence, they may withhold it: an agent of type λ who chooses input e may display in the certification test any pair $(e^d, \pi^d) \leq (e, \pi(e, \lambda))$.

Market A perfectly competitive market compensates agents according to their expected output. As ithe market observes neither inputs nor outputs directly,

the market forms expectations in equilibrium solely based on the public certification design and the agents' realized ratings. Thus, given a test design T, the market offers a pay schedule $W_T(g) = \mathbb{E}[\pi \mid T, g]$. We denote the set of feasible payment for a given test design by $\mathcal{W}(T)$.

Timing The game unfolds as follows:

1. The designer publicly announces a certification test T.
2. Each agent of type λ chooses (i) an input e, yielding output $\pi(e, \lambda)$, and (ii) a pair $(e^d, \pi^d) \leq (e, \pi(e, \lambda)$ to display in the certification test.
3. Given the test T and the reported pair (e^d, π^d), each agent receives a rating $g = T(e^d, \pi^d)$.
4. Observing T and each agent's rating g, the market updates its belief about the agent's output π and makes a payment of $W_T(g) = \mathbb{E}[\pi \mid T, g]$.

Agent's problem Given a certification test T and an anticipated payment schedule $W_T : G_T \to \mathbb{R}_+$, each agent λ chooses an input e_λ and a report $(e^d_\lambda, \pi^d_\lambda) \leq (e_\lambda, \pi(e_\lambda, \lambda))$ to maximize her utility $W_T\left(T(e^d_\lambda, \pi^d_\lambda)\right) - c(e_\lambda)$.

For each test T, we define the agent's best-reply correspondence mapping types and wage schedules into actions by $\psi_T : [\underline{\lambda}, \overline{\lambda}] \times \mathcal{W}(T) \rightrightarrows \mathbb{R}^3_+$:

$$\psi_T(\lambda, W) := \underset{e \in \mathbb{R}_+,\ (e^d, \pi^d) \leq (e, \lambda e)}{\arg\max} \; W_T\Big(T(e^d, \pi^d)\Big) - c(e). \tag{1}$$

Solution concept We apply perfect Bayesian equilibrium (PBE), as defined in Fudenberg and Tirole (1991), as our solution concept. The only additional restriction that we impose on off-path beliefs is the following natural requirement on the consistency of market beliefs and the test design. The market's belief about an agent's output given a rating g must not be lower than the minimal output level required for the lowest type, $\underline{\lambda}$, to achieve this rating.[3]

Note that, for many test designs, the game admits multiple equilibria sustained by more or less pessimistic off-path market beliefs. Let $\mathfrak{E}_T$ denote the set of all PBE that may arise in the continuation game following T, and let $\mathcal{E}_T^{\text{sel}} \in \mathfrak{E}_T$ denote the equilibrium selected under a specified equilibrium-selection criterion. In the following, we will focus on the designer's most- and least-preferred equilibria.

Designer The designer selects a test to maximize total output of the agents.[4] For any test design T, let $\mathcal{E}_T^{\text{sel}}$ denote the equilibrium anticipated by the designer in the continuation game following T, and let $e_\lambda^{\text{sel}}(T) \in \mathbb{R}_+$ denote the input chosen by type λ in $\mathcal{E}_T^{\text{sel}}$. The designer's problem is

$$\max_{T \in \mathcal{T}} \; \mathbb{E}\big[\lambda \, e_\lambda^{\text{sel}}(T)\big] \, . \tag{2}$$

Naturally, the optimal test design depends on whether the designer *expects* her preferred equilibrium to arise in the continuation game or whether she is instead concerned with adverse equilibria.

[3] Our results remain qualitative unchanged if we were to restrict off-path beliefs further to be constructed only from undominated actions.

[4] Note that we assume that the designer is not a "gatekeeper;" that is, the certification is not necessary to participate in the market but only a tool to communicate output to the market. Hence, the designer cares about total output rather than output conditional on obtaining certification.

While the existence of an optimal test is straightforward in the optimistic case, the pessimistic scenario is technically more subtle. Typically, no test T^* attains the designer's worst-case payoff; there exists a sequence of tests $(T_n)_{n\in\mathbb{N}}$ yielding progressively higher worst-case payoffs, with a limit test that admits, in addition to the limit of the associated equilibrium strategies sequence, another, worse, equilibrium. To address this issue and to avoid trivial nonexistence/multiplicity from agent's indifference, we follow Halac (2025) and impose a tie-breaking rule that breaks agents' indifference in favor of higher inputs.[5]

Definition 1 (Robust PBE). *Given a test T, a PBE $\mathcal{E}_T$ is* robust *(an rPBE) if there exists no other PBE $\tilde{\mathcal{E}}_T$ such that the following hold.*

(i) ***Lower payoff.*** *The designer obtains a strictly lower payoff in $\tilde{\mathcal{E}}_T$.*

(ii) ***Favorable tie-breaking.*** *Under the pay schedule $\tilde{W}_T$ in $\tilde{\mathcal{E}}_T$, an agent who is indifferent between two input levels $e' > e$, selects e' in $\tilde{\mathcal{E}}_T$.*

We then distinguish designers by their attitude toward adverse equilibria. An **optimistic designer** chooses T to maximize her payoff in her most-preferred PBE following T. A **pessimistic designer** chooses T to maximize her payoff in the rPBE following T.

2.1 Discussion

Focus on output. In our model, the market care only about output (e.g., product quality), not about inputs or productivity types per se. This captures settings in which output represents the final good traded in the market: consumers value the quality of the product itself, not the investments it required or whether the same firm may produce better products in the future.[6]

[5] Alternatively, one could work with ϵ-equilibria of the full game. If T^* is the test selected in the pessimistic-designer case (according to our analysis), then the equilibrium outcomes we describe can be arbitrarily approximated by a sequence of least-preferred equilibrium outcomes corresponding to a sequence of tests $T_n \to T^*$ as $n \to \infty$.

[6] In a human capital interpretation, output can represent next period's productivity, so heterogeneity reflects accumulated stocks rather than fixed traits. The analysis then interprets certification as incentivizing investment in future productivity rather than signaling a fixed type.

Similarly, the designer cares only about output, not about the cost of producing it. This assumption allows us to capture settings where certification design targets specific goals (e.g., environmental impact, safety, or reliability) whose effects may not be fully internalized by the market. It also allows us to highlight, in the cleanest possible way, the forces and dynamics that would still be present if the designer placed a relatively greater value on output quality than on cost. For instance, higher outputs produce positive externalities (such as human capital or more environmentally friendly products) or input costs may represent transfers to complementary sectors (such as wages for experts who train employees or purchases from upstream quality suppliers) and may themselves be valuable from the designer's perspective.

Production function. While, in line with the literature, our baseline model focuses on the simple production function $\pi(e, \lambda) = \lambda e$, our main insights extend to production functions that are increasing in input and weakly supermodular in type and input (see Section 5.2). This extension is useful for settings in which baseline quality varies across types even without certification, such as $\pi = \lambda + e$.

Joint design. To highlight our key innovation, it is useful to equivalently reformulate the test design problem as one where the designer jointly chooses a *scoring rule*, $s : \mathbb{R}_+^2 \to \mathcal{S}$, that combines reported input–output pairs into a score, and a *rating rule*, $g : \mathcal{S} \to G$, that maps scores into ratings, with $T = g \circ s$. The scoring rule determines *what is tested*; that is, the relative weight placed on inputs and outputs. For instance, a certification test can focus entirely on inputs (e.g., $s(e^d, \pi^d) = e^d$), entirely on outputs (e.g., $s(e^d, \pi^d) = \pi^d$), or combine the two, either linearly (e.g., $s(e^d, \pi^d) = \alpha\pi^d + (1 - \alpha)e^d$) or nonlinearly (e.g., $s(e^d, \pi^d) = \mathbb{I}\{e^d \geq \bar{e}\} + \mathbb{I}\{\pi^d \geq \bar{\pi}\}$). On the other hand, the rating rule determines how finely the score is communicated to the market: it may reveal scores exactly or group multiple scores into the same observable rating.

Previous work has typically treated the test variable, that is, the choice of whether certification evaluates inputs or outputs, as exogenous, optimizing only over the rating rule given the test variable. Our framework endogenizes both: the certifier simultaneously decides *what to test* and *how to rate*.

Underreporting and test monotonicity. We assume that agents may underreport their input–output pairs. From an applied perspective, this captures a realistic feature of certification environments: firms can easily conceal part of their input (e.g., by not documenting some training sessions or omitting certifications that attest to the use of low-impact materials in construction) or underreport, and equivalently derating, their realized output (e.g., by presenting a slightly defective testing unit to the certifier or requiring call center agents to pause before answering). Whether agents actually underreport is an equilibrium outcome. Although no agent underreports in equilibrium, allowing it is nevertheless important because it constrains which tests are implementable. In particular, underreporting rules out certification designs that reward agents for underperforming relative to another agent.

Formally, the underreporting technology delivers a monotonicity property that we view as central to certification: if an agent reports weakly higher input and output than another agent, she cannot receive a strictly worse rating. This *test monotonicity* is the only implication of underreporting that we use in our analysis.

3 Analysis

Before introducing the designer's optimization problem, it is useful to establish two preliminary results. Without loss, we normalize the baseline output to $b = 0$.

Monotonicity. For any test T and pay schedule W_T, each agent λ chooses an input e_λ and a report $(e^d_\lambda, \pi^d_\lambda) \leq (e_\lambda, \lambda e_\lambda)$ to maximize $W_T\big(T(e^d_\lambda, \pi^d_\lambda)\big) - c(e_\lambda)$. Since $c(e)$ is strictly increasing, any agent achieving rating g in test T optimally chooses the smallest input required to obtain g. Denote this minimal input by

$$e^{T,g}_\lambda := \min\Big\{e \in \mathbb{R}_+ : \exists (e^d, \pi^d) \leq (e, \lambda e) \text{ such that } T(e^d, \pi^d) = g\Big\}. \tag{3}$$

Because agents may underreport their input and output levels, any rating g is attainable for all agents. Thus, the minimal input $e_{\lambda}^{T,g}$ is well-defined. As the output is an increasing function of input, the minimal input decreases in λ. However, the output associated with this minimal input, $\lambda e_{\lambda}^{T,g}$, is increasing in λ. Hence, higher-productivity agents can achieve the same rating with weakly less input, yet attain a higher output once they do. This monotonicity property implies that, from the designer's perspective, any test T induces a *downward-sloping* mapping between agents' types and their required minimal input, which we establish in the following lemma.

Lemma 1. *For any test design T and rating $g \in G_T$, the minimal input required by types λ, λ' with $\lambda > \lambda'$ required to achieve rating g satisfies*

$$\frac{e_{\lambda}^{T,g}}{e_{\lambda'}^{T,g}} \in \left[\frac{\lambda'}{\lambda}, 1\right].$$

Proof. Suppose $\frac{e_{\lambda}^{T,g}}{e_{\lambda'}^{T,g}} < \frac{\lambda'}{\lambda}$. Then, it follows that $\pi(e_{\lambda}^{T,g}, \lambda) < \pi(e_{\lambda'}^{T,g}, \lambda')$ and $e_{\lambda}^{T,g} < e_{\lambda'}^{T,g}$. Hence, $e_{\lambda'}^{T,g}$ cannot be the minimal input attaining rating g for type λ'.

Suppose $\frac{e_{\lambda}^{T,g}}{e_{\lambda'}^{T,g}} > 1$. Then, it follows that $\pi(e_{\lambda}^{T,g}, \lambda) > \pi(e_{\lambda'}^{T,g}, \lambda')$ and $e_{\lambda}^{T,g} > e_{\lambda'}^{T,g}$. Hence, $e_{\lambda}^{T,g}$ cannot be the minimal input attaining rating g for type λ. □

Equilibrium payments. Since any agent λ attaining rating g in equilibrium does so by exerting $e_{\lambda}^{T,g}$, it follows that, for any equilibrium $\mathcal{E}_T$ and any rating $g \in \mathcal{G}_T$, the corresponding equilibrium payment W must satisfy

$$W(g) = \mathbb{E}\Big[\lambda\, e_{\lambda}^{T,g} \;\mid\; \lambda \in \Lambda^{g}(\mathcal{E}_T)\Big],$$

where $\Lambda^{g}(\mathcal{E}_T)$ denotes the set of types that obtain rating g in the $\mathcal{E}_T$.

4 Pass-Fail Tests

In this section, we restrict attention to pass-fail tests; that is, tests such that $T : \mathbb{R}_{+}^{2} \to \{0, 1\}$, where $g = 0$ represents failing the test and $g = 1$ represents passing the test. In the context of certification, this implies that the choice is only on the extensive margin, that is, whether an agent is awarded with a certificate or not. This focus is motivated by two considerations. First, pass-fail tests are especially common in practice due to their simplicity and transparency, and thus provide a natural and policy-relevant benchmark. Second, as we show in Section 5.1, the main insights from our analysis extend to environments with multi-rating tests, but the pass-fail formulation makes the key mechanisms particularly transparent and allows for a sharp characterization of the optimal design.

To illustrate the central tradeoffs, we first examine simple tests that assign ratings based on a single test variable (input or output), and then show how the optimal pass-fail test improves upon these benchmarks.

4.1 Output Tests

First, consider pass-fail output tests, where agents pass if and only if their output exceeds $\overline{\pi}$, regardless of their input choice. Formally, $T(e, \pi) = \mathbb{I}[\pi \geq \overline{\pi}]$. Since the market values only output, and neither input nor agents' types per se, these tests provide a natural benchmark.

Since inputs are costly, failing agents exert no input and produce only their baseline output of zero. Agents who choose to pass the test and obtain the certificate invest so that they attain exactly the required threshold $\overline{\pi}$. This uniquely determines the payment schedule $W_{\mathcal{E}^T}$ in any equilibrium $\mathcal{E}^T$ of a pure output test: Conditional on passing, agents receive a payment $W_{\mathcal{E}^T}(1, T) = \overline{\pi}$, whereas, conditional on failing, agents earn $W_{\mathcal{E}^T}(0, T) = 0$.

Although all agents face the same input cost function and payment schedule, their payoff from passing the test still depends on their productivity type λ. Since lower types require higher input to reach the threshold $\overline{\pi}$, $e_{\lambda}^{T,1} = \overline{\pi}/\lambda$, the payoff from passing an output test is strictly increasing in the agent's type. Hence, any equilibrium of the subgame induced by $T(e, \pi) = \mathbb{I}[\pi \geq \overline{\pi}]$ has a cutoff structure. In each equilibrium $\mathcal{E}^T$, there exists a cutoff type $\tilde{\lambda}(\overline{\pi})$ such that all agents with $\lambda \geq \tilde{\lambda}(\overline{\pi})$ pass and all others fail. Since this cutoff type must be indifferent between passing

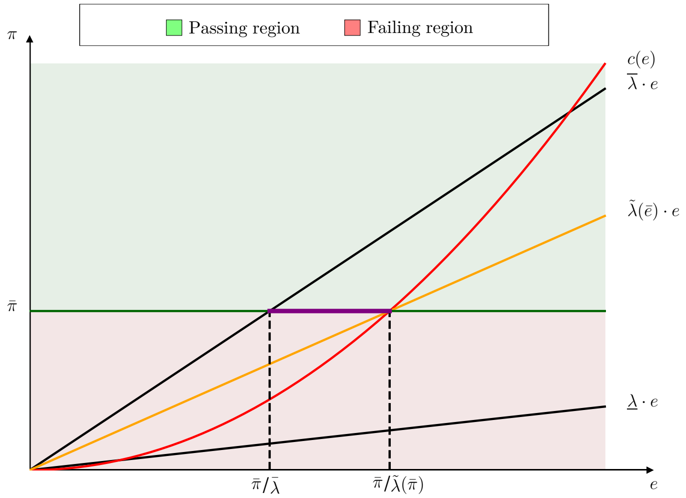
🟩 Passing region
🟥 Failing region
π
$\overline{\pi}$
$c(e)$
$\overline{\lambda} \cdot e$
$\tilde{\lambda}(\bar{e}) \cdot e$
$\underline{\lambda} \cdot e$
$\overline{\pi}/\overline{\lambda}$
$\overline{\pi}/\tilde{\lambda}(\overline{\pi})$
e

Figure 1: Output test with threshold $\overline{\pi}$. *The black diagonal lines show the output of the highest and lowest types at each input e; the yellow line shows the output of the cutoff type $\tilde{\lambda}(\overline{\pi})$, for whom the cost of passing equals $\overline{\pi}$. Finally, the purple line highlights all positive input–output combinations in equilibrium.*

and failing the test, we have $\tilde{\lambda}(\overline{\pi}) = \frac{\overline{\pi}}{c^{-1}(\overline{\pi})}$. Figure 1 illustrates the equilibrium construction in such an output test.

Given this continuation equilibrium $\mathcal{E}^T$, the designer's problem reduces to choosing the threshold $\overline{\pi}$ that maximizes total output. The designer faces a typical marginal-inframarginal tradeoff. Raising $\overline{\pi}$ increases the output of all inframarginal (passing) agents but comes at a marginal loss from shrinking the set of passing agents. The optimal output threshold balances these two effects.

Proposition 1. *The optimal output test $T^*(e,\pi) = \mathbb{I}[\pi \geq \overline{\pi}^*]$ is characterized by the unique threshold $\overline{\pi}^* \geq 0$ solving $\max\limits_{\overline{\pi} \geq 0} \overline{\pi} \left(1 - F\left(\frac{\overline{\pi}}{c^{-1}(\overline{\pi})}\right)\right)$.*
In the unique equilibrium $\mathcal{E}^{T^*}$, the wage schedule is $W_{\mathcal{E}^{T^*}}(g, T^*) = \overline{\pi}^* \mathbb{I}\{g = 1\}$, and agents invest $e_\lambda^{T^*} = \overline{\pi}^*/\lambda$ if $\lambda \geq \frac{\overline{\pi}^*}{c^{-1}(\overline{\pi}^*)}$ and $e_\lambda^{T^*} = 0$ otherwise.

Intuitively, an output test reveals the exact value of agents to the market, but it fails to incentivize higher inputs from high-productivity types, allowing them to *coast* *on their talent*: they pass the test while investing strictly less input and obtaining a strictly higher payoff than the cutoff type. Thus, the agents with the highest returns to input investments end up the least of those who invest at all.

4.2 Input Tests

Next, consider pass–fail input tests $T(e, \pi) = \mathbb{I}[e \geq \overline{e}]$, where agents pass if and only if their input exceeds a threshold $\overline{e} \geq 0$, regardless of the output they produce. These tests do not directly provide the market with information about agents' outputs. Nevertheless, they contain information about their expected about through the equilibrium beliefs about which agents passed the test. While input tests convey information only about agents' input, they prevent high-productivity types from coasting on their talent. All agents must invest the same input to pass. Moreover, because input costs are type independent, all agents face the same incentives.

Under a pure input test, failing agents optimally invest $e_\lambda^{T,0} = 0$, producing no output. Each passing agent λ invests $e_\lambda^{T,1} = \overline{e}$, producing output $\pi(e, \lambda) = \lambda\overline{e}$. Thus, while the payment to agents who fail the test remains $W(0, T) = 0$, the payment to agents who pass the test now depends on market beliefs about the productivity of passing agents:

$$W(1, T) = \overline{e}\, \mathbb{E}_{\mathcal{E}^T}[\lambda \mid g = 1].$$

Hence, whether agents are willing to pass the test or not depends on the markets' beliefs as well. Whenever these beliefs are such that $W(1, T) > c(\overline{e})$, all agents are (strictly) willing to pass the test. In contrast, when the beliefs are such that $W(1, T) < c(\overline{e})$ all agents have a strict incentive to fail the test. Whenever the beliefs induce a wage equal to the cost of investing the required input $\overline{e}$, that is, when $W(1, T) = c(\overline{e})$, all agents are indifferent between passing and failing the test. For many input test designs, we will obtain equilibrium multiplicity generated by the market beliefs. Hence, unlike in output testing, the optimal input test depends on the designer's attitude toward adverse equilibria.

Optimistic designer. An optimistic designer assumes that the market and agents coordinate on its preferred PBE following any input test $T(e, \pi) = \mathbb{I}[e \geq \overline{e}]$. Recall that the designer's objective is to maximize total output. Thus, the designer will never choose an input threshold such that no agent is willing to pass the test even under the most optimistic market belief about passing agents' types, that is, the optimal input threshold must be such that $c(\overline{e}) < \pi(\overline{e}, \overline{\lambda})$. Moreover, the designer will never choose an input threshold such that all agents have a strict incentive to pass the test, that is, the optimal input threshold must be such that $c(\overline{e}) > \mathbb{E}_F[\lambda]\overline{e}$. To see why, note that in this case all agents passing is the unique equilibrium outcome. However, there by slightly raising the the passing threshold all agents still have a strict incentive to pass the test but invest more input, and thus, the total output is higher. Defining $\hat{e}$ implicitly as the solution to $c(\check{e})/\check{e} = \overline{\lambda}$ and $\hat{e}$ implicitly as the solution to $c(\hat{e})/\hat{e} = \mathbb{E}_F[\lambda]$, the set of candidate thresholds for the optimal input tests is $[\check{e}, \hat{e}]$.[7]

It is immediate that for all input thresholds in the candidate set there always exists a PBE with a cutoff structure, that is with the property that only agents with types above a threshold $\tilde{\lambda}(\overline{e})$ pass the test. The following lemma establishes that the cutoff equilibrium is always the designer's preferred equilibrium.

Lemma 2. *Consider an input test $T = \mathbb{I}[e \geq \overline{e}]$ with $\overline{e} \in [\check{e}, \hat{e}]$. Then, the designer's preferred PBE following T is a cutoff equilibrium, that is, there exists a type $\tilde{\lambda}$ such that agents with $\lambda \geq \tilde{\lambda}$ pass the test by choosing $\overline{e}$ and agents with $\lambda < \tilde{\lambda}$ fail the test by choosing $e = 0$.*

Hence, the designer's problem amounts to choosing the best cutoff equilibrium. Denote by $\tilde{\lambda}(\overline{e})$ the cutoff type corresponding to an input threshold $\overline{e}$. The optimistic designer's problem is therefore $\max_{\overline{e} \in [\check{e}, \hat{e}]} \; \overline{e}\left(1 - F(\tilde{\lambda}(\overline{e}))\right)$.

Pessimistic designer. Unlike its optimistic counterpart, a pessimistic designer is concerned with adverse equilibrium outcomes. Thus, a pessimistic designer will

[7] Note that $c(e)/e$ is strictly increasing in e with $\lim_{e \to 0} c(e)/e = 0$ due to the strict convexity of $c(e)$ and $c(0) = c'(0) = 0$. Hence, $\check{e}$ and $\hat{e}$ are well-defined, unique, and satisfy $\check{e} < \hat{e}$.

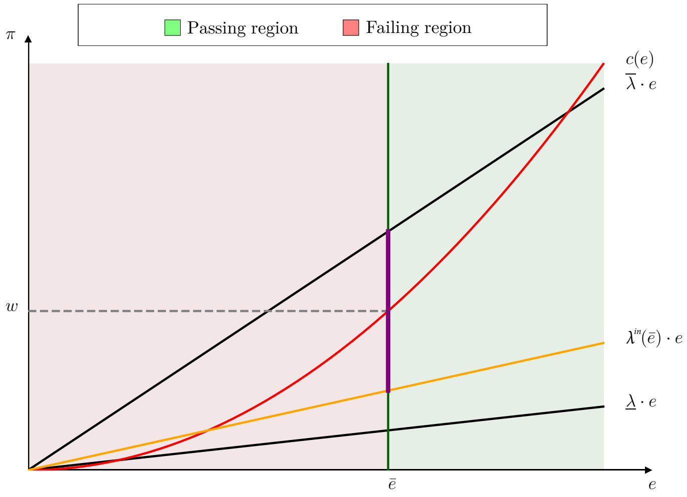

Figure 2: Input test with threshold $\overline{e}$: designer-preferred equilibrium. *While all types are indifferent between passing and failing at $w = \mathbb{E}[\lambda \mid \lambda \geq \lambda^{in}(\overline{e})]\,\overline{e} = c(\overline{e})$, only those with $\lambda > \tilde{\lambda}$ choose to pass. The purple line thus highlights all positive input–output combinations reached in equilibrium.*

never devise a test which is such that an equilibrium exists in which all agents fail. This places an upper bound on the input threshold that the pessimistic designer will optimally choose. The most pessimistic belief of the market about passing agents' productivity is that only the lowest type $\underline{\lambda}$ passes. Hence, an equilibrium in which all agents fail exists whenever $c(\overline{e}) > \underline{\lambda}\overline{e}$.[8] The optimal input threshold for the test designer is therefore to choose the maximum input threshold for which no equilibrium exists in which all agents fail the test, that is, $\overline{e} = \check{e}$. All agents pass this test and total output is $\mathbb{E}[\lambda]\check{e}$.

Comparison A pessimistic designer ensures that all agents have an incentive to pass the test in order to prevent adverse equilibria. This robustness does not come for free to the designer; she has to sacrifice some output production to achieve it.

[8] Recall that by our tie-breaking assumption in Definition 1, $c(\overline{e}) = \underline{\lambda}\overline{e}$ induces the rPBE in which all agents pass the test.

The following proposition summarizes our observations over the optimal input test.

Proposition 2. *The optimal input test has the following properties.*

- *An **optimistic designer** selects* $\overline{e}^o = \arg\max_{\overline{e}\in[\check{e},\hat{e}]} \overline{e}\,(1 - F(\tilde{\lambda}(\overline{e})))$ *and all agents above the cutoff type* $\tilde{\lambda}(\overline{e}^o$ *pass the test. Conditional on passing the test, the agents receive the payment* $W_{\mathcal{E}^{T^o}}(1, T^*) = \mathbb{E}[\lambda \mid \lambda > \tilde{\lambda}(\overline{e}^o)]\overline{e}^o = c(\overline{e}^o)$*.*
  In any rPBE following the input test with $\overline{e}^o$*, all agents fail the test.*

- *A **pessimistic designer** selects* $\overline{e}^p$ *such that* $\underline{\lambda}\,\overline{e}^p = c(\overline{e}^p)$ *and all agents pass the test in the rPBE. Conditional on passing the test, the agents receive the payment* $W_{\mathcal{E}^{T^p}}(g, T^p) = \mathbb{E}[\lambda] = c(\overline{e}^p)$*.*

4.3 Comparing Input and Output Tests

For the pessimistic designer who aims at incentivizing output production, the key tradeoff is between mitigating the risk of adverse equilibria and curbing high-productivity types' coasting. Pure output tests certify the exact value of agents to the market removing all uncertainty about their outputs, but they allow high-productivity agents to pass with minimal input investments relative to their lower-productivity peers. This inefficiency is particularly costly since the designer values the inputs from high-productivity agents relatively more. In contrast, pure input tests induce uniform inputs from all agents and can incentivize high input levels even from productive agents. However, unless the threshold is very low ($\overline{e} \leq \hat{e}$), they expose the designer to adverse equilibria: because they do not guarantee a sufficient market value for passing agents, a pessimistic market may offer low payments and discourage agents from passing. Thus, the optimal test design ultimately depends on the designer's attitude toward equilibrium uncertainty.

Optimistic designer. An optimistic designer is unconcerned about potential adverse equilibria and, as a result, strictly prefers input tests. To see this, consider an optimal output test $T(e, \pi) = \mathbb{I}[\pi \geq \overline{\pi}]$, and let $\tilde{\lambda} < \overline{\lambda}$ be the cutoff type who is indifferent between passing and failing the output test in equilibrium. As lower types

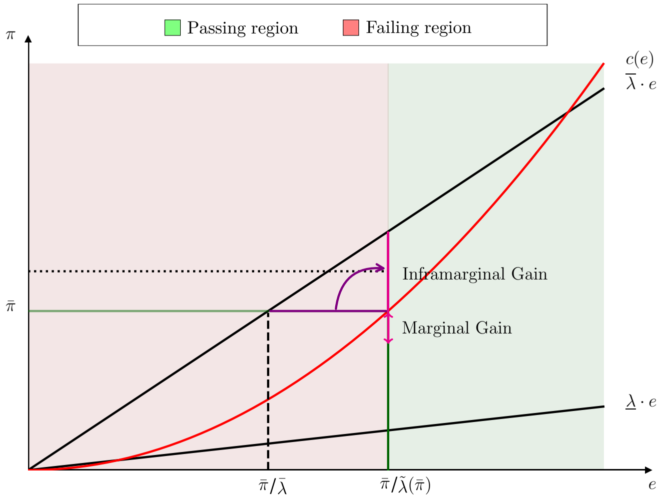
🟩 Passing region
🟥 Failing region
π
$\bar{\pi}$
$c(e)$
$\overline{\lambda} \cdot e$
Inframarginal Gain
Marginal Gain
$\underline{\lambda} \cdot e$
$\bar{\pi}/\bar{\lambda}$
$\bar{\pi}/\tilde{\lambda}(\bar{\pi})$
e

Figure 3: Improving on an output test in designer-preferred equilibrium. *Replacing an output test with threshold $\overline{\pi}$ by an input test with $\overline{e} = \overline{\pi}/\tilde{\lambda}$ yields both inframarginal gains (existing passers move from the purple horizontal line to the magenta vertical line) and marginal gains (additional types pass, shown by the vertical double arrow).*

must invest more input to reach the output threshold, this cutoff type must be the agent investing the most, $e_{\tilde{\lambda}}^{T,1} = \overline{\pi}/\tilde{\lambda}$, and obtains payoff $\overline{\pi} - c(\overline{\pi}/\tilde{\lambda}) = 0$. It turns out that there always exists a pure input test, with the corresponding designer-preferred equilibrium, that outperforms such an optimal output test. Consider the input test $T(e,\pi) = \mathbb{I}[e \geq \overline{e}]$ with threshold $\overline{e} = \overline{\pi}/\tilde{\lambda}$. Since

$$\mathbb{E}[\lambda \mid \lambda \geq \tilde{\lambda}]\,\overline{e} - c(\overline{e}) \;>\; \overline{\pi} - c\left(\frac{\overline{\pi}}{\tilde{\lambda}}\right) \;=\; 0,$$

it follows that all types above some new cutoff $\lambda' < \tilde{\lambda}$ invest $\overline{e}$ and pass the input test. The designer gains both on the extensive and on the intensive margin: (i) agents with $\lambda \in [\lambda', \tilde{\lambda})$, who failed the output test, now pass the input test;and (ii) agents with $\lambda > \tilde{\lambda}$, who already passed the output test, also pass the input test but now invest $\overline{e} > \overline{\pi}/\lambda$.

Pessimistic designer. A pessimistic designer, by contrast, is concerned with adverse equilibria and may avoid pure input tests despite their superiority in incentive-provision. The reason being that input tests are prone to equilibrium multiplicity and thus to the existence of low-investment equilibria. Indeed, the maximal input that can be robustly induced in equilibrium by an input test is $\check{e}$, the unique solution to $\underline{\lambda}\, e = c(e)$, which converges to 0 as $\underline{\lambda} \to 0$. Thus, while optimism always favors input tests, pessimism often favors output tests.

4.4 Optimal Pass-Fail Test

In this section, we derive the optimal pass-fail test for both an optimistic and a pessimistic designer.

Optimistic Designer

We first observe that an optimistic designer cannot gain from going beyond pure input tests. Thus, a simple test based on a single input threshold is sufficient to attain the maximal payoff for the designer. In light of the previous results, this result is intuitive. Input tests are efficient at providing investment incentives. An optimistic designer is not concerned with the existence of adverse equilibria. Thus, input tests are the ideal tool for optimistic designers. Nevertheless, it is worth noting that such an optimal test does not directly involve any testing of output, that is, the outcome that the designer and the market care about.

Theorem 1. *The optimal pass-fail test for an* ***optimistic designer*** *is* $T^o = \mathbb{I}[e \geq \overline{e}^o]$*, where* $\overline{e}^o = \arg\max_{\overline{e} \in [\check{e}, \hat{e}]} \overline{e}\,(1 - F(\tilde{\lambda}(\overline{e})))$*, with* $\tilde{\lambda}(\overline{e})$ *defined by* $\overline{e} \int_{\tilde{\lambda}(\overline{e})}^{\overline{\lambda}} \lambda\, dF(\lambda) = c(\overline{e})$*. In the designer-preferred PBE following this test, all agents above type* $\tilde{\lambda}(\overline{e})$ *pass the test by investing the threshold input and the equilibrium payment satisfies* $W_{\mathcal{E}^{T^o}}(1, T^*) = \mathbb{E}[\lambda \mid \lambda > \tilde{\lambda}(\overline{e}^o)] = c(\overline{e}^*)$.

In addition, the test T^o *features an rPBE in which all agents fail the test.*

Intuitively, in the optimistic designer scenario, equilibrium selection eliminates the need to test output in order to secure favorable market beliefs and thus favorable payoffs for passing agents. The designer therefore prefers a pure input test, which provides the strongest investment incentives across all types by preventing high types from coasting on their talent. In particular, following the same steps as in Section 4.3, we can show that an optimistic designer always benefits from switching from an arbitrary test T to a pure input test, which requires a passing input equal to that exerted by the lowest type passing type in under T in the designer's preferred equilibrium.

Pessimistic Designer

In stark contrast to the optimistic case, a pure input test is never optimal for a pessimistic designer. Recall from Section 4.2 that the pessimistic designer's best input test is $T^p = \mathbb{I}[e \geq \check{e}]$, where $\check{e}$ is implicitly defined from the equation $\underline{\lambda}\check{e} = c(\check{e})$. In the rPBE induced by T^p, all agents choose input level $\check{e}$, the market offers the passing payment $W(1,T) = \check{\pi} := \mathbb{E}[\lambda]\check{e}$, and the designer's payoff is $\check{\pi}$. We argue that a pessimistic designer can obtain a higher rPBE payoff by combining the relative strengths of the two testing variables. By adopting an *L-test*—that is, a test that combines one input and one output threshold and requires agents to meet both thresholds to pass the test—she protects herself against adversarial equilibria using the output threshold, but provides strong investment incentives using the input threshold.

Consider, for instance, $T'(e,\pi) := \mathbb{I}[e \geq \check{e}] \cdot \mathbb{I}[\pi \geq \tilde{\pi}]$, where $\tilde{\pi}$ is the unique solution to $c(\tilde{\pi}/\underline{\lambda}) = \check{\pi}$. Passing the test T' requires more output than passing T (as $\tilde{\pi}$ is constructed through the lowest types' productivity). Thus, the induced rPBE $\mathcal{E}^{T'}$ must yield a higher passing payment as well, $W(1,T') \geq W(1,T) = \check{\pi}$. Moreover, by construction of $\tilde{\pi}$, even the lowest type $\underline{\lambda}$ prefers passing T' to failing whenever $W(1,T') \geq \check{\pi}$. Thus, $\mathcal{E}^{T'}$ necessarily features all agents passing T' and there does not exist a worse equilibrium for the principal. Finally, because $c(\check{e}) = \underline{\lambda}\check{e} < \mathbb{E}[\lambda]\check{e} = \check{\pi} = c(\tilde{\pi}/\underline{\lambda})$, the input required from the lowest type to reach the output threshold $\tilde{\pi}$ is strictly higher than $\check{e}$. Consequently, even a pessimistic designer benefits from switching from T to T': T' induces weakly higher output from all types and strictly higher output from lower types.

Also the optimal pure-output test $T^o = \mathbb{I}[\pi \geq \overline{\pi}]$ from Proposition 1 can be improved upon through an L-test. Let $\tilde{\lambda}$ denote the lowest passing type in the rPBE induced by T^o, and consider $T'(e, \pi) = \mathbb{I}[e \geq \overline{\pi}/\tilde{\lambda}] \cdot \mathbb{I}[\pi \geq \overline{\pi}]$. Since the passing wage under T' is at least $\overline{\pi} = c(\overline{\pi}/\tilde{\lambda})$, all types $\lambda \geq \tilde{\lambda}$ pass in the induced rPBE. Moreover, because $\overline{\pi}/\lambda$ decreases in λ, the new input threshold binds for all types above $\tilde{\lambda}$, requiring them to raise their input from $e_\lambda = \overline{\pi}/\lambda$ under T to $e = \overline{\pi}/\tilde{\lambda} > \overline{\pi}/\lambda$. Hence, switching to T' strictly benefits the designer.

In general, however, even L-tests are suboptimal. Our main result provides the characterization of the optimal pass-fail test. In particular, we show that the optimal test is a *step*-test $T(e, \pi) = \mathbb{I}[\pi \geq \overline{\pi}_H] + \mathbb{I}[e \geq \overline{e}]\, \mathbb{I}[\overline{\pi}_H > \pi \geq \overline{\pi}_L]$, i.e., a test combining one input threshold $\overline{e}$ and two output thresholds $\overline{\pi}_H > \overline{\pi}_L > 0$, where agents pass either by meeting the high output bar $\overline{\pi}_H$ or by meeting the lower output standard $\overline{\pi}_L$ while investing at least $\overline{e}$.

The optimal pass-fail test highlights how a pessimistic designer optimally combines the two strengths of input and output tests. The output components are devised to effectively prevent the existence of adverse equilibria. The input threshold then provides maximal incentives for input investments within the limits that the output components allow.

Theorem 2. *The optimal pass–fail test for a pessimistic designer is a step-test*

$$T(e, \pi) = \mathbb{I}[\pi \geq \overline{\pi}_H] + \mathbb{I}[e \geq \overline{e}]\, \mathbb{I}[\overline{\pi}_H > \pi \geq \overline{\pi}_L],$$

that induces all agents with $\lambda \geq \tilde{\lambda}_T \in [\underline{\lambda}, \overline{\lambda})$ to pass in the induced rPBE. Moreover, the optimal thresholds satisfy $\overline{\lambda}\, \overline{e} > \overline{\pi}_H > \overline{\pi}_L$ and, in the induced rPBE, yield $\tilde{\lambda}_T e^{T,1}_{\tilde{\lambda}_T} = c(e^{T,1}_{\overline{\lambda}})$.

To illustrate why a step-test is optimal, we proceed in several steps. Assume that test T is optimal and let $\tilde{\lambda}_T$ denote the lowest passing type in the induced rPBE. First, observe that requiring an output threshold $\pi_L = \tilde{\lambda}_T e^{T,1}_{\tilde{\lambda}_T}$ cannot reduce the designer's rPBE payoff. Hence, we can assume that the optimal test T imposes this minimal output threshold and that, in the induced rPBE, the lowest passing type attains exactly this output level. This threshold places a lower bound on market beliefs, and thus on the (on- or off-path) passing payment, in particular, we obtain $W(T, 1) \geq \pi_L$. Note that introducing this threshold makes it weakly harder for types below the cutoff to pass, and hence, does not affect participation. Types above the cutoff continue passing the test in the same way. Therefore, the new output threshold has the sole purpose of disciplining the most pessimistic market beliefs if passing the test is an off-path event. Thus, a test with the additional output threshold π_L is less prone to adverse equilibria than any other test that induces the same output level from the cutoff type $\tilde{\lambda}_T$. We summarize this insight in the following lemma.

Lemma 3. *Consider two pass-fail tests, $\tilde{T}$ and $T = \tilde{T}\,\mathbb{I}[\pi \geq \pi_L]$, and denote by $\mathcal{E}^{\tilde{T}}$ and $\mathcal{E}^{T}$ their respective rPBEs. If π_L is the output produced by the lowest passing type $\tilde{\lambda}_{\tilde{T}}$ in $\mathcal{E}^{\tilde{T}}$, then the designer's payoff and the agents' choices and payments are identical in $\mathcal{E}^{\tilde{T}}$ and $\mathcal{E}^{T}$.*

Second, consider an arbitrary test T' imposing the same minimal output requirement π_L as the test T. If T' requires from the high type a passing input level $e^{T',1}_{\overline{\lambda}}$ such that the associated cost exceeds the minimum output threshold (that is, if $c(e^{T',1}_{\overline{\lambda}}) > \pi_L$), then T' features an all-agents-fail rPBE sustained by the most pessimistic off-path passing payment $W(T, 1) = \pi_L$. Thus, such a test T' is suboptimal.

Suppose instead that $c(e^{T',1}_{\overline{\lambda}}) < \pi_L$, implying that the highest type is always willing to pass the test, even under the most pessimistic market belief. By Lemma 1, the required passing input is decreasing in agents' productivity, and therefore, the payoffs are increasing in the agents' productivity as well. Therefore, the fact that the highest type's costs are below the input threshold π_L implies that $e^{T',1}_{\overline{\lambda}} < e^{T,1}_{\overline{\lambda}}$. Further, recall that in any test, the underreporting incentive constraint requires that the highest type attains an output level of at least π_L, implying that $e^{T',1}_{\overline{\lambda}} \geq \pi_L/\overline{\lambda}$. However, the designer can improve on the test T' by raising the implied required passing investment of the highest type marginally while still ensuring incentive compatibility.[9] With this modified test, the designer realizes an inframarginal gain, due

[9] To see that this is feasible, note that, due to incentive compatibility, there exists a type $\lambda^0 = \overline{\lambda} - \varepsilon$

to higher required inputs to pass the test, and a marginal gain, due to a positive wage response to the higher investments of high types.

It follows from this reasoning that that an optimal pass–fail test T must be devised such that the input investment required by the highest type, $e^{T,1}_{\overline{\lambda}}$, comes at a cost equal to the lower input threshold, which coincides with the output produced by the cutoff type.[10]

Lemma 4. *Any optimal pass-fail test T for a pessimistic designer must satisfy*

$$c(e^{T,1}_{\overline{\lambda}}) = \tilde{\lambda}_T e^{T,1}_{\tilde{\lambda}_T}. \tag{4}$$

The previous to lemmata pin down two constraints that hold in an optimal pass-fail test for the pessimistic designer; that is, (i) there is an output threshold π_L equal to the output attained by the indifferent agent, and (ii) the input required from the highest type is such that its associated cost is equal to the output threshold π_L.

To find the set of candidate optimal tests, it suffices to pick an input threshold π_L, and identify the associated input required from the highest type. Then, the only properties of the optimal test to be identified is which input is to be required by the types below the highest. Consider a test T with threshold π_L and a candidate indifferent type $\tilde{\lambda}$. Then, indifference requires that the agents with types $\lambda \in (\tilde{\lambda}, \overline{\lambda})$ choose inputs such that the indifference condition holds:

$$\frac{\int_{\tilde{\lambda}}^{\overline{\lambda}} \lambda e^{T,1}_{\lambda} dF(\lambda)}{1 - F(\tilde{\lambda})} = c\left(e^{T,1}_{\tilde{\lambda}}\right). \tag{5}$$

Note that an implication of the incentive compatibility constraint in Lemma 1 for

with $e^{T',1}_{\lambda^0} \in (e^{T',1}_{\tilde{\lambda}}, e^{T',1}_{\overline{\lambda}})$ and $c(e^{T',1}_{\lambda^0}) < \pi_L$. By introducing an input threshold at $e^{T',1}_{\lambda^0}$, the required input of all types above λ^0 increases but the highest type's cost is still below π_L.

[10] Recall that by Definition 1, our tie-breaking rule ensures that even if the highest type is indifferent between passing and failing, the highest type will pass the test.

how the required input can vary whenever the minimum input is differentiable is

$$\frac{d}{d\lambda} e^{T,1}_{\lambda} \in \left[-\frac{e^{T,1}_{\lambda}}{\lambda}, 0 \right].$$

Hence, the test T must induce a mapping from type-dependent required inputs to induced outputs that is a weakly decreasing function from $(e^{T,1}_{\overline{\lambda}}, \pi(e^{T,1}_{\overline{\lambda}}, \overline{\lambda}))$ to $(e^{T,1}_{\tilde{\lambda}}, \pi_L)$.

Finally, note that for any such mapping, the designer can strictly improve her objective unless there exists a type λ^k such that for all types $\lambda \in (\lambda^k, \overline{\lambda})$, the induced output is equal to the highest type's output, that is, $\pi(e^{T,1}_{\lambda}, \lambda) = \pi(e^{T,1}_{\overline{\lambda}}, \overline{\lambda})$, and for all $\lambda \in [\tilde{\lambda}, \lambda^k]$, the induced input is equal to the input of the indifferent type, that is, $e^{T,1}_{\lambda} = e^{T,1}_{\tilde{\lambda}}$. When this condition does not hold, the designer can change the test to such a step-test T' by introducing an input threshold $\overline{e} = e^{T,1}_{\tilde{\lambda}}$ and an output threshold $\overline{\pi}_H = \pi(e^{T,1}_{\overline{\lambda}})$. First, by construction, this test will be robust in that it does not feature an equilibrium in which all agents fail. Second, a strictly positive measure of agents above the indifferent type $\tilde{\lambda}$ will be required a strictly higher input to pass the test. These agents are willing to invest the additional input as long as the indifferent type has an incentive to pass the test, as their payoff is at least as high as the indifferent type's payoff. Third, the indifferent type now has a strict incentive to pass the test, as the higher types invest more, putting upward pressure on the market payment. Finally, the designer has an additional marginal gain from new agents passing this modified test.

Thus, the only candidate for an optimal pass-fail test remains the step-test proposed in Theorem 2 and illustrated in Figure 4. While capturing the same logic, the formal proof of Theorem 2 in the appendix takes a different approach. It recasts the designer's optimal test design problem as an optimal control problem. This approach has the advantage of being more easily adaptable to general production technologies, which we exploit in Section 5.2.

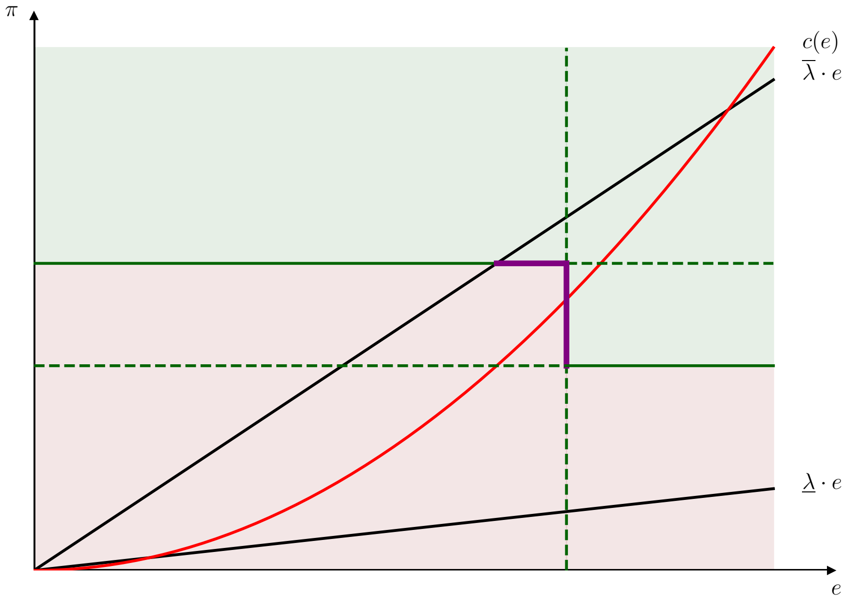

Figure 4: The optimal test for a pessimistic designer is a step-test. *The purple segments represent the input-output combinations generated by the passing agents in the induced rPBE.*

5 Extensions

In this final section, we show that our main insights are robust to different generalizations of our baseline model. First, we show that even when going beyond simple pass-fail tests, the optimal test design must rely on combinations of input and output-testing components. Second, we show that the characterization of optimal pass-fail tests in Theorem 2 holds for more general production functions than the one we have studied so far.

5.1 Beyond Pass-Fail Testing

Our analysis so far has focused on pass–fail tests. While binary ratings capture many relevant applications, it is natural to ask how richer testing technologies might affect the designer's optimal testing approach. In this section, we show that the main trade-offs identified for binary tests extend to arbitrary deterministic rating schemes.

A *multi-rating test* is a right-continuous function $T : \mathbb{R}^2_+ \to G$, mapping any reported input–output pair $(e^d, \pi^d) \in \mathbb{R}^2_+$ to a rating $g = T(e^d, \pi^d) \in G$, where G is a (possibly infinite) set of ratings. Such a test can be coarse (e.g., pass–fail) or arbitrarily fine (e.g., $T(e, \pi) = e$ or $T(e, \pi) = \pi$). In this context, a test T is called a *pure-input test* if $T(e, \pi)$ is independent of π, a *pure-output test* if it is independent of e, and a *composite test* otherwise.

The main result of this section is twofold, mirroring our findings in the pass–fail case. If the designer can coordinate the market and agents on her preferred equilibrium, a pure-input test suffices to induce the highest total output. If instead the designer seeks to maximize her payoff subject to robustness concerns, then the optimal test must combine input and output components to discipline market beliefs while preventing higher types from excessively coasting on their talent.

Theorem 3.

1. *The optimal multi-rating test for an optimistic designer is a pure-input test.*
2. *The optimal multi-rating test for a pessimistic designer is a composite test, necessarily depending on both input- and output-components.*

5.2 General Production Technology

In line with the human capital accumulation and certification literature, our analysis has so far considered a multiplicative technology $\pi(e, \lambda) = \lambda e$. This section shows that the main insights from Theorem 2 are significantly more general, extending to all settings where input is productive ($\pi(\lambda, e)$ increasing in e) and where input and productivity type act as weak complements (π weakly supermodular in (e, λ)). These mild conditions cover, among others, the additive specification $\pi(e, \lambda) = \lambda + e$.

Theorem 4. *Suppose output is generated by a production function $\pi(e, \lambda)$ that is increasing in e and weakly supermodular in (e, λ).*

1. *The optimal pass–fail test for an optimistic designer is a pure input test,*

$$T(e, \pi) = \mathbb{I}[e \geq \overline{e}].$$

2. *The optimal pass–fail test for a pessimistic designer is a step test,*

$$T(e, \pi) = \mathbb{I}[\pi \geq \overline{\pi}_H] + \mathbb{I}[e \geq \overline{e}]\, \mathbb{I}[\overline{\pi}_H > \pi \geq \overline{\pi}_L].$$

This generalization highlights that our results are not an artifact of the multiplicative formulation in our baseline model, but instead reflect deeper forces in environments where inputs and types interact more flexibly. The optimistic designer continues to favor pure input tests because equilibrium selection disciplines market beliefs, making output testing redundant. By contrast, the pessimistic designer relies on output thresholds to discipline beliefs and employs step tests to curb coasting on talent without sacrificing robustness.

6 Conclusion

In this paper, we highlight that the efficacy of a certification system depends as much on what is tested as it does on how certification is assigned. In particular, the choice between testing inputs and outputs is not merely a technical decision about what the market cares about, but a strategic choice that fundamentally alters agents' incentives and equilibrium outcomes. While pure output tests effectively anchor market beliefs, they suffer from a "coasting" inefficiency, allowing high-productivity agents to pass with minimal input investments. Conversely, pure input tests maximize incentive provision by standardizing input requirements, yet they remain fragile to pessimistic market expectations. Our analysis resolves this tension by deriving the optimal pass-fail certification. By conditioning certification on a specific combination of input and output thresholds, a designer can simultaneously secure robustness against adverse equilibria and provide strong incentives for input investments.

These findings open various avenues for future research. One promising direction is to study the commercialization of certification when the testing variable can be designed flexibly. Our current framework assumes a designer maximizing aggregate output, yet many real-world certifiers, from credit rating agencies to software compliance auditors, are profit-maximizing entities selling certification as a product. This shift in objective introduces a pricing dimension that may interact complexly with the test design and the optimal testing variable. For example, high-productivity agents might be willing to pay a premium for "output-only" tracks that allow them to leverage their talent, reveal high output levels to the market, while avoiding costly input monitoring.

A Omitted Proofs

A.1 Proof of Lemma 2

Proof. Note that in any PBE $\mathcal{E}^T$ following an input test with $\overline{e} \in [\check{e}, \hat{e}]$ it must hold that $W_{\mathcal{E}^T}(T,1) = c(\overline{e})$. Moreover, the market belief must be consistent with the agents' behavior, implying that for a passing set of agents $\Lambda_{\mathcal{E}^T} \subseteq [\underline{\lambda}, \overline{\lambda}]$, we must have $W_{\mathcal{E}^T}(T,1) = \mathbb{E}[\lambda \mid \lambda \in \Lambda_{\mathcal{E}^T}]$. Denote the set of agents passing the test in a cutoff equilibrium by $C = [\tilde{\lambda}, \overline{\lambda}]$ and the set of agents passing the test in any other equilibrium by $N \subset [\underline{\lambda}, \overline{\lambda}]$. Both being equilibria following the same input test with implies that they must induce the same expected productivity of passing agents

$$\lambda^e := \mathbb{E}[\lambda \mid \lambda \in C] = \mathbb{E}[\lambda \mid \lambda \in N]. \tag{6}$$

We want to show that the designer prefers the set C of passing agents over any other set N, which is equivalent to

$$\int_C \lambda dF(\lambda) \geq \int_N \lambda dF(\lambda). \tag{7}$$

By the definition of conditional expectations combined with C and N inducing the same conditional expectation, showing this inequality is equivalent to showing that

$$P(C) := \int_C dF(\lambda) \geq \int_N dF(\lambda) =: P(N), \tag{8}$$

that is, that the amount of passing agents in the cutoff equilibrium is higher than in any other equilibrium.

Denote by $A := C \setminus N$ and $B := N \setminus C$. Then, we obtain that showing $P(C) \geq P(N)$ is equivalent to showing $P(A) \geq P(B)$.[11] Note that $\lambda^e > \tilde{\lambda}$, as the density has full support and $C = [\tilde{\lambda}, \overline{\lambda}]$. By C and N having the same conditional expectation,

[11] This follows as $A \cup (C \cap N) = C$ and $B \cup (C \cap N) = N$.

we obtain the following sequence of manipulations

$$\int_C (\lambda - \lambda^e) dF(\lambda) = \int_N (\lambda - \lambda^e) dF(\lambda) \tag{9}$$

$$\int_{C \setminus N} (\lambda - \lambda^e) dF(\lambda) = \int_{C \setminus N} (\lambda - \lambda^e) dF(\lambda) \tag{10}$$

$$\int_A (\lambda - \lambda^e) dF(\lambda) = \int_B (\lambda - \lambda^e) dF(\lambda). \tag{11}$$

Observe that $A \subset C = [\tilde{\lambda}, \overline{\lambda}]$ and $B \subset C^c = [\underline{\lambda}, \tilde{\lambda})$ Thus, for all $\lambda \in A$, we have $\lambda \geq \tilde{\lambda}$ and for all $\lambda \in B$ we have $\lambda < \tilde{\lambda}$. Denote $\Delta := \tilde{\lambda} - \lambda^e < 0$, so that on A, $\lambda - \lambda^e \geq \tilde{\lambda} - \lambda^e = \Delta$, and on B, $\lambda - \lambda^e \leq \tilde{\lambda} - \lambda^e = \Delta$. These imply together that

$$\Delta P(A) \leq \int_A (\lambda - \lambda^e) dF(\lambda) = \int_B (\lambda - \lambda^e) dF(\lambda) \leq \Delta P(B). \tag{12}$$

As $\Delta < 0$, we obtain $P(A) \geq P(B)$ proving the result. □

A.2 Proof of Theorems 1, 2 and 4

Note that Theorems 1 and 2 are special cases of Theorem 4. Theorem 1 is the special case with production function $\pi(e, \lambda) = e \cdot \lambda$ and without the constraint that no equilibrium exists in which all agents fail, condition $(rPBE)$ below. Theorem 2 is just Theorem 4 with the production function $\pi(e, \lambda) = e \cdot \lambda$. As we prove all results using the same proof technique, leveraging optimal control methods, we prove them jointly.

To simplify notation, we omit the test- and rating-dependence of the minimum required effort $e_\lambda^{T,g}$ in the following and write $e(\lambda)$ instead. The following lemma is the immediate analogue of Lemma 3 under the general production function.

Lemma 5. *Given any optimal pass-fail test $\tilde{T}$ with an associated rPBE $\mathcal{E}^{\tilde{T}}$, there exists an outcome equivalent pass-fail test T with associated rPBE $\mathcal{E}^T$ that has an output threshold $\bar{\pi} = \pi(e(\tilde{\lambda}), \tilde{\lambda})$, where $\tilde{\lambda}$ is the marginal type in $\mathcal{E}^{\tilde{T}}$.*

Next, we state the incentive compatibility condition for the general production function, analogous to Lemma 1.

Lemma 6. *Any incentive compatible test must be such that the minimal required input of type λ $e(\lambda)$ is continuous and, at each point of differentiability, it satisfies* $\dot{e}(\lambda) \in \left[-\frac{\pi_\lambda(e(\lambda),\lambda)}{\pi_e(e(\lambda),\lambda)}, 0\right]$.

Proof. The result obtains directly by contradiction using the underreporting property whenever $e(\lambda)$ were not continuous or the slope outside the bounds stated in the lemma, as in Lemma 1. □

Optimization Problem We obtain the following optimization problem for the designer.[12]

$$\text{(OBJ)} \qquad \max_{e(\lambda)} \int_{\tilde{\lambda}}^{\bar{\lambda}} \pi(e(\lambda),\lambda) f(\lambda)\, d\lambda$$

$$\text{(IC)} \qquad \text{s. t. } \forall \lambda \in [\tilde{\lambda}, \bar{\lambda}],\ \dot{e}(\lambda) \in \left[-\frac{\pi_\lambda(e(\lambda),\lambda)}{\pi_e(e(\lambda),\lambda)}, 0\right]$$

$$\text{(PC)} \qquad \int_{\tilde{\lambda}}^{\bar{\lambda}} \pi(e(\lambda),\lambda) \frac{f(\lambda)}{1-F(\tilde{\lambda})}\, d\lambda - c(e(\tilde{\lambda})) \geq 0$$

$$\text{(rPBE)} \qquad \pi(e(\tilde{\lambda}), \tilde{\lambda}) - c(e(\bar{\lambda})) \geq 0.$$

Note that constraint (PC) ensures participation of the marginal type $\tilde{\lambda}$, and (rPBE) ensures that there is no equilibrium in which every agent fails.

In the next step, we rewrite this optimization problem as an optimal control problem with control $u(\lambda)$ and state variables $e(\lambda), w(\lambda)$ that evolve according to

$$\begin{aligned} \dot{e}(\lambda) &= u(\lambda) \\ \dot{w}(\lambda) &= \left(\pi(e(\lambda),\lambda) - \int_{\tilde{\lambda}}^{\lambda} \pi_e(e(\ell),\ell) \frac{f(\ell)}{F(\lambda)-F(\tilde{\lambda})}\, d\ell\right) \frac{f(\lambda)}{F(\lambda)-F(\tilde{\lambda})} \end{aligned}$$

$w(\lambda)$ is introduced as a state variable to handle the participation constraint of the marginal type $\tilde{\lambda}$, which includes an integral (isoperimetric) constraint, and is the

[12] In the following, as we adopt an optimal control approach, derivatives should be interpreted as right-derivatives whenever there is a kink in $e(\lambda)$.

wage conditional on types $\lambda \in [\tilde{\lambda}, \lambda]$ passing the test. Hence, $w(\bar{\lambda})$ represents the equilibrium wage.

Constraint (IC) implies that we have a bounded control $u(\lambda) \in \left[-\frac{\pi_\lambda(e(\lambda),\lambda)}{\pi_e(e(\lambda),\lambda)}, 0\right]$. The constraints (PC) and (rPBE) enter the control problem through initial- and terminal-value conditions. We choose the initial value of the state $e(\tilde{\lambda})$ as part of the optimization problem (which gives rise to the corresponding transversality conditions given the constraints of the problem. The Hamiltonian is

$$\begin{aligned}\mathcal{H} &= p_0 \cdot \pi(e(\lambda), \lambda) f(\lambda) + p_e(\lambda) u(\lambda) \\ &\quad + p_w(\lambda) \left(\pi(e(\lambda), \lambda) - \int_{\tilde{\lambda}}^{\lambda} \pi_e(e(\ell), \ell) \frac{f(\ell)}{F(\lambda) - F(\tilde{\lambda})} \, d\ell \right) \frac{f(\lambda)}{F(\lambda) - F(\tilde{\lambda})}\end{aligned}$$

where $p_0 \in \{0, 1\}$ ensures the well-definedness of the maximum principle and p_e and p_w are the co-states of the state variables e and w. (PC) and (rPBE) imply the following boundary constraints

$$\begin{aligned} w(\bar{\lambda}) - c(e(\tilde{\lambda})) &\geq 0 \\ \pi(e(\tilde{\lambda}), \tilde{\lambda}) - c(e(\bar{\lambda})) &\geq 0. \end{aligned}$$

The necessary conditions for optimality follow from Pontryagin's maximum principle

$$\begin{aligned}
u(\lambda) & \begin{cases} = 0, & \text{if } p_e(\lambda) > 0 \\ \in \left[-\frac{\pi_\lambda(e(\lambda),\lambda)}{\pi_e(e(\lambda),\lambda)}, 0\right], & \text{if } p_e(\lambda) = 0 \\ = -\frac{\pi_\lambda(e(\lambda),\lambda)}{\pi_e(e(\lambda),\lambda)}, & \text{if } p_e(\lambda) < 0 \end{cases} \\
\dot{p}_e(\lambda) &= -\pi_e(e(\lambda),\lambda) f(\lambda) \left(p_0 + \frac{p_w(\lambda)}{F(\lambda) - F(\tilde{\lambda})} \right) \\
p_e(\tilde{\lambda}) &= \gamma_p \pi_e(e(\tilde{\lambda}), \tilde{\lambda}) - \gamma_m c_e(e(\tilde{\lambda})) \\
p_e(\bar{\lambda}) &= -\gamma_p c_e(e(\bar{\lambda})) \\
\dot{p}_w(\lambda) &= 0 \\
p_w(\bar{\lambda}) &= \gamma_m \\
w(\tilde{\lambda}) &= \pi(e(\tilde{\lambda}), \tilde{\lambda}) \\
& \gamma_m \geq 0,\ w(\bar{\lambda}) - c(e(\tilde{\lambda})) \geq 0,\ \gamma_m \left(w(\bar{\lambda}) - c(e(\tilde{\lambda})) \right) = 0 \\
& \gamma_p \geq 0,\ \pi(e(\tilde{\lambda}), \tilde{\lambda}) - c(e(\bar{\lambda})) \geq 0,\ \gamma_p \left(\pi(e(\tilde{\lambda}), \tilde{\lambda}) - c(e(\bar{\lambda})) \right) = 0 \\
& p_0 \in \{0, 1\}.
\end{aligned}$$

Proof of Theorem 1 Note that we can recover the case without the robustness constraint ($rPBE$), by setting $\gamma_p = 0$. Thus, we immediately obtain $p_e(\overline{\lambda}) = 0$ and $p_e(\tilde{\lambda}) = -\gamma_m c_e(e(\tilde{\lambda}))$. As $\gamma_m \geq 0$, and because, whenever we can find a feasible path $e(\lambda)$, we obtain $p_0 = 1$, and hence, $\dot{p}_e(\lambda) < 0$, implying that $p_e(\lambda) > 0$ for all $\lambda \in [\tilde{\lambda}, \overline{\lambda})$, and hence $u(\lambda) = 0$. By appropriately choosing $\tilde{\lambda}$ and $e(\tilde{\lambda})$, a feasible solution exists and the optimal test is a pure input test.

Proof of Theorems 2 and 4 Note that it follows from $\dot{p}_w(\lambda) = 0$ that $p_w(\lambda) = \gamma_m \geq 0$. As $p_0 \in \{0, 1\}$, we obtain immediately that $\dot{p}_e(\lambda) \leq 0$. As for any $\tilde{\lambda}$ for which we can find a function $e(\lambda)$ such that all constraints are met, we obtain that $p_0 = 1$, and $\dot{p}_e(\lambda) < 0$. Hence, the co-state of $e(\lambda)$ is strictly decreasing throughout. Note that $p_e(\bar{\lambda}) \leq 0$ as $\gamma_p \geq 0$. Hence, there are only three candidate test structures (since $p_e(\lambda)$ can cross zero at most once and if it does, it does so from above):

(i) An L-test; if $\gamma_p = 0$, as then $p_e(\bar{\lambda}) = 0$, which combined with the optimally chosen singular control and $\dot{p}_e(\lambda) < 0$ implies that $u(\lambda) = 0$ for all $\lambda \in [\tilde{\lambda}, \bar{\lambda}]$), and we set (weakly) optimally $\pi(e(\lambda), \lambda) = \pi(e(\tilde{\lambda}), \tilde{\lambda})$ for all $\lambda < \tilde{\lambda}$.

(ii) A pure output test; if $\gamma_p > 0$ and $\gamma_m \geq \gamma_p \frac{\pi_e(e(\tilde{\lambda}),\tilde{\lambda})}{c_e(e(\tilde{\lambda}))}$, as then $p_e(\bar{\lambda}) < 0$ and $p_e(\tilde{\lambda}) \leq 0$, which together with $\dot{p}_e(\lambda) < 0$ implies that $p_e(\lambda) < 0$ for all $\lambda \in [\tilde{\lambda}, \bar{\lambda}]$, and hence, $u(\lambda) = -\frac{\pi_\lambda(e(\lambda),\lambda)}{\pi_e(e(\lambda),\lambda)}$ for all λ.

(iii) The staircase test; if $\gamma_p > 0$ and $\gamma_m < \gamma_p \frac{\pi_e(e(\tilde{\lambda}),\tilde{\lambda})}{c_e(e(\tilde{\lambda}))}$, as then $p_e(\bar{\lambda}) < 0$ and $p_e(\tilde{\lambda}) > 0$, which together with $\dot{p}_e(\lambda) < 0$ implies that there is a $\hat{\lambda} \in (\tilde{\lambda}, \bar{\lambda})$ such that $p_e(\lambda) > 0$, and thus, $u(\lambda) = 0$ for all $\lambda \in [\tilde{\lambda}, \hat{\lambda})$, and $p_e(\lambda) < 0$, and thus, $u(\lambda) = -\frac{\pi_\lambda(e(\lambda),\lambda)}{\pi_e(e(\lambda),\lambda)}$ for all $\lambda \in (\hat{\lambda}, \bar{\lambda}]$.

Next, note that both constraints, (PC) and $(rPBE)$ must be binding, implying that $\gamma_p > 0$ and $\gamma_m > 0$. Suppose the robustness constraint is not binding, implying that $\gamma_p = 0$ and $\pi(e(\tilde{\lambda}), \tilde{\lambda}) > c(e(\bar{\lambda}))$. In this case, we must be in (i), that is, have the L-test structure. However, if this were the case, $w(\bar{\lambda}) > c(e(\tilde{\lambda}))$ and the constraint on the marginal type is slack, implying $\gamma_m = 0$. This implies that $p_e(\tilde{\lambda}) = 0 = p_e(\bar{\lambda})$, which contradicts $\dot{p}_e(\lambda) < 0$.

Further, suppose that the robustness constraint $(rPBE)$ is binding, implying $\gamma_p > 0$ and $f(e(\tilde{\lambda}), \tilde{\lambda}) = c(e(\bar{\lambda}))$, but that the constraint on the marginal type (PC) is slack, implying that $\gamma_m = 0$ and $w(\bar{\lambda}) > c(e(\tilde{\lambda}))$. In this case, we must already feature a staircase test (see conditions for item (iii)). However, we can additionally conclude that it is suboptimal not to have the marginal type's participation constraint binding. By raising $e(\tilde{\lambda})$ marginally, no constraints are violated while the objective must have increased, contradicting optimality.

Hence, we conclude that $\gamma_p > 0$ and $\gamma_m > 0$ with both constraints binding. What remains to be shown is that it is never optimal to feature a pure output test (item (ii)). Note that in a pure output test $\pi(e(\tilde{\lambda}), \tilde{\lambda}) = \pi(e(\bar{\lambda}), \bar{\lambda})$ while $e(\bar{\lambda}) < e(\tilde{\lambda})$, implying that if $w(\bar{\lambda}) - c(e(\tilde{\lambda})) = 0$, then $\pi(e(\tilde{\lambda}), \tilde{\lambda}) - c(e(\bar{\lambda})) > 0$, as $w(\bar{\lambda}) = \pi(e(\tilde{\lambda}), \tilde{\lambda})$. Thus, both constraints cannot be binding simultaneously in a pure output test. It follows that the optimal test must be a staircase test as in item (iii).

Note that with an appropriate choice of $\tilde{\lambda}$ and $e(\tilde{\lambda})$ a feasible path $e(\lambda)$ always exists. Thus, in this case, we obtain $p_0 = 1$ and the optimal test for a pessimistic designer obtains as a step case.

The necessary conditions are sufficient to identify the optimal solution in our case, as have existence of an optimal solution for given initial conditions, and moreover, we obtain a unique solution to the necessary conditions implied by $\dot{p}_e(\lambda) < 0$.

A.3 Proof of Theorem 3

Optimistic Designer. Suppose the designer can coordinate the agents and the market on her preferred equilibrium in the subgame following any test T. Fix an arbitrary test $T : \mathbb{R}^2_+ \to G$ (with a possibly infinite grade set G), and let $\mathcal{E}^T$ denote the designer–preferred equilibrium in the subgame that follows it.

For every grade $g \in G$ attained in $\mathcal{E}^T$, denote by $\Lambda^T_g := \{\lambda : T(e^T_\lambda, \lambda e^T_\lambda) = g\}$ the set of agents attaining that grade, by $\lambda^T_g := \min \Lambda^T_g$ the lowest type achieving that grade, and by $e^T_g := e^T_{\lambda^T_g}$ the maximal input invested in $\mathcal{E}^T$ to attain g. As higher types are more productive, e^T_λ is weakly decreasing in λ, and $e^T_g \geq e^T_\lambda$ for all $\lambda \in \Lambda^T_g$.[13]

For any grade g with $\Lambda^T_g \neq \emptyset$, define

$$m^T(g, e) := \mathbb{E}\left[\lambda | \lambda \in O^T(g)\right] e - c(e), \tag{13}$$

where $O^T(g) := \bigcup_{g' \in G : e^T_{g'} = e^T_g} \Lambda^T_{g'}$ is the set of all types that attain grades associated with the same maximal effort as grade g in the equilibrium $\mathcal{E}^T$. Moreover, let

$$\underline{m}^T := \inf_{g \in G : O^T(g) \neq \emptyset} m^T(g, e^T_g). \tag{14}$$

Since $\mathcal{E}^T$ is an equilibrium and agents can always obtain a zero outside option by

[13] To see this, suppose otherwise. Then, there exists a type $\lambda \in \Lambda^T_g$ with $\lambda > \lambda^T_g$ and $e^T_\lambda > e^T_g$. However, type λ can do better by choosing the lower input e^T_g and reporting output $\lambda^T_g e^T_g$, which is strictly lower than the actual output and thus feasible. With this deviation, λ obtains the same grade at a lower cost.

not exerting effort, it follows that $\underline{m}^T \geq 0$.[14]

As $c(e)$ is strictly convex and satisfies $c(0) = 0,\ c'(0) = 0,\ m^T(g,e)$ is strictly concave in e. Thus, the equation $m^T(g,e) = \underline{m}^T$ admits two solutions, $e''_{T,g} < e'_{T,g}$, with the property that $e'_{T,g} \geq e^T_g$, as $m^T(g, e^T_g) \geq \underline{m}^T$. For each g attained in $\mathcal{E}^T$, we use the *largest* solution $e'_{T,g}$ of $m^T(g,e) = \underline{m}^T$ to construct an improvement for the designer.

Consider the test[15]

$$T'(e,\pi) := \begin{cases} e'_{T,g} & \text{if } e = e'_{T,g} \text{ for } g \in G : O^T(g) \neq \emptyset \\ 0 & \text{ otherwise.} \end{cases} \tag{15}$$

In this test, for each g attained in $\mathcal{E}^T$, every type $\lambda \in O^T(g)$ is willing to invest $e^{T'}_\lambda = e'_{T,g}$ if the composition of $O^T(g)$ remains unchanged, as, by construction, achieving any grade $g' = e'_{T,g}$ costs $c(e'_{T,g})$ to every type and delivers an expected payoff

$$\underline{m}^T = \mathbb{E}\left[\lambda \mid \lambda \in O^T(g)\right] e'_{T,g} - c\left(e'_{T,g}\right) \geq 0. \tag{16}$$

Thus, all agents are indifferent among all input levels e such that $e = e'_{T,g}$—that is, indifferent among all available grades. Note that if $\mathcal{E}^T$ features some agents achieving a failing grade with $e^T_g = 0$, then $\underline{m}^T = 0$. It follows that there exists an equilibrium $\mathcal{E}^{T'}$ in the subgame following T' such that the types $\lambda \in O^T(g)$ achieve the grades associated with $e'_{T,g}$. Note that in the equilibrium $\mathcal{E}^{T'}$, for any attained grade g and any $\lambda \in O^T(g)$, we have $e'_{T,g} \geq e^T_g \geq e^T_\lambda$. Therefore, this tests improves the designer's payoff:

$$\int_{\underline{\lambda}}^{\overline{\lambda}} \lambda e^{T'}_\lambda dF(\lambda) \geq \int_{\underline{\lambda}}^{\overline{\lambda}} \lambda e^T_\lambda dF(\lambda), \tag{17}$$

[14] Indeed, in $\mathcal{E}^T$, the lowest type in $O^T(g)$ invests input e^T_g and obtains an expected payoff lower than $m^T(g, e^T_g)$ (as $e^T_g \geq e^T_\lambda$ for all $\lambda \in \Lambda^T_g$). Thus $m^T(g, e^T_g) \geq \underline{m}^T \geq 0$ for all g with $\Lambda^T_g \neq \emptyset$.

[15] If there are more than one grade with the same $e_{T,g}$ in $\mathcal{E}^T$, we pick merge these into a single grade.

with a strict inequality whenever $e'_{T,g} > e_g^T$ for some grade g. Consequently, the designer weakly improves by switching to the pure-input test T', proving the first part of Theorem 3.

Pessimistic Designer To prove the Pessimistic Designer result, we proceed in several steps.

Step 0: Pure–input tests are never optimal. We first show that, among pure-input tests—that is, tests satisfying $T(e,\pi) = T(e,\pi')$ for all π, π'—the best possible outcome is achieved by the pass/fail test $T'(e,\pi) = \mathbb{I}(e \geq e^*)$, where e^* solves $\underline{\lambda} e^* = c(e^*)$. Indeed, while T' induces only equilibria in which all agent types choose e^*, any other pure-input test admits an equilibrium where all types choose $e_\lambda \leq e^*$. To see this, note that if there were some grade that requires an input level $e_\lambda > e^*$, there always exists an off-path belief such that only the lowest type $\underline{\lambda}$ achieves this grade. Under this off-path belief, there exists an equilibrium in which all types invest no more than e^*. Thus, the optimal pure input test under designer-worst equilibrium selection is the pass/fail input test with input threshold e^*. Since Theorem 2 implies that T' is strictly dominated by some input-output test, no pure-input test can be optimal.

Step 1: Pure–output tests are never optimal. Consider any pure output test T—that is, a test such that $T(e,\pi) = T(e',\pi)$ for all $e, e' \in \mathbb{R}_+$. Note that we can represent any such test as $T(e,\pi) = \pi\mathbb{I}(\pi \in \Pi_T)$, where Π_T is the set of output reports that T allows to reveal; that is, each of these output levels is associated with a different grade in the test. The following properties hold for any equilibrium $\mathcal{E}^T$ following T:

a If $T(e_\lambda^T, \lambda e_\lambda^T) = \pi_t$, then (i) $e_\lambda^T = \pi_t/\lambda$, and (ii) $T(e_{\lambda'}^T, \lambda' e_{\lambda'}^T) \geq \pi_t$ for all $\lambda' \geq \lambda$; that is, higher types attain weakly higher output levels in equilibrium. Indeed, since the production function is $\pi = \lambda e$, reaching higher output levels is more costly, but less so for higher types. To see this, note that the difference in payoffs for two grades associated with output $\pi \geq \pi'$ for any type λ is

$$\Delta(\lambda) = \left(\pi - c\left(\frac{\pi}{\lambda}\right)\right) - \left(\pi' - c\left(\frac{\pi'}{\lambda}\right)\right), \tag{18}$$

which increases in λ due to the convexity of the cost function

$$\frac{d\Delta(\lambda)}{d\lambda} = c'\left(\frac{\pi}{\lambda}\right)\frac{\pi}{\lambda^2} - c'\left(\frac{\pi'}{\lambda}\right)\frac{\pi'}{\lambda^2} > 0. \tag{19}$$

b For each type $\lambda \in [\underline{\lambda}, \overline{\lambda}]$, denote by e^{**}_λ the input that maximizes agent λ's payoff if the market were to observe her input-output combination directly, i.e., $e^{**}_\lambda := \arg\max_{e \geq 0}\big\{\lambda e - c(e)\big\}$. If the corresponding output level $\pi^{**}_\lambda := \lambda e^{**}_\lambda$ is such that $\pi^{**}_\lambda \in \Pi_T$, then exerting e^{**}_λ is strictly dominant for λ in the subgame following T. Note that both e^{**}_λ and π^{**}_λ are strictly increasing and continuous in $\lambda \in [\underline{\lambda}, \overline{\lambda}]$.

c If $T(e^T_{\lambda'}, \lambda' e^T_{\lambda'}) = T(e^T_{\lambda''}, \lambda'' e^T_{\lambda''}) = \pi_t$ for $\lambda' < \lambda''$, then (i) $T(e^T_\lambda, \lambda e^T_\lambda) = \pi_t$ for all $\lambda \in [\lambda', \lambda'']$ (a direct consequence of (a)), and (ii) there cannot be $\pi_1 < \pi_t < \pi_2$ such that $[\pi_1, \pi_2] \subseteq \Pi_T$. To see this, note $\pi - c\left(\frac{\pi}{\lambda}\right)$ is strictly concave in π with a maximum at π^{**}_λ, and $\pi^{**}_{\lambda''} > \pi^{**}_{\lambda'}$. Hence, types λ' and λ'' cannot both prefer the same π_t over all output levels in $[\pi_1, \pi_2]$ if those output level could also be revealed, $[\pi_1, \pi_2] \subseteq \Pi_T$.

d If $\mathcal{E}^T$ is an optimal test, at least two grades associated with two different output levels must be attained in $\mathcal{E}^T$. This observation follows from the suboptimality of a pure output pass/fail test, as shown in theorem 2.

Endowed with these properties, we can prove that a pure output test T can never be optimal in the Pessimistic Designer case.

STEP 1.0 Consider the unique equilibrium $\mathcal{E}^T$ following T, where uniqueness follows from the test being a pure output test. Denote by $u_\lambda : \Pi \to \mathbb{R}$ the function mapping any hypothetical grade g in a pure output test associated with the output level

π to the corresponding payoff for type λ of achieving that grade:

$$u_\lambda(\pi) := \pi - c\left(\frac{\pi}{\lambda}\right).$$

Moreover, given any grade g associated with an output level π_H achieved in $\mathcal{E}^T$, denote by λ_H^{marg} and λ_H^{max} the lowest and highest types, respectively, that achieve that grade in $\mathcal{E}^T$, i.e., for which $u_\lambda(\pi_H) \geq u_\lambda(\pi)$ for all $\pi \in \Pi_T$. Finally, for every λ and π such that $u_\lambda^T(\pi) \geq 0$ and $\pi > \pi_\lambda^{**}$, denote by $\pi_\lambda^{low} \neq \pi$ the output level for which $u_\lambda^T(\pi) = u_\lambda^T(\pi_\lambda^{low})$. Note that strict convexity of c implies the following:

- π_λ^{low} exists, is unique, and increases in λ;
- $\pi_\lambda^{low} < \pi_\lambda^{**} < \pi$;
- $u_\lambda^T(\pi') > u_\lambda^T(\pi)$ for all $\pi' \in (\pi_\lambda^{low}, \pi)$, and $u_\lambda^T(\pi') < u_\lambda^T(\pi)$ for all $\pi' \in \mathbb{R}_+ \setminus [\pi_\lambda^{low}, \pi]$.

STEP 1.1 *If output test T is optimal, there exists $\lambda_l < \overline{\lambda}$ such that in $\mathcal{E}^T$ the output attained by type λ, λe_λ^T is strictly increasing and continuous in $\lambda \in [\lambda_l, \overline{\lambda}]$.*

Strict monotonicity. The fact that it is weakly increasing is a direct consequence of property (a) above. To prove strict monotonicity, suppose, toward a contradiction, that there exists $\lambda_l < \overline{\lambda}$ such that $\lambda_l e_{\lambda_l}^T = \overline{\lambda} e_{\overline{\lambda}}^T := \bar{\pi}$. Then:

(i) $\bar{\pi} \geq \pi_{\overline{\lambda}}^{**}$. Indeed, suppose toward a contradiction that $\bar{\pi} < \pi_{\overline{\lambda}}^{**}$, and denote by $\lambda_A < \overline{\lambda}$ the agent type such that $\bar{\pi} = \pi_{\lambda_A}^{**}$. Then, the designer could profitably deviate to an alternative test that reveals output levels above $\bar{\pi}$:

$$T'(e, \pi) := \begin{cases} \pi & \text{if } \pi \geq \bar{\pi} \\ T(e, \pi) & \text{otherwise.} \end{cases}$$

Indeed, for any type λ both the cost of attaining a grade associated with $\pi_t \in \Pi_T$ such that $\pi_t \leq \bar{\pi}$, $c\left(\frac{\pi_t}{\lambda}\right)$, and the associated wage $W_{\pi_t} = \pi_t$ under T', coincide with those under T. Thus, no agent reduces her equilibrium input under test T' relative to T. Moreover, by (b), each $\lambda \in [\lambda_A, \overline{\lambda}]$ that was attaining $\bar{\pi}$ with input $\frac{\bar{\pi}}{\lambda}$ under T optimally invests $\frac{\pi^{**}_{\lambda}}{\lambda}$ under T'. Because $\pi^{**}_{\lambda} > \pi^{**}_{\lambda_A}$ for all $\lambda \in (\lambda_A, \overline{\lambda}]$, the modified test T' generates a gain for the designer equal to

$$\int_{\lambda_A}^{\overline{\lambda}} (\pi^{**}_{\lambda} - \bar{\pi})\,\mathrm{d}F(\lambda) > 0,$$

contradicting the optimality of T.

(ii) $(\bar{\pi}^{low}_{\lambda_l}, \bar{\pi}) \cap \Pi_T = \emptyset$; i.e., no output level $\pi \in (\bar{\pi}^{low}_{\lambda_l}, \bar{\pi})$ is associated with a grade in test T. Since $\bar{\pi} \geq \pi^{**}_{\overline{\lambda}} > \pi^{**}_{\lambda_l}$ and, by assumption, l obtains $\bar{\pi}$ in $\mathcal{E}_T$, $\pi^{low}_{\lambda_l}$ is well defined, and $(\bar{\pi}^{low}_{\lambda_l}, \bar{\pi}) \cap \Pi_T = \emptyset$: otherwise λ_l would optimally deviate to a grade associated with such an output level.

However, it follows from (i) and (ii) that the designer can profitably deviate to an alternative test T' (a contradiction). Indeed, denote by $\pi_B \leq \bar{\pi}^{low}_{\lambda_l} < \bar{\pi}$ the second highest grade in Π_T that is attained in $\mathcal{E}^T$. Moreover, define by e^{dev}_{λ} the input such that $\bar{\pi} - c(e^{dev}_{\lambda}) = u_{\lambda}(\pi_B)$. Define the test T'

$$T'(e,\pi) := \begin{cases} \bar{\pi} & \text{if } \pi \geq \bar{\pi} \text{ and } e \geq e^{dev}_{\overline{\lambda}} \\ T(e,\pi) & \text{if } \pi < \bar{\pi} \\ T(e,0) & \text{otherwise,} \end{cases}$$

which features an L-threshold at the top that requires an output level of at least $\bar{\pi}$ and an input level of at least $e^{dev}_{\overline{\lambda}}$. For all output levels below $\bar{\pi}$ the test T' is identical to T, and for all output levels above $\bar{\pi}$ together with input levels below $e^{dev}_{\overline{\lambda}}$, the test assigns the lowest grade from test T, which delivers a payoff of zero.

For any type λ, attaining output $\pi_t < \bar{\pi}$, with $\pi_t \in \Pi_T$, has the same cost $c\left(\frac{\pi_t}{\lambda}\right)$ and the same benefit π_t under T and T'. Thus, no agent obtaining a grade $\pi_t < \bar{\pi}$ selects a lower input under T' than under T.

Moreover, all types achieving grade $\bar{\pi}$ in $\mathcal{E}_T$ will still prefer attaining the grade labeled $\bar{\pi}$ over any grade associated with output levels $\pi_t < \bar{\pi}$ under T'. To see this, note that, by construction of $e^{dev}_{\overline{\lambda}}$, the highest type $\overline{\lambda}$ will not benefit from deviating to lower output levels. As $e^{dev}_{\overline{\lambda}}$ decreases in λ and $\pi_B \leq \pi^{low}_{\lambda_l}$ (by (ii)), all other types $\lambda \in [\lambda_l, \overline{\lambda})$ will also attain the grade associated with output level $\bar{\pi}$ and input threshold $e^{dev}_{\overline{\lambda}}$.

Finally, since $\pi_B \leq \pi^{low}_{\lambda_l} < \pi^{low}_{\overline{\lambda}}$, we know that $e^{dev}_{\overline{\lambda}} > \frac{\bar{\pi}}{\overline{\lambda}}$. Thus, there exists $\lambda' \in (\underline{\lambda}, \overline{\lambda})$ such that each agent $\lambda \in [\lambda', \lambda]$ chooses higher input under T' than under T. This results in a gain for the designer of at least

$$\int_{\lambda'}^{\overline{\lambda}} \left(\lambda e^{dev}_{\overline{\lambda}} - \bar{\pi}\right) \mathrm{d}F(\lambda) > 0,$$

contradicting the optimality of T. Thus, λe^T_λ is strictly increasing in λ over the interval $[\lambda_l, \overline{\lambda}]$.

Continuity. If λe^T_λ is strictly increasing in $\lambda \in [\lambda_l, \overline{\lambda}]$, it must also be left-continuous on $\lambda \in (\lambda_l, \overline{\lambda}]$. Suppose not. Then there exists $\lambda' \in (\lambda_l, \overline{\lambda}]$ such that $\lim_{\lambda \to \lambda'^-} \lambda e^T_\lambda < \lambda' e^T_{\lambda'}$. But, since λe^T_λ is strictly increasing on $\lambda \in [\lambda_l, \overline{\lambda}]$, there cannot be a mass of types attaining grade $\pi' = \lambda' e^T_{\lambda'}$. By continuity of c and point (a) above, λ' must then be indifferent between attaining $\lambda e^T_{\lambda'}$ and $\lim_{\lambda \to \lambda'^-} \lambda e^T_\lambda$ (if $\lim_{\lambda \to \lambda'^-} \lambda e^T_\lambda \in \Pi_T$). Given the strict concavity of $\pi - c(\pi/\lambda)$ in π, we have $\lim_{\lambda \to \lambda'^-} \lambda e^T_\lambda = (\pi')^{low}_{\lambda'} < \pi^{**}_{\lambda'} < \pi'$, implying the existence of $\lambda'' \in (\lambda_l, \lambda')$ such that $\lambda e^T_\lambda < \pi^{**}_\lambda$ for all $\lambda \in (\lambda'', \lambda')$. This yields a contradiction as all types in (λ'', λ') would prefer a grade closer to $\lim_{\lambda \to \lambda'^-} \lambda e^T_\lambda \in \Pi_T$, all types in (λ'', λ'):

- if $\lim_{\lambda \to \lambda'^-} \lambda e^T_\lambda \in \Pi_T$, all types in (λ'', λ') would pool at this grade, violating strict monotonicity;
- if $\lim_{\lambda \to \lambda'^-} \lambda e^T_\lambda \notin \Pi_T$, then types in (λ'', λ') would have no best reply, contradicting the existence of the candidate equilibrium.

Thus, λe^T_λ is left-continuous in λ over $(\lambda_l, \overline{\lambda}]$. An analogous argument shows right-continuity over the same range

STEP 1.2 *For any $\lambda_L \in [\underline{\lambda}, \overline{\lambda}]$ such that λe^T_λ is strictly increasing and continuous on $[\lambda_L, \overline{\lambda}]$, we have $\lambda e^T_\lambda = \pi^{**}_\lambda$ for all $\lambda \in [\lambda_L, \overline{\lambda}]$.*

By step (1.1), we know that there exists $\lambda_L < \overline{\lambda}$ such that λe^T_λ is strictly increasing and continuous on $\lambda \in [\lambda_L, \overline{\lambda}]$. Using the notation $\pi^T_\lambda := \lambda e^T_\lambda$, we have that $\pi^T_{\lambda_L} < \pi^T_{\overline{\lambda}}$ and $\left[\pi^T_{\lambda_L}, \pi^T_{\overline{\lambda}}\right] \subseteq \Pi_T$.

We first show that $\pi^T_{\overline{\lambda}} = \pi^{**}_{\overline{\lambda}}$. Indeed, if $\pi^T_{\overline{\lambda}} < \pi^{**}_{\overline{\lambda}}$, there would exist $\lambda' < \overline{\lambda}$ such that all $\lambda \in \left[\lambda', \overline{\lambda}\right]$ would profitably deviate to attain $\pi^T_{\overline{\lambda}}$; and if $\pi^T_{\overline{\lambda}} > \pi^{**}_{\overline{\lambda}}$, agent $\overline{\lambda}$ would profitably deviate to some $\pi \in \left[\pi^T_{\lambda_L}, \pi^T_{\overline{\lambda}}\right)$.

Combining $\left[\pi^T_{\lambda_L}, \pi^T_{\overline{\lambda}}\right] \in \Pi_T$ with $\pi^T_{\overline{\lambda}} = \pi^{**}_{\overline{\lambda}}$ and an analogous optimality reasoning, we obtain $\lambda e^T_\lambda = \pi^{**}_\lambda$ for all $\lambda \in \left(\lambda_L, \overline{\lambda}\right]$.

Finally, note that by continuity of π^{**}_λ and strictly increasing λe^T_λ on $\left[\lambda_L, \overline{\lambda}\right]$, we must also have that $\pi^T_{\lambda_L} = \pi^{**}_{\lambda_L}$.

STEP 1.3 For any $\lambda_L \in [\underline{\lambda}, \overline{\lambda}]$ such that λe^T_λ is strictly increasing and continuous on $[\lambda_L, \overline{\lambda}]$, we have $\lim_{\lambda \to \lambda_L^-} \lambda e^T_\lambda = \lambda_L e^T_{\lambda_L} = \pi^{**}_{\lambda_L}$.

By Step 1.2, $\lambda e^T_\lambda = \pi^{**}_\lambda$ for all $\lambda \in \left[\lambda_L, \overline{\lambda}\right]$. Suppose, toward a contradiction, $\lim_{\lambda \to \lambda_L^-} \lambda e^T_\lambda < \pi^{**}_{\lambda_L}$. Then there must exist $\lambda'' \in (\underline{\lambda}, \lambda_L)$ such that $[\pi^{**}_\lambda, \pi^{**}_{\lambda_L}) \cap \Pi_T = \emptyset$ and $\lambda e^T_\lambda < \pi^{**}_\lambda$ for all $\lambda \in (\lambda'', \lambda_L)$. However, since $u^T_\lambda(\pi)$ is continuous in λ and, for every λ, strictly increasing over $[0, \pi^{**}_\lambda]$, this implies that there exists $\lambda''' \in [\lambda'', \lambda_L)$ such that all $\lambda \in [\lambda''', \lambda_L)$ optimally select $\pi^{**}_{\lambda_L}$ following T; a contradiction to $\lim_{\lambda \to \lambda_L^-} \lambda e^T_\lambda < \pi^{**}_{\lambda_L}$.

STEP 1.4 *For any $\lambda_L \in [\underline{\lambda}, \overline{\lambda}]$ such that λe^T_λ is strictly increasing and continuous on $[\lambda_L, \overline{\lambda}]$, we have that there is no $\lambda'' < \lambda_L$ such that $\lambda'' e^T_{\lambda''} = \lambda_L e^T_{\lambda_L} =: \pi_{\lambda_L}$.*

Suppose otherwise, and denote by λ'' the lowest type for whom $\lambda'' e^T_{\lambda''} = \pi_{\lambda_L}$. Then $\lambda e^T_\lambda = \pi_{\lambda_L}$ for all $\lambda \in [\lambda'', \lambda_L]$. Moreover, because $\pi_{\lambda_L} = \pi^{**}_{\lambda_L} > \pi^{**}_{\lambda''}$, $(\pi_{\lambda_L})^{low}_{\lambda''} < \pi_{\lambda_L}$ is well defined, and $((\pi_{\lambda_L})^{low}_{\lambda''}, \pi_{\lambda_L}) \cap \Pi_T = \emptyset$; otherwise, λ'' could profitably deviate. However, this implies that the designer could profitably deviate to an alternative test.

If $\lambda e_\lambda^T = 0$ for all $\lambda < \lambda''$, then it is immediate to see that it is profitable, for the designer, to deviate to $T'(e,\pi) := T(e,\pi)\mathbb{I}(e \geq e_{\lambda''}^T)$.

If instead $\lambda e_\lambda^T > 0$ for some $\lambda < \lambda''$, let $\pi_B := \min\{\pi \in \Pi_T : \pi < \pi_{\lambda_L}^{**}\} \leq (\pi_{\lambda_L})_{\lambda''}^{low}$ denote the highest attainable grade strictly below $\pi_{\lambda_L}^{**}$. For all $\lambda \in [\lambda'', \lambda_L]$, define $(\pi_B)_\lambda^{\text{high}}$ as the unique output level $\pi > \pi_\lambda^{**}$ that makes type λ indifferent between attaining π_B and $(\pi_B)_\lambda^{\text{high}}$, that is, $(\pi_B)_\lambda^{\text{high}} - c\left(\frac{(\pi_B)_\lambda^{\text{high}}}{\lambda}\right) = u_\lambda(\pi_B)$. Finally, for all $\lambda' \in [\lambda'', \lambda_L]$, let $\lambda_{\lambda'}^B$ denote the unique type such that $\pi_{\lambda_{\lambda'}^B}^{**} = (\pi_B)_{\lambda'}^{\text{high}}$.

Note that $(\pi_B)_\lambda^{high}$ is strictly increasing in λ and, since λ'' must be indifferent between $\pi_{\lambda_L}^{**}$ and π_B, we must have $(\pi_B)_{\lambda''}^{high} = \pi_{\lambda_L}^{**}$. Thus $e_{\lambda_\lambda^B}^{**}$ is also strictly increasing in $\lambda \in [\lambda'', \lambda_L]$, with $e_{\lambda_{\lambda''}^B}^{**} = e_{\lambda_L}^{**}$. Combined with the fact that e_λ^T is strictly decreasing in $\lambda \in [\lambda'', \lambda_L]$, with $e_{\lambda_L}^T = e_{\lambda_L}^{**}$, this implies there exist a unique $\tilde{\lambda} \in [\lambda'', \lambda_L)$ such that $e_{\lambda_{\tilde{\lambda}}^B}^{**} = e_{\tilde{\lambda}}^T$ Denoting $\hat{\lambda} := \lambda_{\tilde{\lambda}}^B$, we can show that the designer could profitably deviate to an alternative test

$$T'(e,\pi) := \begin{cases} T(e,\pi) & \text{if } \pi < \pi_{\lambda_L}^{**} \text{ or } \pi \geq \pi_{\hat{\lambda}}^{**} \\ \pi_L & \text{if } \pi \in \left[\pi_{\lambda_L}^{**}, \pi_{\hat{\lambda}}^{**}\right] \text{ and } e \geq e_{\hat{\lambda}}^{**} \\ T(e,0) & \text{otherwise} \end{cases}$$

Indeed, for any type λ, the payoff from attaining any grade $\pi_t \in \Pi_T \setminus \left[\pi_{\lambda_L}^{**}, \pi_{\hat{\lambda}}^{**}\right]$ under T' is the same as under T. Thus no agent obtaining grade $\pi < \pi_L^{**}$ under T will invest lower input under T'. Moreover, all types $\lambda \leq \tilde{\lambda}$ who attain grade $\pi_{\lambda_L}^{**}$ in $\mathcal{E}_T$ will still prefer, under T', to achieve $\pi_{\lambda_L}^{**}$ rather than any $\pi_t \leq \pi_{\lambda_L}^{**}$: their payoff from $\pi_{\lambda_L}^{**}$ is (weakly) higher under T' than under T (as, by construction, $e_\lambda^T = \frac{\pi_{\lambda_L}^{**}}{\lambda} \geq e_{\tilde{\lambda}}^T = e_{\hat{\lambda}}^{**}$). Additionally, since (by construction) $\tilde{\lambda}$ is indifferent between $\pi_{\hat{\lambda}}^{**}$ and π_B, all types $\lambda \in [\tilde{\lambda}, \hat{\lambda}]$ prefer $\pi_{\hat{\lambda}}^{**}$ to any $\pi \leq \pi_B$. Thus, under T', they will either aim for $\pi_{\hat{\lambda}}^{**}$ or $\pi_{\lambda_L}^{**}$, both of which now require them to invest higher input (as $e_\lambda^T = \frac{\pi_{\lambda_L}^{**}}{\lambda} \leq e_{\tilde{\lambda}}^T = e_{\hat{\lambda}}^{**}$ for all $\lambda \in [\tilde{\lambda}, \hat{\lambda}]$)—benefiting the designer. Finally, all agents who attain $\pi > \pi_{\hat{\lambda}}^{**}$ under $\mathcal{E}^T$ will continue to choose π under T', since $\pi^{**}_{\lambda_L}$ still yields a return below $\pi^{**}_{\hat{\lambda}}$.

Step 1.5 Taken together, Steps 1.0-1.4 imply that if a pure output test is optimal, then $T(\pi, e) = \pi\mathbb{I}\left(\pi \in [\pi^{**}_{\underline{\lambda}}, \pi^{**}_{\overline{\lambda}}]\right)$is optimal. The last step is thus to show that this test could be improved upon. In particular, it is straightforward to see that the designer would obtain a strictly higher payoff by deviating to $T(\pi, e) = \pi\mathbb{I}\left(\pi \in [\underline{\pi}, \pi^{**}_{\overline{\lambda}}]\right)$, where $\underline{\pi} > \pi^{**}_{\underline{\lambda}}$ is the output level such that $\underline{\pi} - c\left(\frac{\underline{\pi}}{\underline{\lambda}}\right) = 0$.

References

Augias, Victor, and Eduardo Perez-Richet. 2023. "Non-Market Allocation Mechanisms: Optimal Design and Investment Incentives." *arXiv preprint arXiv:2303.11805.*

Ball, Ian. 2019. "Scoring strategic agents." *arXiv preprint arXiv:1909.01888.*

Becker, William E, and Sherwin Rosen. 1992. "The learning effect of assessment and evaluation in high school." *Economics of Education Review* 11 (2): 107–118.

Bergemann, Dirk, and Stephen Morris. 2009. "Robust implementation in direct mechanisms." *The Review of Economic Studies* 76 (4): 1175–1204.

Bizzotto, Jacopo, and Adrien Vigier. 2021. "Optimal school design." *Available at SSRN 3877063.*

Boleslavsky, Raphael, and Christopher Cotton. 2015. "Grading standards and education quality." *American Economic Journal: Microeconomics* 7 (2): 248–279.

Costrell, Robert M. 1994. "A simple model of educational standards." *The American Economic Review*: 956–971.

Daley, Brendan, and Brett Green. 2014. "Market signaling with grades." *Journal of Economic Theory* 151: 114–145.

Dubey, Pradeep, and John Geanakoplos. 2010. "Grading exams: 100, 99, 98,... or a, b, c?" *Games and Economic Behavior* 69 (1): 72–94.

Dworczak, Piotr, and Alessandro Pavan. 2022. "Preparing for the worst but hoping for the best: Robust (bayesian) persuasion." *Econometrica* 90 (5): 2017–2051.

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

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

Fudenberg, Drew, and Jean Tirole. 1991. “Perfect Bayesian equilibrium and sequential equilibrium.” *journal of Economic Theory* 53 (2): 236–260.

Halac, Marina. 2025. “Contracting for coordination.” *Journal of the European Economic Association* 23 (3): 815–844.

Halac, Marina, Elliot Lipnowski, and Daniel Rappoport. 2024. “Pricing for Coordination.” *Working Paper.*

Kapon, Sam. 2023. “Persuasion in Evidentiary Mechanisms.” *Working Paper.*

Ma, Ching-To. 1988. “Unique implementation of incentive contracts with many agents.” *The Review of Economic Studies* 55 (4): 555–572.

Mishra, Debasis, Sanket Patil, and Alessandro Pavan. 2025. “Robust Procurement Design.” *Working Paper.*

Perez-Richet, Eduardo, and Vasiliki Skreta. 2022. “Test design under falsification.” *Econometrica* 90 (3): 1109–1142.

Perez-Richet, Eduardo, and Vasiliki Skreta. 2023. “Fraud-proof non-market allocation mechanisms.”Technical report, Working paper.

Popov, Sergey V, and Dan Bernhardt. 2013. “University competition, grading standards, and grade inflation.” *Economic inquiry* 51 (3): 1764–1778.

Spence, Michael. 1978. “Job market signaling.” In *Uncertainty in economics*, 281–306: Elsevier.

Wolinsky, Asher. 1993. “Competition in a market for informed experts’ services.” *The RAND Journal of Economics*: 380–398.

Xiao, Peiran. 2025. “Incentivizing Agents Through Ratings.” *working paper.*