

Attempt Limits in Standardized Testing: Theory and Evidence ^{*}

Francesco Billari [†]

December 6, 2025

[Please click here for the latest version.](#)

Abstract

Attempt limits are widely used in standardized tests for personnel selection and licensing. Using data from the U.S. Bar Exam, I provide descriptive evidence that introducing attempt limits increases passing rates. This is consistent with the rationale that limiting attempts imposes a test-taking cost that induces candidates to self-select into test sessions based on their level of preparation: candidates sit only when they expect a sufficiently high probability of passing. In this sense, attempt limits can improve the average quality of successful candidates. Motivated by this evidence, I develop a principal–agent model revealing a trade-off: attempt limits deter weaker candidates but also prevent some stronger ones from eventually succeeding, so governments strongly averse to human capital loss may oppose them. Thus, I provide a normative framework for the design of attempt limits. First, I show that the limit is optimal only if the test is not difficult. Second, when the limit is chosen jointly with the passing score, the optimal design depends on the test’s information structure: if the test is better at identifying strong candidates, the attempt limit is optimal; if instead the test is better at identifying weak candidates, the principal prefers a difficult exam with unlimited retakes. Finally, when the information structure itself can be chosen, the principal should adopt a test that is better at identifying strong agents if candidate’s responsiveness to the attempt limit is high.

Keywords: Attempt Limits, Standardized Tests, Personnel, License.

JEL Codes: D02, D04, D82, D86, H40, H83.

^{*}I am very grateful to Marco Battaglini, Stephen Coate and Tommaso Denti for advising me on this paper. I extend my gratitude to Takuma Habu and Jacob Dorn for their constructive feedback. I thank various audiences at Cornell University for their valuable comments on the paper.

[†]Department of Economics, Cornell University. Email: fb293@cornell.edu

1 Introduction

Tests are widely used by institutions for personnel selection and professional licensing. For example, the Indian Civil Service Examination attracts about 1.3 million applicants each year, while the Chinese Civil Service Examination draws between 1.5 and 3.5 million candidates annually. In the realm of licensing, passing the United States Medical Licensing Examination (USMLE) is a prerequisite for obtaining a medical license, and the Bar Exam is a state-specific requirement to practice law.

These examinations differ substantially in their test-taking restrictions. The Indian and Chinese governments, for instance, generally allow candidates up to six attempts. The USMLE, a three-step examination, limits candidates to four attempts per step. Bar Exam regulations vary across states: New York and California impose no limits, whereas Iowa and Arizona allow at most three and two attempts, respectively.

A common justification for limiting attempts points to capacity constraints: testing institutions have limited resources and can therefore examine only a finite number of candidates in each period. Once an individual has failed repeatedly, it may be reasonable to consider their likelihood of eventual success negligible, and to allocate scarce testing capacity to more promising candidates instead.

This paper develops an alternative rationale for the existence of attempt limits. Standardized tests are subject to two kinds of errors: false positives, in which weak candidates pass, and false negatives, in which strong candidates fail. In the presence of false positives, weak candidates have an incentive to retake the test indefinitely when no limit exists, since each attempt offers a nonzero chance of success. Imposing an attempt limit introduces a strategic cost to test-taking: with only a finite number of opportunities, candidates choose to sit for the test only when they believe their probability of passing is sufficiently high, that is, when they are adequately prepared. In this sense, attempt limits can improve the average quality of successful candidates.

Consistent with this theory, I provide descriptive evidence that the introduction of attempt limits raises Bar Exam passing rates. Using data I collected across U.S. jurisdictions, I document that states enforcing an attempt limit achieve higher passing rates, measured as the share of candidates who pass the examination in a given session. The identification strategy relies on the institutional persistence of test rules. Since 2011, U.S. jurisdictions have transitioned from state-specific Bar Exams to the

Uniform Bar Exam (UBE), under which all candidates in participating jurisdictions are evaluated on identical materials using a common grading standard. Owing to institutional persistence, jurisdictions retained many of their preexisting regulations after adopting the new format; thus, within UBE states, candidates face the same test structure but are subject to different rules. As a result, institutional persistence generates policy variation on a uniform testing platform.

Because the National Conference of Bar Examiners reports separate statistics for first-time test takers and repeaters, I can observe and compare the performance of these two groups. After establishing the exogeneity of UBE adoption and controlling for state fixed effects, I find that jurisdictions enforcing attempt limits exhibit higher passing rates among repeaters. In contrast, the distribution of passing rates among first-time takers is unaffected by the existence of an attempt limit. This evidence aligns with the theory of the attempt limit as a sorting device: first-time test takers, who have multiple opportunities ahead of them, are unconstrained by the rule, whereas repeaters nearing the limit tend to sit for the exam only when adequately prepared.

When the likelihood of false negatives is high, however, strong candidates may fail several times and eventually exhaust their permitted attempts, leading to a permanent loss of human capital. In such environments, institutions have an incentive to relax or eliminate attempt limits to avoid excluding competent individuals.

This trade-off motivates the analysis that follows. Using the framework of mechanism design, I model attempt limits as a screening device that sorts candidates according to competence and characterize the conditions under which such restrictions are welfare improving. I interpret the attempt limit as one of several design parameters available to the testing institution, alongside the minimum passing score and the type of evidence the test is able to disclose. Because these dimensions are interdependent, I first analyze the attempt-limit decision holding other parameters fixed, and then extend the analysis to the joint design problem. This broader framework reveals when the optimal testing policy includes an attempt limit.

I begin the analysis by considering a principal-agent model with two periods: the agent (testee) is hired only by passing a test that is run in each of the two periods. There are two types of agents: weak and strong candidates. Strong candidates have higher chances than weak agents to pass the test. At the

beginning of the first period weak candidates, who can become competent in the next period if the test is not passed/taken, choose between two options: taking the test in the first session or wait one period before trying the test in the second session. The principal (tester), who prefers hiring strong candidates over weak candidates, decides whether or not to introduce a test restriction, keeping fixed the other test settings.

The first result is that the enforceability of the test depends on the relative return from one-period wait, $\Delta p_\theta/p_B$. The absolute return, $\Delta p_\theta = p_G - p_B$, captures the informativeness of the test—its ability to distinguish competence by giving the good type a higher probability of passing. The term p_B represents test difficulty: the higher p_B , the easier the test. The attempt limit is implementable if and only if the relative return is sufficiently high. If the candidate takes the test in the first period, they pass with probability p_B ; if they wait and become competent, their probability of passing increases by Δp_θ . The higher the relative return, the stronger the agent’s incentives to postpone taking the test by one period. The second result is that the attempt limit is optimally implemented if and only if the test difficulty is intermediate. It cannot be too easy, since p_B close to p_G would make the attempt limit non-implementable; nor can it be too difficult, since p_B close to zero would imply that the candidate fails regardless, making the attempt limit unnecessary. In this case, the benefits of the attempt limit—preventing bad candidates from being hired in the first period—would be limited compared to the cost of reducing the number of available attempts for good candidates.

A closely related result concerns the case in which weak candidates have a relatively high probability of passing the test. In this situation, the attempt limit would no longer be implementable. When the test is too easy, it cannot effectively distinguish between competent and incompetent candidates, and early testing provides little useful information to the principal. In such a context, the principal is better off by temporarily suspending testing and allowing all candidates to accumulate human capital before being evaluated. In practice, this means turning the first testing period into a mandatory training phase and shifting screening to a later stage, when candidates are more likely to be competent. Therefore, when the test is too lenient, the optimal policy is not to tighten attempt limits but to postpone evaluation altogether, transforming the testing institution into a two-stage process: training first, selection later.

If the principal is let to choose the minimum passing score, she can affect the probabilities of both

good and bad agents to pass the test by setting the minimum passing score. In fact candidate's performance in the exam can be measured by a test score, that is random variable whose distribution is type-dependent: good-agent test score distribution stochastically dominates the bad-agent one. In other terms, for any specified passing threshold, strong candidates have more chances to pass the test than weak ones, even though raising the passing threshold reduces the passing probabilities of both types of agents.

In this framework, a second key result is that the optimal choice of test restrictions, and in particular the adoption of an attempt limit, depends on the type of information the test reveals about candidates. To illustrate this intuition, I consider two benchmark cases: the good-certifying test and the bad-certifying test. The good-certifying case is depicted in Figure 1, while the bad-certifying case is shown in Figure 2.

The left panel of Figure 1 displays the test-score distributions of weak (red-shaded area) and strong (green-shaded area) candidates. In the good-certifying test, weak candidates' scores are uniformly distributed on the interval $[0,1]$, whereas strong candidates never score below 0.5. Hence, any individual scoring above 0.5 is certified to be competent. From the right panel of Figure 1, it can be noticed that the one-period relative return $\Delta p_\theta/p_B$ is increasing in σ , for $\sigma \in [0,0.5]$ and goes from 0 to infinite: indeed Δp_θ is zero at $\sigma = 0$ and increases over $[0,0.5]$, while p_B decreases from 1 to 0. For the attempt limit to serve as an effective deterrent, the passing threshold must be moderately high, though still below 0.5. When candidates are highly impatient, the threshold must be set sufficiently high to make waiting worthwhile. However, this also raises the failure rate among strong candidates, thereby increasing the risk of human-capital loss. Alternatively, the principal may set the threshold at $\sigma = 0.5$ and allow retakes. In this case, weak candidates never pass, while strong candidates retain the maximum number of attempts. Both policies prevent the hiring of weak candidates, but they do so through different mechanisms: the former regime relies on the no-retake threat, while latter policy relies on a difficult test. When principal's aversion to human-capital loss is high, she prefers setting a difficult test with retakes.

In the bad-certifying test, illustrated in Figure 2, weak candidates' test scores are uniformly distributed on the interval $[0,1]$, whereas strong candidates never score below 0.5. Consequently, any

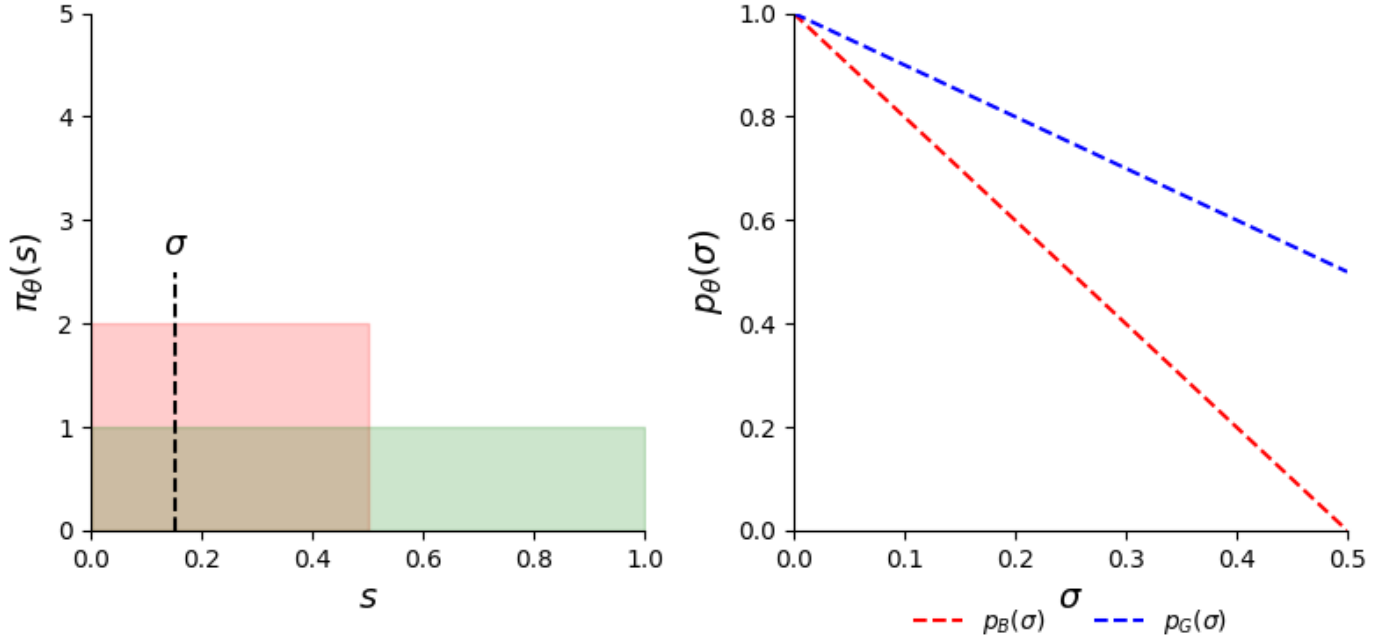


Figure 1: Good-certifying test

candidate scoring below this threshold is conclusively identified as weak. When the passing threshold σ equals one, all candidates fail the test. From the right panel of Figure 2, it's easy to notice that the one-period wait relative return is decreasing in σ over $[0.5, 1]$ and it has its maximum value at $\sigma = 0.5$. Thus, either the attempt limit is implementable at $\sigma = 0.5$ or is not implementable at all (which is equivalent to setting $\sigma = 1$, that is, failing everyone). If enforceable, the model assumptions make sure that the attempt limit is the most favorable configuration for the principal, who can postpone the selection of weak candidates while avoiding false negatives among strong ones. To conclude, given the relative return $\frac{\Delta p_\theta(0.5)}{p_B(0.5)}$, the attempt limit is the optimal policy if and only if the agent is patient enough.

In some environments, the information structure of the test, that is, the test signal, is exogenous, and the principal chooses only the optimal combination of the test restriction and the passing threshold. This is the case of the Uniform Bar Exam, where the test signal is centrally designed by the National Conference of Bar Examiners, while individual states independently determine the passing threshold and the number of permitted attempts.

More generally, it is natural to consider settings in which the principal can freely choose all aspects of the test design, including the test signal itself. I therefore assume that the principal can select

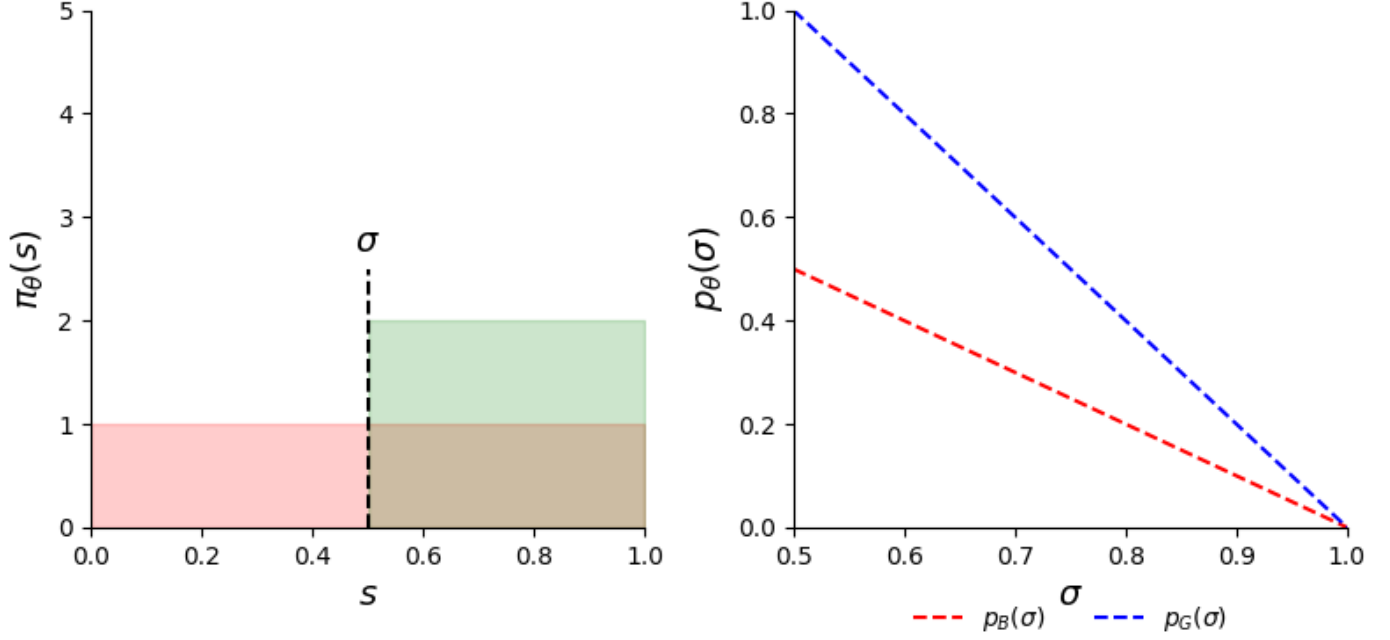


Figure 2: Bad-certifying test

any type of test i from $\{b, \tau, g\}$: they are respectively the bad-certifying test, the symmetric two-sided certifying test and the good-certifying test. They are represented in Figure 3 and are characterized by the following technological constraint: the overlap between the supports of the two conditional distributions is fixed at 0.5. Intuitively, this constraint captures the idea that the principal can trade off accuracy in identifying strong candidates against accuracy in identifying weak ones, but cannot achieve perfect classification of both types simultaneously. The two-sided certifying test represents a combination of the other two types of test, providing conclusive evidence for both types of agents. The third result is that the good-certifying test is the optimal test if both the principal and the agent are reluctant to one-period wait; otherwise the optimal test is the bad-certifying test. So the optimal test is either the good or the bad certifying test, but never the two-sided certifying test. To illustrate the intuition, I parametrize agent's reluctance to one-period wait by $\delta_A \in (0, 1)$: the higher is δ_A , the less reluctant the agent is. Also, I assume that the principal is an instant-utility maximizer; hence, if the competent agents are at least half - as I assume - of the cohort of candidates, it's possible to show that the optimal setting always employs the attempt limit. Therefore, given that the probability of the false positives is reduced to zero, the principal chooses the test signal with the highest $p_G(\hat{\sigma}_i)$, where $\hat{\sigma}_i$ is the passing

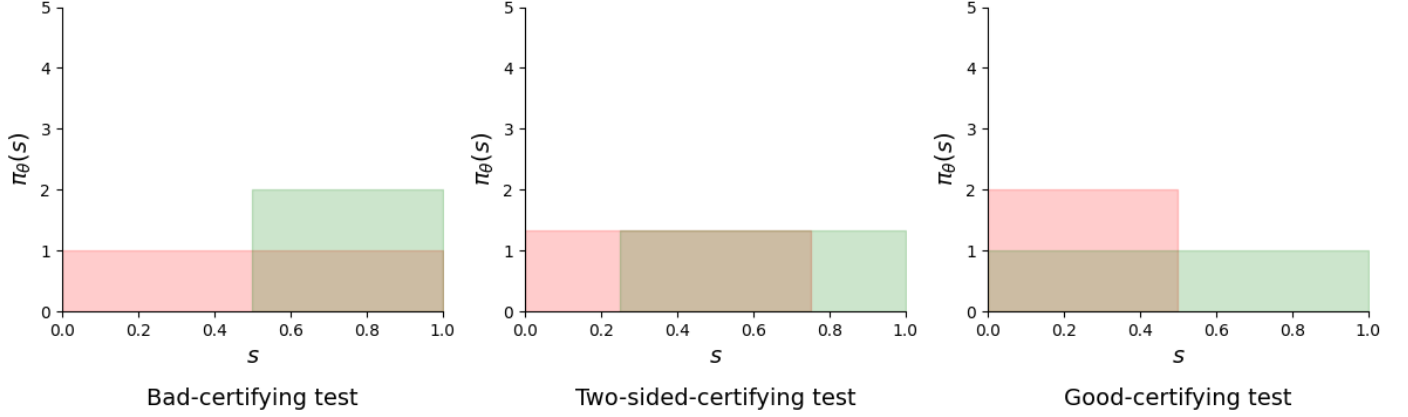
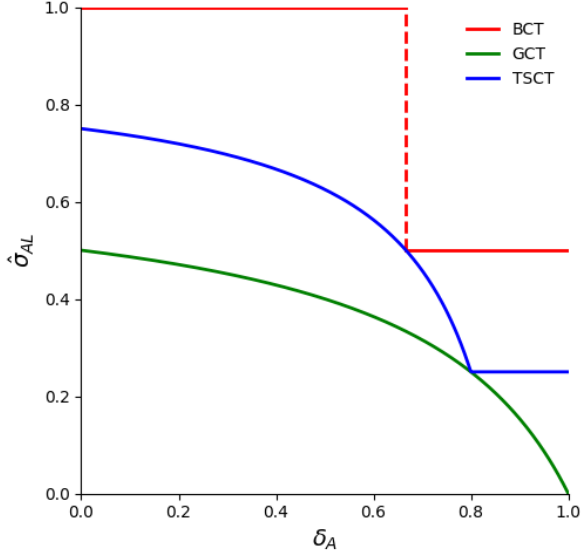


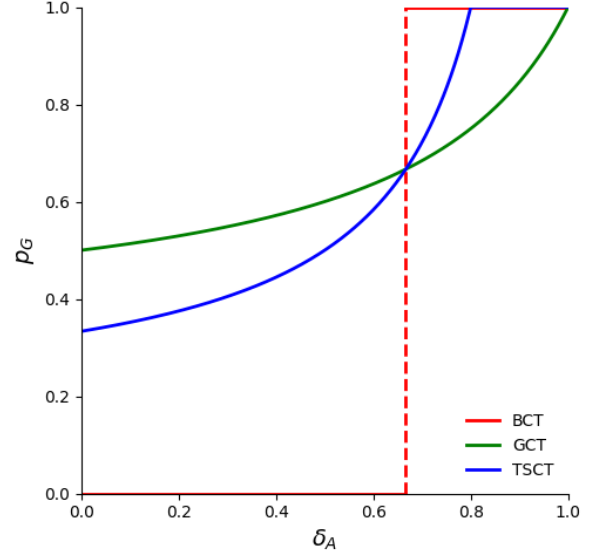
Figure 3: Types of test

threshold for the test signal i that makes $\frac{\Delta p_{\theta}^i(\sigma)}{p_B^i(\sigma)}$ large enough to convince the weak agent to wait one period. Thus, the principal chooses the bad-certifying test for $\delta_A \geq \frac{2}{3}$: the value $\frac{2}{3}$ is indeed the level of δ_A at which the agent is indifferent between taking the bad-certifying test with difficulty $p_B^b(0.5)$ and wait one period; for any $\delta_A \geq \frac{2}{3}$, Figure 4 shows that the passing probabilities for the good type are $p_G^b(0.5) = 1 \geq p_G^r(\hat{\sigma}_r) > p_G^g(\hat{\sigma}_g)$. This is because the bad-certifying test is the most accurate about the good-candidate's performance and the poorest performance of the strong candidate is still too high for the bad candidate who ultimately chooses to wait one period. If $\delta_A < \frac{2}{3}$, the attempt limit is implementable in the bad-certifying test by setting $\hat{\sigma}_b = 1$ implying $p_G^b(1) = 0$: thus, it is not chosen. Since $p_G^r(\sigma_r) < p_G^g(\sigma_g)$ for $\delta_A < \frac{2}{3}$, the principal chooses the good-certifying test. This is because good-certifying test signal is the most accurate about the bad-type agent's performance at the bottom of the test score range: therefore, a small passing threshold yields p_B very low while keeping p_G pretty high: it minimizes the probability of false negatives while keeping weak candidates out of the test session.

The remainder of the paper is organized as follows. Section 2 presents the empirical evidence on the Bar Exam and the effects of attempt-limit policies. Section 3 introduces the baseline theoretical model. Section 4 derives the main results concerning the optimal choice of an attempt limit when other test parameters are held fixed. Section 5 extends the analysis to the joint determination of the attempt limit and the passing threshold. Section 6 further generalizes the framework to the simultaneous design of the attempt limit, the passing threshold, and the test signal. Before proceeding, I review the related literature to which this paper contributes.



(a) Attempt Limit passing thresholds



(b) Good-type passing probabilities

Figure 4: Optimal test signal

Related Literature This paper contributes to a broad literature on the design of tests and screening mechanisms under information frictions. It combines evidence from the U.S. Bar Exam with a mechanism-design framework that interprets attempt limits as a policy instrument that shapes candidates' incentives and welfare. The analysis brings together several strands of research: the economics of education and standardized testing, the literature on college admissions and optimal test design, mechanism design with verifiable or partially verifiable information, knowledge screening and optimal test design, and multidimensional and dynamic screening.

A first line of related work lies in the economics of education and standardized testing. Research in this area has examined how institutional design affects human capital accumulation, equity, and efficiency. Studies on accountability systems and testing reforms, including Tyler (2020), Goodman (2021), and Angrist et al. (2023), show that institutional rules can significantly shape effort, learning incentives, and labor-market outcomes. Related work by Figlio and Loeb (2011), Dee and Jacob (2006), and Kane and Staiger (2002) documents how testing rules and accountability mechanisms affect achievement and selection. The empirical analysis in this paper complements this literature by identifying a causal relationship between attempt limits and test outcomes. Jurisdictions that restrict the number of retakes exhibit higher passing rates among repeaters, suggesting that attempt limits induce self-selection and improved preparation. This result contributes to a growing body of evidence linking policy design in

testing and licensing to behavioral sorting and to the informational efficiency of performance signals.

A closely connected strand examines college admissions and standardized testing as selection mechanisms under asymmetric information. This literature analyzes the role of test difficulty, grading rules, and admission thresholds in shaping selection, incentives, and fairness (Chade, Lewis, and Smith, 2011; Fryer and Loury, 2013; Banerjee, Dubey, and Tadelis, 2023). Recent theoretical work by Dessein, Frankel, and Kartik (2024) unifies these perspectives, showing that institutions may optimally choose less informative or coarser testing mechanisms to balance decision accuracy and incentive alignment. The present framework extends this logic to a dynamic setting in which candidates can learn, delay, and retake exams. By endogenizing both the number of attempts and the informativeness of the test signal, the analysis generalizes the static models of optimal admissions testing to environments where information about ability is revealed gradually over time.

Another body of research relevant to this analysis focuses on mechanism design with verifiable or partially verifiable information. Foundational contributions include Townsend (1979) and Gale and Hellwig (1985) on costly state verification, and Green and Laffont (1986), Forges and Koessler (2005), and Deneckere and Severinov (2008) on partial verifiability. Within this literature, Bull and Watson (2007) introduced the notion of hard evidence, while Glazer and Rubinstein (2004) and Carroll and Egorov (2019) analyzed mechanisms in which the principal can verify or audit an agent’s claims at a cost. In the present framework, the test itself serves as the verification technology, and the associated verification cost arises endogenously from the expected loss due to imperfect testing. When the principal allows a candidate to take the exam, she accepts a probabilistic form of verification: strong candidates may fail and weak candidates may pass. This misclassification risk constitutes an expected loss, which plays the same conceptual role as the verification cost in classical models. The principal can influence this expected loss through institutional design—by setting the number of attempts and the passing threshold—thereby controlling both the precision and the frequency of verification. The analysis also relates to recent work on mechanism design with evidence and persuasion, including Sher and Vohra (2015), Krämer and Strausz (2024), and Dasgupta, Krasikov, and Lamba (2022), in which the directionality of deviations—the ways in which agents can suppress or reveal information—determines optimal design. Similarly, attempt limits add an intertemporal dimension to disclosure: candidates reveal their readiness

not through direct reporting, but through the timing of their participation.

Insights from the emerging literature on knowledge screening and optimal test design are also closely related. Recent theoretical work by Dasgupta (2024) models tests as mechanisms for eliciting private information about knowledge or beliefs, showing that the optimal mechanism takes the form of a “pick-the-correct-answer” test with at most two thresholds. Dasgupta and Xia (2024) extend this framework to environments with verifiable evidence, demonstrating that when agents can hide but not fabricate information, the optimal test combines correctness with a minimal-evidence requirement, rewarding both accuracy and justification. The present analysis adds a dynamic dimension to this framework. Rather than designing the mapping from answers to outcomes within a single test, the institution designs the timing and frequency of tests. Attempt limits function as dynamic screening instruments: they separate weak from strong candidates by deterring early participation by the unprepared, while preserving the possibility of selection in later periods. This temporal mechanism parallels the static evidence thresholds in the knowledge-screening models but operates across test attempts rather than across levels of reasoning or justification. The logic also connects to research on optimal test informativeness in signaling environments, such as Rosar (2017), Harbaugh and Rasmusen (2018), Weksler and Zik (2022), and Hancart (2022), which show that less informative or coarser tests may be desirable when more precise evaluation induces strategic distortion. Similarly, limiting retake opportunities introduces coarseness in timing: by reducing the frequency of testing, institutions can increase the reliability of observed success rates.

A further connection arises with the literature on multidimensional and dynamic screening. In models where agents possess multiple private attributes—such as verifiable training and unverifiable talent—principals face a joint problem of verification and inference. Vravosinos (2025) analyzes this environment and shows that when verification can occur along several dimensions, the principal must balance precision across them: improving verification of one attribute limits the informativeness of the other. Related work by Egorov (2021) and Carroll and Egorov (2019) extends this logic to dynamic settings in which agents can gradually disclose verifiable evidence over time. These papers highlight how principals optimally restrict information revelation to manage incentives and maintain screening power. The present framework can be viewed as an institutional analogue of these mechanisms. Rather than

using continuous transfers or allocations as screening instruments, the principal here employs discrete institutional rules—attempt limits and thresholds—that determine when and how information about ability is revealed. This approach connects to dynamic mechanism design (Battaglini, 2005; Pavan, Segal, and Toikka, 2014), where timing and repeated opportunities act as endogenous screening devices. In this sense, attempt limits function as intertemporal participation constraints that regulate the timing and precision of information disclosure in standardized testing.

Finally, a broader conceptual parallel can be drawn with the literature on information design and Bayesian persuasion, in which a designer chooses a signal structure—the conditional distribution of messages given the true state—to influence receivers’ actions (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2016; Kolotilin et al., 2017). In persuasion models, the sender designs signals to alter the receiver’s posterior beliefs about an exogenous state. In contrast, the principal in this paper chooses the conditional distributions of test scores for strong and weak candidates through the institutional design of the testing process. These choices determine how informative the test is and how information about ability unfolds over time, but the objective is behavioral rather than informational: the principal does not seek to manipulate beliefs, only to regulate participation and selection through the structure of the testing environment. This perspective situates test design within the broader class of controlled information-release problems while highlighting its distinct institutional mechanism.

2 Motivating Evidence

This section aims to provide descriptive evidence that motivate the theoretical analysis by illustrating how bar-exam outcomes vary systematically with the presence of attempt limits. States with attempt limits tend to exhibit higher pass rates and sharper distinctions between first-time and repeat takers, providing motivation for the mechanisms developed in the theoretical analysis.

$$Attempt\ Limit \Rightarrow \uparrow\ Passing\ rate = \frac{\#Admitted}{\#Takers}$$

This pattern is consistent with the theory proposed in this paper, that is, the attempt limit working as a sorting device. By introducing a penalty for taking the test, jurisdictions would induce a self-selection

among candidates and only strong candidates show up to take the test, resulting in higher passing rates. More formally, assume a cohort of N candidates: a fraction of them $\nu = \frac{n}{N}$ are strong candidates, whereas the reminder is made up of weak candidates. Also, it seems natural to think that good students have higher chances to pass the test than bad candidates, $p_G > p_B$. Thus, if the attempt limit is able to narrow the population of test takers to the group of competent people, I would expect a higher passing rate in presence of an attempt limit (AL) than in absence (UT):

$$\begin{aligned} \text{Passing rate}_{AL} &= \frac{np_G}{n} = p_G > \nu p_G + (1 - \nu)p_B \\ &= \frac{np_G + (N - n)p_B}{N} = \text{Passing rate}_{UT} \end{aligned}$$

2.1 Background

The U.S. Bar Exam. The bar exam is the licensing test a law graduate (or equivalent) must pass in order to practice law in a given U.S. state or territory and it ensures a baseline competency in legal reasoning, ethics and lawyering tasks. In the United States, the practice of law is not regulated at the national level but rather by the individual states and territories. Each jurisdiction establishes its own requirements for admission to the bar. While each state has its own licensing authority and specific requirements, the central component of this admission process is the Bar Examination, a comprehensive test designed to assess whether candidates possess the knowledge and skills required to practice law competently and ethically. The overarching purpose of the bar exam is to ensure that all newly admitted attorneys meet a minimum standard of professional competence. The exam is not intended to identify exceptional lawyers or scholars, but rather to protect the public by screening out individuals who lack the basic legal knowledge and reasoning ability necessary for the responsible practice of law. Today, the bar exam is administered under the authority of each state's highest court or designated board of bar examiners. While the specific format and requirements vary across jurisdictions, most states rely heavily on materials developed by the National Conference of Bar Examiners (NCBE), a nonprofit organization that creates several standardized components used nationwide. This structure allows for a degree of uniformity while preserving state autonomy over admission policies. Although the precise format of the

exam differs by state, most jurisdictions employ some combination of the following NCBE components:

- ▶ **The Multistate Bar Examination (MBE):** a multiple-choice test consisting of 200 questions covering core legal subjects such as Constitutional Law, Contracts, Criminal Law and Procedure, Civil Procedure, Evidence, Real Property, and Torts. The MBE assesses general legal principles and analytical reasoning across fact patterns.
- ▶ **The Multistate Essay Examination (MEE):** a series of essay questions requiring candidates to analyze complex hypothetical scenarios and craft coherent, legally sound written responses. The MEE evaluates legal reasoning and the ability to apply abstract principles to realistic problems.
- ▶ **The Multistate Performance Test (MPT):** a practical exercise that simulates a lawyering task, such as drafting a memorandum, client letter, or legal brief, based on a file of factual materials and governing law. The MPT aims to test professional skills rather than rote knowledge.
- ▶ **The Multistate Professional Responsibility Examination (MPRE):** a separate two-hour test that assesses understanding of professional ethics and the rules governing attorney conduct. Nearly all states require a passing score on the MPRE in addition to the main bar exam.

In total, the exam typically spans two or three days and is offered twice yearly, in February and July. Although the bar exam has a broadly similar structure across the United States, the way in which results are evaluated, combined, and interpreted varies substantially among jurisdictions. Most jurisdictions assign separate numerical scores to three main components of the examination—the multiple-choice section, the essay section, and the performance test—and then compute a composite score. While the multiple-choice portion typically accounts for around half of the total, the precise weighting scheme is left to each jurisdiction. Another area of variation concerns how many times a candidate is allowed to attempt the bar exam. Most jurisdictions permit multiple attempts. Fees for registering and sitting for the bar exam also vary widely. Application fees, though relatively modest in comparison to total legal education costs, symbolize the decentralized and market-driven nature of U.S. professional licensing.

The Uniform Bar Exam. The Uniform Bar Examination (UBE) represents one of the most significant reforms in the modern history of legal licensing in the United States. Developed and coordinated by the

National Conference of Bar Examiners (NCBE), the UBE was created to establish a more consistent and portable system of assessing professional competence among aspiring lawyers.

Before the introduction of the UBE in 2011, each state designed and graded its own bar examination, often including both national and locally drafted components. This fragmentation produced wide disparities in format, difficulty, and scoring standards, making it challenging for lawyers to transfer their credentials from one jurisdiction to another. In an increasingly mobile and interconnected legal market, such barriers appeared increasingly anachronistic. The NCBE proposed the UBE as a means of harmonizing assessment across state lines while preserving local control over licensing decisions. The central idea was that, if all jurisdictions measured competence using the same standardized components and scoring methods, the resulting scores could be mutually recognized—allowing lawyers greater mobility and simplifying the process of admission to multiple bars. The UBE thus embodies an attempt to reconcile the historical principle of state autonomy with the practical need for national coherence in professional standards.

The UBE is composed of three examinations that had already been developed and administered by the NCBE on a national basis. The innovation of the UBE was not in inventing new tests, but in combining existing ones under a common structure, weighting, and grading methodology. The multistate bar examination, the multistate essay examination and the multistate performance test are administered uniformly across participating jurisdictions over a two-day period. The MBE typically constitutes 50 percent of the total score, the MEE 30 percent, and the MPT 20 percent. Scores are expressed on a 400-point scale, and each jurisdiction determines its own cut score for passing.

The hallmark feature of the Uniform Bar Examination is score portability. A candidate who earns a passing score in one UBE jurisdiction may transfer that score to another UBE jurisdiction, subject to local eligibility requirements and time limits. For instance, a lawyer who passes the bar in Colorado may seek admission in New York or the District of Columbia without retaking the entire examination, provided the score meets the recipient jurisdiction's minimum standard and remains valid (usually for three to five years).

Although the UBE standardizes the exam content and relative weights of its components, each jurisdiction retains discretion over scoring and passing standards. Scores are calculated using scaled

and equated methods to ensure comparability across test administrations. The multiple-choice section (MBE) is standardized nationally, providing an anchor for the scaling of essay and performance test results. Each state sets its own minimum passing score, which typically ranges from 260 to 280 on the 400-point scale. This variation reflects differing philosophies of professional gatekeeping: jurisdictions with lower thresholds prioritize access and mobility, whereas those with higher cut scores emphasize rigorous quality control. Despite these differences, the use of uniform testing materials ensures that a score earned in one jurisdiction has a consistent interpretive meaning across all others.

Since its first administration in 2011, the UBE has been adopted by the vast majority of U.S. jurisdictions. Today, over forty states ¹ territories, and the District of Columbia administer the UBE, encompassing a broad cross-section of the American legal landscape—from large, competitive markets such as New York, Illinois, and Washington, D.C., to smaller states such as Wyoming and Vermont. A few jurisdictions, notably California and Florida, continue to administer their own examinations. While the UBE provides a common framework, jurisdictions may impose supplemental requirements to preserve their distinctive legal traditions and professional standards. Many states require candidates to complete a brief course or online module on local law and procedure before admission. Others require separate character and fitness evaluations, oaths of office, or continuing legal education obligations. These additions allow states to tailor the licensing process to local needs without fragmenting the core assessment of legal competence.

2.2 Data

My analysis combines two categories of variables. The first group, examination statistics, includes measures of participation and performance, such as the number of candidates taking and passing the bar examination. The second group, test characteristics, captures institutional aspect of the test setting including format and grading policies.

¹Source: image in Figure 5 is reproduced from the official website of the National Conference of Bar Examiners (NCBE).

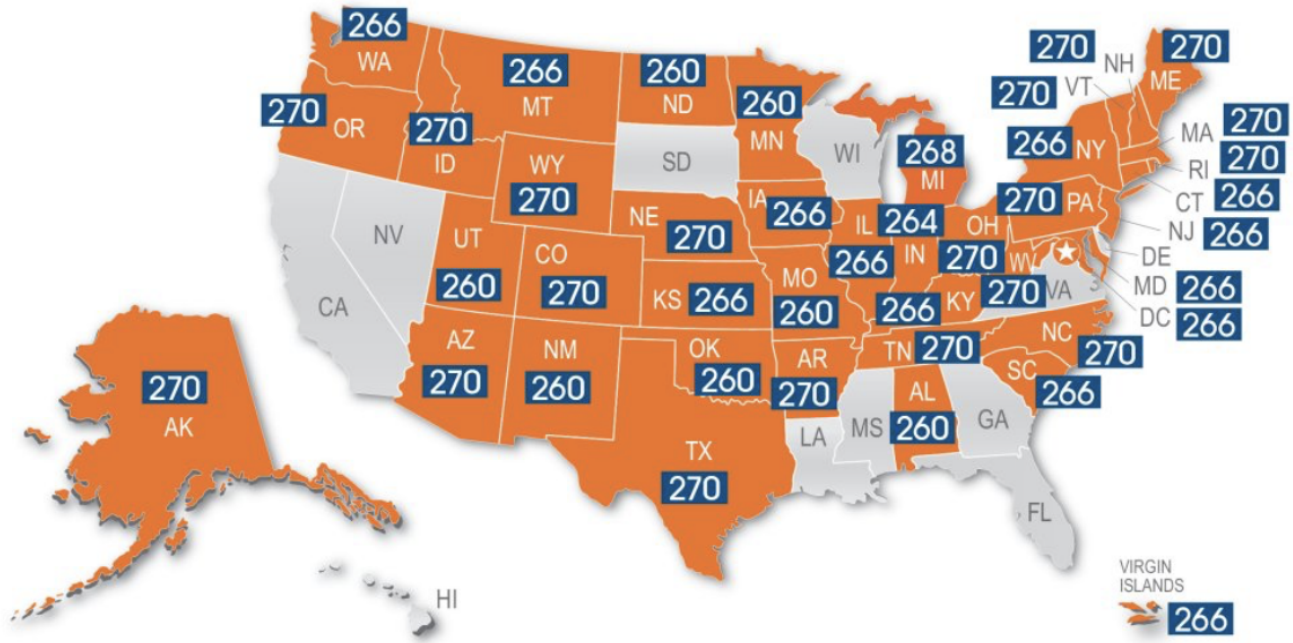


Figure 5: Current UBE jurisdiction and passing scores

2.2.1 Examination Statistics

The analysis relies on administrative data on bar examination participation and performance collected by the National Conference of Bar Examiners (NCBE). The dataset provides information for each jurisdiction, year, and test session, reporting both the number of examinees and the number of successful candidates. The outcome variable of interest is the passing rate, defined as the percentage of test-takers passing the test in a given session.

Crucially, the NCBE distinguishes examinees by their testing experience, dividing candidates into two categories: first-time test takers and repeat test takers. Accordingly, for every jurisdiction, year, and session, I observe both the number of first-time examinees and repeaters, together with the corresponding number of candidates in each group who passed the examination. This structure enables an analysis of performance dynamics across candidate types and over time. The unit of analysis is the jurisdiction–year–session–testing experience combination. In other words, for each jurisdiction and test session within a given year, the dataset includes separate observations for first-time and repeat test

takers. Thus the outcome variable is

$$PassRate_{jtse} = \frac{Passing_{jtse}}{Takers_{jtse}}$$

The sample spans the period 2007–2024, covering nearly two decades of bar examination data under relatively stable institutional conditions. The analysis focuses on U.S. states only, excluding U.S. territories and associated jurisdictions. In particular, I omit Guam, the Northern Mariana Islands, Palau, Puerto Rico, and the U.S. Virgin Islands from the sample. This restriction ensures consistency in institutional context, given that territorial bar examinations often follow distinct rules, administrative structures, and eligibility criteria that differ from those of the fifty states.

A caveat applies to the classification of testing experience. As noted by the NCBE, the designation of “first-time” and “repeat” examinees refers solely to candidates’ prior attempts within the same reporting jurisdiction. The data do not capture attempts made in other jurisdictions. Consequently, some “first-time” examinees in a given state may, in fact, have previously sat for the bar examination elsewhere. This limitation should be kept in mind when interpreting differences in pass rates between first-time and repeat candidates.

2.2.2 Test Characteristics

In addition to information on participation and performance, I compile a complementary dataset describing the institutional and administrative characteristics of the bar examination across jurisdictions, years, and sessions. These variables capture heterogeneity in exam design, grading standards, and costs, which may influence both participation and pass rates.

Scaled Passing Score. The first structural feature is the scaled passing score, which represents the minimum threshold required to pass the bar examination. This variable is invariant across testing-experience categories. Because jurisdictions employed different scoring systems prior to the adoption of the Uniform Bar Examination (UBE), I rescale all pre-UBE scores to a standardized 400-point scale. This harmonization ensures comparability and consistency across states and years, allowing the passing score to serve as a uniform measure of exam difficulty both before and after UBE adoption.

Maximum Allowed Attempts. The second structural variable is the Maximum Allowed Attempts, which defines the total number of times candidates with a given level of testing experience are permitted to sit for the bar examination in a specific jurisdiction and year. This variable changes across jurisdictions, years, and testing-experience categories. For instance, if jurisdiction j sets a limit of four attempts for first-time takers, it implies that repeaters in that same jurisdiction and year may sit for the exam at most three times. This variable therefore captures the effective number of remaining opportunities available to candidates of different experience levels.

Attempt Limit Indicator. I also construct an Attempt Limit indicator variable that takes the value 1 if a jurisdiction specifies a limit on the number of attempts allowed for first-time test takers, and 0 otherwise. This indicator varies by jurisdiction and year, reflecting whether a state imposes any explicit restriction on the number of times a first-time candidate may attempt the exam. Taken together, the Maximum Allowed Attempts variable and this indicator capture both the existence and stringency of attempt restrictions across states and over time.

UBE Adoption. The dataset also records the adoption of the Uniform Bar Examination (UBE) for each jurisdiction. The UBE indicator variable equals 1 if jurisdiction j administers the UBE format in year t and session s , and 0 otherwise. This variable does not vary across testing-experience categories within the same jurisdiction-year-session cell. Tracking UBE adoption allows identification of major policy shifts in exam format and enables the analysis of their implications for candidate performance. The UBE introduced standardized content, scoring procedures, and score portability, substantially transforming both the structure of the exam and the strategic environment faced by candidates.

Application Fees. Finally, I collect detailed data on application fees, which represent the direct financial cost of taking the bar examination. This variable varies by jurisdiction, year, session, and testing-experience category. In many jurisdictions, repeaters are charged lower fees than first-time takers, and fees often differ between the February and July sessions. This structure allows me to capture both temporal and cross-jurisdiction variation in exam costs, as well as systematic differences in pricing by experience level and session timing.

Institutional Stickiness Variables. To capture policy persistence in test-setting behavior, I construct variables that reflect the degree to which jurisdictions retain their pre-UBE rules after adopting the standardized format. For each test-setting feature—the passing score, maximum allowed attempts, and application fee—I define a variable X_{jtse} representing the last rule in effect before UBE adoption in jurisdiction j , session s , and experience group e . These variables measure how much jurisdictions rely on historical baselines when transitioning into the UBE system, providing an empirical proxy for institutional stickiness.

Benchmark Variables. To capture policy diffusion and peer imitation, I construct benchmark variables that describe how jurisdictions adjust their test settings relative to their peers. Two benchmarks are geographical, based on the Census region and Census division to which each jurisdiction belongs. For a given feature X and jurisdiction j , I compute the average rule among all other jurisdictions in the same region or division, excluding j , in year t , session s , and experience group e :

$$\bar{X}_{jtse}^{\text{geo}} = \frac{1}{|C(j)| - 1} \sum_{i \in C(j) \setminus \{j\}} X_{itse},$$

where $|C(j)|$ is the total number of jurisdictions in the Census region or division $C(j)$ that jurisdiction j belongs to. The third benchmark is institutional, based on UBE membership. For each test feature X and jurisdiction j , I compute the average rule among other jurisdictions that have already adopted the UBE, excluding j :

$$\bar{X}_{jtse}^{\text{UBE}} = \frac{1}{|U_{ts}| - 1} \sum_{i \in U_{ts} \setminus \{j\}} X_{itse},$$

where $|U_{ts}|$ is the number of UBE jurisdictions in session s and year t . Unlike the geographical benchmarks, which remain fixed over time, the UBE benchmark evolves dynamically as new jurisdictions join the UBE, reflecting the spread of institutional convergence across states.

2.3 Empirical Strategy

2.3.1 Fixed-Effects Model

To estimate the effect of the attempt limit on bar exam pass rates, I begin with the following fixed-effects specification:

$$\begin{aligned} PassRate_{jtse} = & \alpha_j + \alpha_t + \alpha_s + \alpha_e + \beta_1(AL_{jts} \times R_e) + \beta_2 AL_{jts} \\ & + \beta_3 PassScore_{jts} + \beta_4 fee_{jtse} + \beta_5 covid_{ts} + \varepsilon_{jtse}, \end{aligned} \quad (1)$$

where $PassRate_{jtse}$ denotes the passing rate of candidates with testing experience e who took the exam in jurisdiction j , year t , and session s . The existence of an attempt limit is captured by an indicator variable equal to one if a jurisdiction enforces a limit on the number of attempts, and zero otherwise (AL_{jts}). The minimum passing score required to pass the exam in jurisdiction j , year t , and session s is denoted by $PassScore_{jts}$. The variable fee_{jtse} measures the application fee charged to a candidate with testing experience e sitting for the exam in jurisdiction j , year t , and session s . The indicator R_e equals one for repeaters and zero for first-time takers. A control variable, $covid_{ts}$, equals one for the July 2020 and February 2021 sessions—when the COVID-19 pandemic disrupted exam administration—and zero otherwise. Finally, α_j , α_t , α_s , and α_e represent jurisdiction, year, session, and testing-experience fixed effects, respectively.

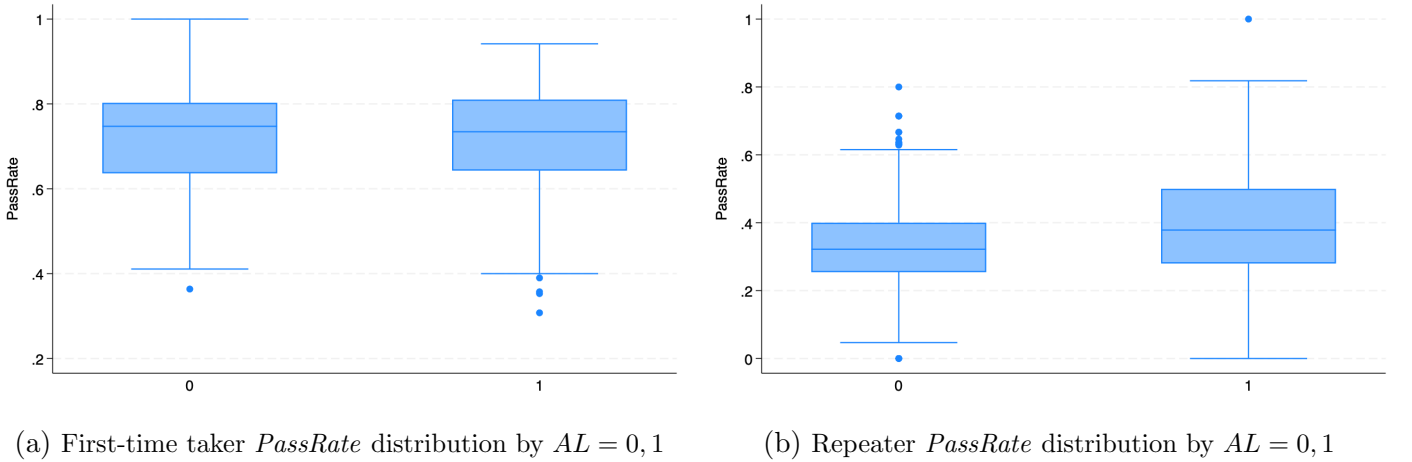


Figure 6: PassRate distribution by testing experience

The identification strategy relies on the interaction between the presence of an attempt limit and candidates’ testing experience. Because examinees are typically allowed multiple attempts, first-time takers are unaffected by the limit and tend to sit for the exam regardless of difficulty, whereas repeaters—being closer to the limit—are more responsive to it and take the exam only when sufficiently prepared. Accordingly, the expected sign of the interaction coefficient is positive, $\beta_1 > 0$. This is evident from Figure 6, where the distribution of passing rates among first-time test takers appears similar regardless of the existence of an attempt limit. By contrast, the passing rates of repeaters in jurisdictions with an attempt limit are, on average, higher than those of repeaters in states without any such limit.

2.3.2 Addressing Endogeneity

The specification introduced in the previous section may be subject to *endogeneity concerns*, as test settings are not randomly assigned but are determined by the jurisdictions themselves. Jurisdictions may adjust exam features in anticipation of future pass rates, thereby creating a correlation between policy choices and unobserved factors affecting performance. This concern does not apply uniformly to all variables—for example, application fees are likely influenced by broader market or administrative considerations rather than by expected pass rates.

To address potential endogeneity, I estimate a two-stage least squares (2SLS) model that exploits institutional features underlying how jurisdictions determine their exam settings. First, when jurisdictions adopt the UBE format, they typically retain many of the rules from their previous examination system. I therefore use these pre-UBE rules as instruments to predict the corresponding post-UBE test settings. Second, jurisdictions often display peer effects in rule-making, benchmarking their policies against those of neighboring or comparable jurisdictions—either geographically or institutionally (for example, other UBE states). Accordingly, I use the average rule adopted by neighboring or peer jurisdictions as an additional source of exogenous variation. Although the use of benchmark variables does not fully eliminate the risk of endogeneity, it offers two advantages: it helps mitigate potential bias arising from endogenous regressors and introduces additional variation that strengthens the identification strategy.

$$PassRate_{jtse} = \alpha_j + \alpha_t + \alpha_s + \alpha_e + \beta_1(\widehat{AL_{jts} \times R_e}) + \sum_{k=2}^4 \beta_k \widehat{X_{k,jtse}} + \varepsilon_{jtse},$$

$$X_{k,jtse} = \alpha_j + \alpha_t + \alpha_s + \alpha_e + \gamma_1 Z_{1,jse} + \gamma_2 Z_{2,jste} + \gamma_3 Z_{3,jste} + u_{jtse},$$

where the outcome variable $X_{k,jtse}$ in the second equation is the k test feature regressor appearing in equation (1) as well as the interaction term $AL_{jts} \times R_e$; instruments $Z_{1,jse}$ denote the test-setting variables derived from the pre-UBE rules in jurisdiction j for session s and experience group e ; instruments $Z_{2,jste}$ capture the geographical and institutional benchmark rules for candidates with experience e , in jurisdiction j , session s , and year t ; instrumental variables $Z_{3,jste}$ include additional control and interaction variables.

2.3.3 Empirical Validation of the Identification Strategy

To validate the identification strategy underlying the 2SLS specification, I conduct two complementary exercises. First, I examine whether the timing of UBE adoption is exogenous to bar exam pass rates. If jurisdictions adopted the UBE in response to prior increases or declines in pass rates, the adoption decision would be correlated with unobserved determinants of exam performance, thus undermining the causal interpretation of the 2SLS estimates. To evaluate this possibility, I estimate an event-study regression of pass rates around the year of UBE adoption. The absence of systematic pre-trends in pass rates would indicate that adoption timing was not driven by prior changes in exam performance, supporting the exogeneity of the UBE switch.

Second, I test whether exam rules changed before UBE adoption, which is central to the “institutional retaining” assumption underpinning the 2SLS framework. If jurisdictions had begun modifying their test settings in anticipation of the transition, the pre-UBE rules—used as instruments for post-UBE test settings—would be endogenous, violating the exclusion restriction. To assess this, I estimate separate event-study regressions for the main test-setting variables—the number of allowed attempts, the passing score, and the application fees. Stable rule configurations prior to adoption would confirm that jurisdictions did not adjust their policies in advance of the switch, validating the use of pre-UBE

rules as exogenous instruments for post-UBE test settings.

$$Y_{jtse} = \alpha_j + \alpha_t + \alpha_s + \alpha_e + \sum_{k=-6}^6 \delta_k D_{jts}^k + covid_{ts} + \varepsilon_{jtse}$$

where Y_{jtse} denotes the variable whose exogeneity with respect to the UBE format is being tested; $\alpha_j, \alpha_t, \alpha_s$ and α_e are jurisdiction, year, session, and test-experience fixed effects; and D_{jts}^k is an event-time dummy (in session units) equal to 1 if jurisdiction j is k test sessions away from its adoption of the UBE format.

Table 1: Regression Results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	<i>PassRate</i>	<i>PassRate</i>	<i>PassRate</i>	<i>PassScore</i>	<i>Tightness</i>	<i>AL</i>	<i>fee</i>
<i>ALxR</i>	0.0524*** (0.0177)	0.0592** (0.0268)					
<i>AL</i>	-0.00488 (0.0186)	0.00893 (0.0290)					
<i>PassScore</i>	-0.0116** (0.00574)	0.00794 (0.00747)					
<i>fee</i>	-0.0000434 (0.0000536)	-0.0000448 (0.0000539)					
<i>covid</i>	0.0741*** (0.0121)	0.0761*** (0.0121)	0.0687*** (0.00899)	-0.131* (0.0701)	-0.0170 (0.0163)	-0.00334 (0.00335)	3.170 (1.693)
<i>D₋₆</i>			-0.00867 (0.0125)	-0.0451 (0.117)	-0.00616 (0.0102)	-0.0343 (0.0349)	4.172 (14.34)
<i>D₋₅</i>			0.00589 (0.0111)	-0.0258 (0.123)	-0.00592 (0.0104)	-0.0368 (0.0360)	12.18 (15.57)
<i>D₋₄</i>			-0.0111 (0.0133)	-0.00548 (0.130)	-0.00636 (0.0107)	-0.0420 (0.0377)	15.58 (15.34)
<i>D₋₃</i>			0.00790 (0.0127)	0.00794 (0.133)	-0.00557 (0.0105)	-0.0407 (0.0369)	23.74 (15.86)
<i>D₋₂</i>			0.00259 (0.0111)	0.0215 (0.136)	-0.00894 (0.00934)	-0.0399 (0.0371)	15.67 (15.53)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	<i>PassRate</i>	<i>PassRate</i>	<i>PassRate</i>	<i>PassScore</i>	<i>Tightness</i>	<i>AL</i>	<i>fee</i>
D_0			0.0135 (0.0155)	-0.108 (0.184)	-0.00943 (0.00897)	-0.0486 (0.0375)	40.80** (17.62)
D_{+1}			-0.0193 (0.0121)	-0.0776 (0.181)	-0.0186 (0.0121)	-0.0781* (0.0455)	41.37** (16.81)
D_{+2}			-0.0153 (0.0114)	-0.0406 (0.175)	-0.0191 (0.0121)	-0.0805 * (0.0466)	45.39** (18.26)
D_{+3}			0.00378 (0.0148)	-0.0524 (0.179)	-0.0250** (0.0122)	-0.105** (0.0474)	41.01** (17.75)
D_{+4}			0.00527 (0.0145)	-0.0720 (0.179)	-0.0253** (0.0126)	-0.108** (0.0491)	42.89** (16.84)
D_{+5}			-0.0107 (0.0121)	-0.120 (0.188)	-0.0197 (0.0126)	-0.0854 * (0.0497)	35.83** (16.55)
D_{+6}			0.00778 (0.0127)	-0.227 (0.219)	-0.0155 (0.0110)	-0.0606 (0.0376)	32.23** (15.71)
N	1336	1336	3634	3634	3634	3634	3634

*Note: The first column reports regression results of the first specification. Column (2) reports second stage regression results of the second specification. Columns (3)-(7) show the event study results. Standard errors are clustered at state level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$*

2.4 Results

Column (1) of Table 1 reports the estimated β coefficients from the baseline specification, and the results are consistent with Figure 6. The coefficient on the interaction between the attempt-limit indicator and the dummy variable for repeaters is positive and statistically significant, confirming the intuition that attempt limits help increase pass rates. Column (2) presents the second-stage coefficients from the 2SLS specification, which point in the same direction. Interestingly, the coefficients on the application fees are not significantly different from zero, indicating that exam fees have no measurable impact on pass rates. This finding supports the rationale behind the attempt limits: they work as a screening device in contexts where monetary transfers are unavailable to sort candidates.

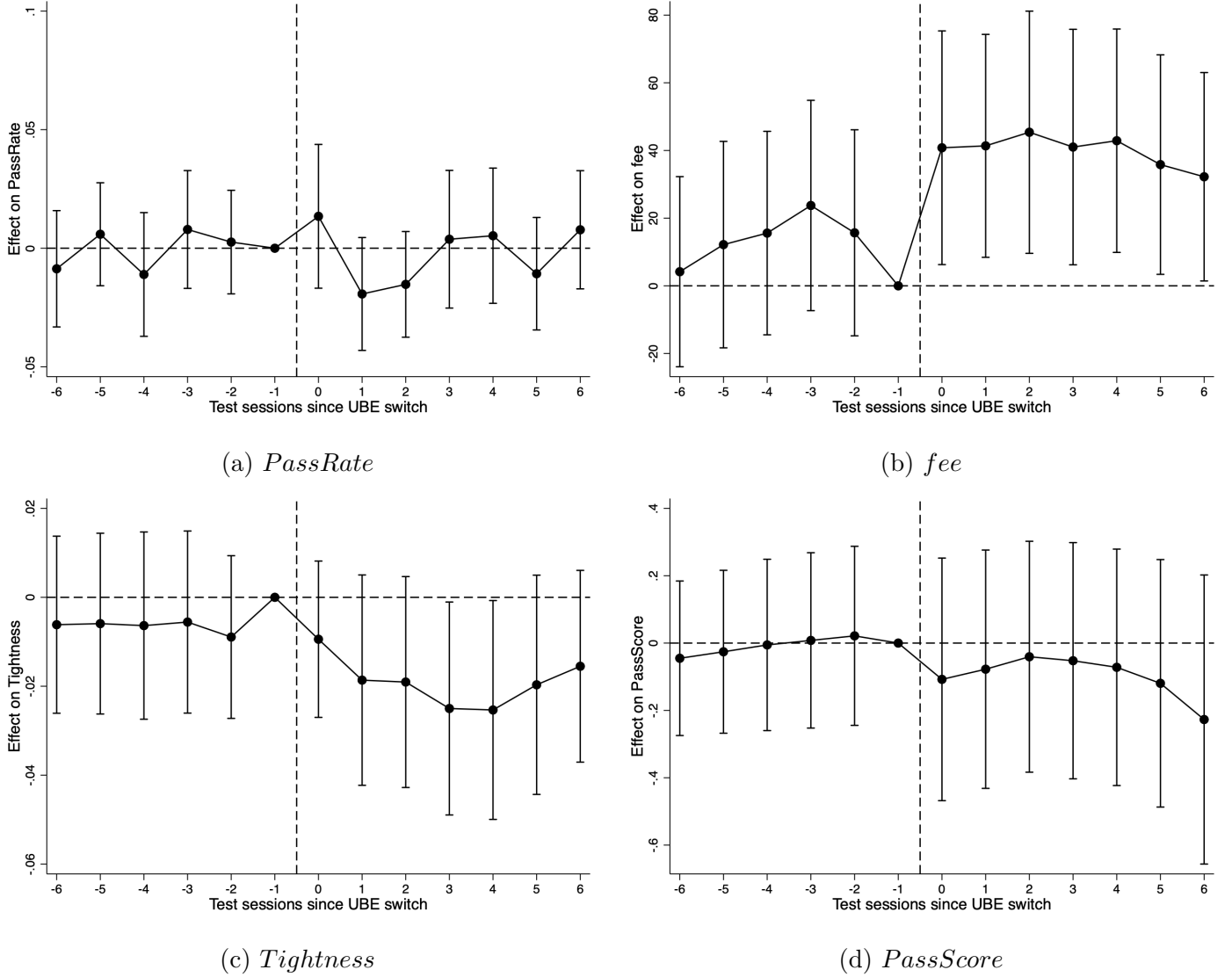


Figure 7: Event study results

Note: Figures plot main event study results. Error bars report the 95% CI of the estimates. Standard errors are clustered at state level.

Finally, it is worth discussing the diagnostic results. Column (3) shows that the transition to the UBE format is exogenous to pass rates, which is clearly illustrated by the top left panel of Figure 7. Even more interesting are the results concerning test features including limit tightness, that takes value 0 if there is not attempt limit and $\frac{1}{\#Attempts}$ otherwise. Thus, test settings do not change before jurisdiction's switch to the UBE, and any observed adjustment in the variables of interest occurs only after the adoption of the new test format. The other three panels in Figure 7 highlight that there is no substantial change in application fees, passing scores and limit tightness before the UBE adoption. All the changes, if any, happen after switching to the common test format. Fees change dramatically

after adopting the UBE format, the change in *PassScore* and *Titghness* is less marked. This pattern helps validate the first specification because it supports the hypothesis that states modify their exam rules only in response to the new format. Furthermore they illustrate the institutional persistence of jurisdictions motivating the second specification.

The empirical results from the UBE highlight that standardized testing institutions are multi-dimensional: attempt limits, passing thresholds, and other structural features jointly shape who takes the test, when they take it, and how well the test identifies competence. Motivated by this evidence, I now develop a theoretical framework that formalizes these interactions and characterizes the optimal design of test restrictions in combination with other elements of the testing environment.

3 General Model

Consider a principal-agent model with two periods: the agent is hired only by passing a test administered by the principal in each period. In particular, there are two types of agents, $\theta \in \{B, G\}$, and the principal prefers G over B . Since type θ is agent's private information, the principal resorts to a test to select candidates. However, the test makes use of an imperfect technology, so that the good agent could fail the test and the bad agent could pass it. If the agent passes the test, he gets unitary wage for infinite periods discounted by $\delta_A \in (0, 1)$: so the payoff of the agent succeeding in period $t = 1, 2$, is $\frac{\delta_A^{t-1}}{1-\delta_A}$; if he fails the test over the two periods, he gets nothing. Principal's payoff depends on the test outcome and the agent's type: if type θ agent passes the test, the principal obtains utility u_θ for infinite periods discounted by $\delta_P \in (0, 1)$ and, if the test is never passed, principal does not get anything. So if type θ agent passes the test in period $t = 1, 2$, principal's payoff is $u_\theta \frac{\delta_P^{t-1}}{1-\delta_P}$; otherwise, principal's payoff is normalized to 0. Here, I assume symmetric payoffs: $u_G = -u_B = 1$. Agent's types are equally likely in the first period. If the good agent doesn't pass the test in the first period, he remains competent. By contrast, if the bad agent doesn't pass the test in the first period, he becomes competent with probability $\frac{1}{2}$. The probability of passing the test is type-dependent: $p_G = Pr(Pass|\theta = G) \geq Pr(Pass|\theta = B) = p_B$. Parameter p_B is a measure of test difficulty: the higher is p_B , the easier is the test. A test is considered *difficult* if $p_B = 0$. In every period the agent decides whether or not to take the test, based on test settings established by the principal. Thus the timing of the events is

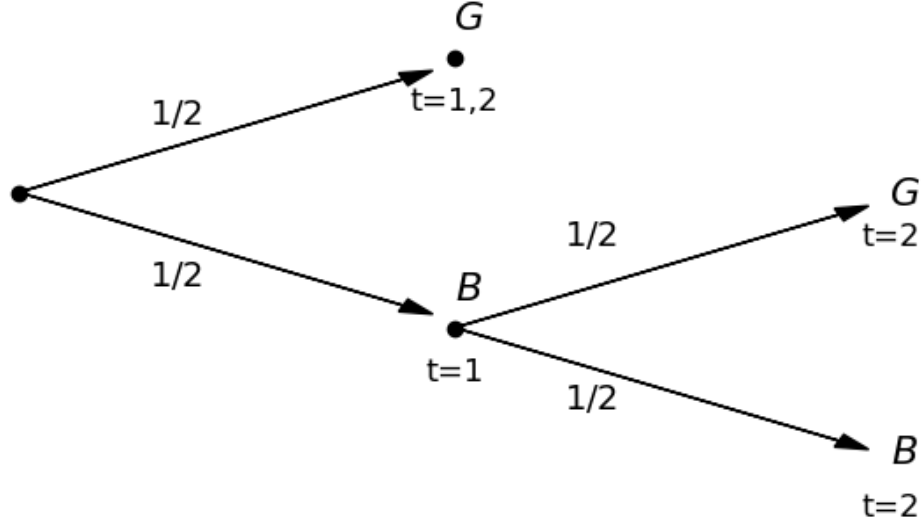


Figure 8: Type distribution over time

1. Principal sets test settings (ξ, σ, π) ;
2. Competence level $\theta \in \{B, G\}$ is privately observed by the agent;
3. Agent decides whether or not to take the test;
4. Test outcome is realized, $\omega \in \{P, F\}$. If $\omega = P$, the game ends; if $\omega = F$, a new test session might take place, depending on test settings.

Test settings are represented by the triple (ξ, σ, π) . Test restriction ξ determines when test (re)taking is allowed, the minimum passing score σ is the score required for the test to be passed, while π indicates the kind of evidence the test provides about candidates. In the next section I start analyzing ξ , while keeping σ and π fixed.

4 Optimal test restrictions

In this section I formally introduce test restrictions. They can be modeled as a direct mechanism in which the testee - the agent - is asked to report his competence level θ . If the testee reports that his

type is θ , the tester - the principal - commits to allowing or not allowing him to take the test in period 1, $\xi_\theta^1 \in \{0, 1\}$, in exchange for allowing or denying test taking in period 2, $\xi_\theta^2 \in \{0, 1\}$. In other terms, test restrictions can be viewed as a contract menu $\xi = \{\xi_\theta\}_{\theta=B,G}$, where each contract $\xi_\theta = (\xi_\theta^1, \xi_\theta^2)$ specifies if the agent - whose initial competence level is θ - is allowed to take the test in period t , $\xi_\theta^t \in \{0, 1\}$. Therefore, the timing of the events can be rephrased as follows:

1. Principal offers contract menu $\xi = \{\xi_\theta\}_{\theta=B,G}$;
2. Competence level $\theta \in \{B, G\}$ is privately observed by the agent;
3. Agent report his type;
4. Test outcome is realized, $\omega \in \{P, F\}$. If $\omega = P$, the game ends; if $\omega = F$, a new test session might take place, depending on test settings.

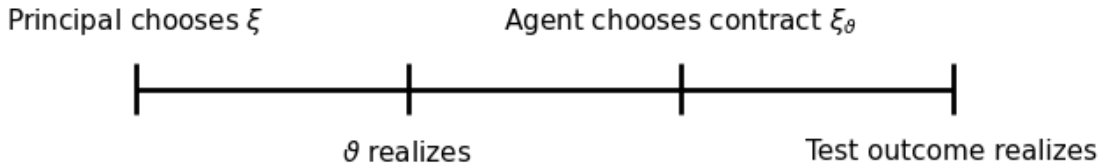


Figure 9: Timing of events

Depending on the type of contract menu offered by the principal to the agent, there are several types of test restrictions. Here, I focus on three types of test restrictions.

Definition 1. *Test restriction $\xi = (\xi_B^1, \xi_B^2, \xi_G^1, \xi_G^2)$ is a contract menu that can be defined by the contracts offered to the agent. Here, three types of test restrictions are defined, $\xi \in \{AL, UT, MT\}$:*

► *Attempt Limit, (AL):*

$$\xi = (0, 1, 1, 0)$$

► *Unrestricted Testing, (UT):*

$$\xi = (1, 1, 1, 1)$$

► *Mandatory Training, (MT):*

$$\xi = (0, 1, 0, 1)$$

When attempt limit is the test restriction selected by the principal, no retakes are allowed: agents are free to choose the test session but they can take the test at most once. This would avoid to hire someone incompetent in the first period, who can become competent in the next one. However competent agents failing the test cannot retake it and this would represent a human capital loss. If the test restriction chosen is unrestricted testing, agents are free to take the test at any time. The benefit of such a policy is to have more chances to hire a competent agent, but the other side of the coin is to be more likely to hire someone incompetent in the first period. Finally, mandatory training restriction always forces the agent to wait one period and take the test at the end of the game: this would allow for the principal to be more likely to hire someone competent at the end of the game, at the cost of period one expected utility.

Unrestricted testing and mandatory training are pooling contract menus: the agent gets the same contract, regardless of the type reported. By contrast, attempt limit is a separating restriction, whereby the contract assigned depends on the competence level reported by the agent. Here, I analyze incentive compatible contract menus, that is, the testee finds it always optimal to truthfully report his type.

Proposition 1. *Any optimal incentive compatible contract menu must be one of the following test restrictions: mandatory training, unrestricted testing and attempt limit.*

While mandatory training and unrestricted testing are always incentive compatible contract menu, the attempt limit restrictions is not always incentive compatible. Its incentive compatibility relies on the expected relative return from one-period wait, $\frac{\Delta p_\theta}{2p_B}$. Difference $\Delta p_\theta = p_G - p_B$ measures how informative the test is or, in other terms, the ability of the test to distinguish between a strong and a weak candidate in terms of passing probabilities.

Proposition 2. *Attempt limit is an incentive compatible test restriction if and only if $\delta_A \geq \frac{2p_B}{p_G + p_B}$. Incentive compatibility threshold $\delta_{IC} = \frac{2p_B}{p_G + p_B}$ is strictly decreasing in the expected relative return $\frac{\Delta p_\theta}{2p_B}$.*

Proof. Attempt limit $\xi = (\xi_B^1, \xi_B^2, \xi_G^1, \xi_G^2) = (0, 1, 1, 0)$ recommends the strong candidate to the test in the first period, whereas the weak candidate to wait one session before taking it. If the agent is competent $\theta = G$, he always reports his true competence level. By assumption, he remains competent in period $t = 2$ and his chances to pass the test are unchanged in next session. Then he prefers to face the same lottery in period $t = 1$ rather than in $t = 2$ and he reveals his type by taking the test in period $t = 1$. If the agent is incompetent $\theta = B$, he's more likely to pass the test in the next period as he becomes competent with probability $\frac{1}{2}$. Thus he truthfully reports his type if and only if he's willing to wait one period before taking the test:

$$\delta_A \left(\frac{1}{2} \frac{1}{1 - \delta_A} p_G + \frac{1}{2} \frac{1}{1 - \delta_A} p_B \right) \geq \frac{1}{1 - \delta_A} p_B, \quad (2)$$

which is equivalent to $\delta_A \geq \delta_{IC} = \frac{2p_B}{p_G + p_B}$. The incentive compatibility threshold can be rearranged as a decreasing function of $\frac{\Delta p_\theta}{2p_B}$, $\delta_{IC} = \frac{2p_B}{p_G + p_B} = \frac{2p_B}{p_G - p_B + 2p_B} = \left(1 + \frac{\Delta p_\theta}{2p_B}\right)^{-1}$. \square

When the attempt limit is imposed, the weak agent chooses between two actions. He could take the test anyway and he would pass with probability p_B . Otherwise he could wait: with probability $\frac{1}{2}$, he remains incompetent and he gets zero return from the one-period wait; with probability $\frac{1}{2}$, he becomes competent and he increase the probability to pass the test by Δp_θ . So the expected absolute return is $\frac{\Delta p_\theta}{2}$, which is compared with the probability to pass by taking the test in the first period, p_B . The greater is the relative return, the lower is the incentive compatibility threshold δ_{IC} . The incentive compatibility condition is only a necessary condition for the attempt limit to be the optimal policy, but it does not guarantee its optimality. So even if the test is informative enough, the principal might prefer letting candidates take the test anytime to any other restriction.

Proposition 3. *There are $\bar{\delta}_A(p_G, \delta_P)$, $p_{AL}(p_G, \delta_P)$, $p_{MT}(p_G, \delta_P)$ and $p_{IC}(p_G)$ such that:*

- *Attempt limit is the optimal IC policy if and only if $\delta_A \geq \bar{\delta}_A(p_G, \delta_P)$ and $p_B \in [p_{AL}, p_{IC}]$;*
- *Mandatory training is the optimal IC policy if and only if $p_B > \max\{p_{MT}, p_{IC}\}$;*
- *Unrestricted testing is the optimal IC policy in all the other cases.*

For the attempt limit to be the optimal policy, the probability of a weak candidate passing the test must take intermediate values $p_B \in [p_{AL}, p_{IC}]$: on the one hand, it cannot be too high because it would violate the incentive compatibility constraint given a certain δ_A ; on the other hand it cannot be too small, because the benefit of avoiding to hire bad candidates would be too small and would not justify the high cost of no retakes for good agents failing the test. So one important result of Proposition 3 is that the attempt limit is optimal only if the test is not difficult. When weak candidate's probability to pass the test is too high, then it would be convenient for the principal to wait one period before administering the test and allow for the growth of human capital. This way, she reduces the probability to have a weak candidate taking the test. It's worth noticing that threshold $\bar{\delta}_A$ does not coincide with δ_{IC} : indeed $\delta_A \geq \delta_{IC}$ guarantees that the attempt limit is incentive compatible, but it does not imply its optimality. Condition $\delta_A \geq \bar{\delta}_A$ guarantees the existence of some p_B making the attempt limit the best incentive compatible restriction. Notice that, even if incentive compatibility is not sufficient to guarantee the optimality of the attempt limit restriction, it suffices to rule out mandatory training as the optimal policy: when the attempt limit is incentive compatible, the principal can successfully postpone testing the bad agent as much as the mandatory training does and, at the same time, she can start testing strong candidates in the first period. Given that the incentive compatibility of the attempt limit rules out the optimality of the mandatory training, the exercise of comparative statics helps understand better the trade-off between unrestricted testing and attempt limit - if implementable - and the trade-off between unrestricted testing and mandatory training.

Proposition 4. *Lower bound p_{AL} is increasing in δ_P for any value of δ_P and it decreases in p_G if and only if $p_G > \bar{p}_G(\delta_P)$. Threshold p_{MT} is decreasing in δ_P and increases in p_G if and only if $p_G \delta_P < \frac{4}{9}$. Threshold $\bar{\delta}_A(p_G, \delta_P)$ is increasing in δ_P and is decreasing in p_G .*

From Figure 10 it's possible to notice the way the discount rate of the principal δ_P affects the optimal test restriction choice. On the one hand, lower bound p_{AL} is increasing in δ_P , meaning that higher δ_P requires higher p_B . This is because the principal gets less focused on selecting the strong candidates for the present solely and he cares more about the future. As a result, he is more averse to the loss of human capital caused by no-retakes for strong candidates failing the test. On the other hand, threshold p_{MT} is increasing in δ_P because, for any p_B , the principal is more willing to wait one period before hiring

through the test.

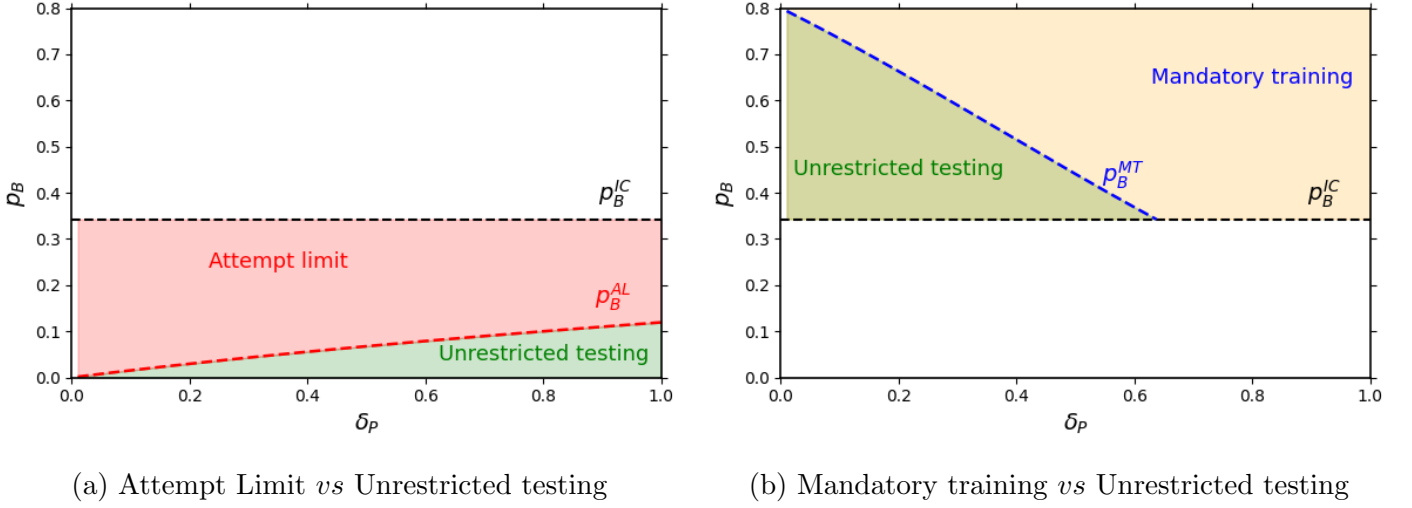


Figure 10: Optimal test restrictions

If the effect of the principal's discount rate δ_P is unambiguous on the optimal test restriction choice, this is not true for p_G . As far as the lower bound of the attempt limit p_{AL} is concerned, when p_G is sufficiently high, the test is good at passing good candidates and the principal is not concerned about the loss of human capital brought about by no-retakes; therefore an increase in p_G motivates the principal to adopt the attempt limit for lower values of p_B . By contrast, a small value of p_G implies two test features: (i) the test is good at failing weak candidates, which means that the principal is not concerned about letting bad agents take the test multiple times - a small value of p_G implies a small value of p_B by assumption; (ii) the test is not good at passing good candidates, who will fail the test regardless of the number of attempts a candidate has. An increase in p_G means that good candidates have more chances per test session to get passed, while weak candidates are still very likely to fail. This motivates the principal to give an additional try to the agents by choosing unconstrained testing over attempt limit. Likewise, the effect of p_G on p_{MT} is ambiguous. Threshold p_{MT} represents the value of p_B for which the principal is indifferent between unrestricted testing and mandatory training. Mandatory training allows for the growth of human capital but the associated cost is twofold: principal gives up period 1 expected utility as well as reducing the number of attempts for strong candidates to pass the test. A small value of $\delta_P p_G$ implies that either p_G or δ_P is too small. In the former case, a small value of p_G implies a small value of p_B meaning that the test is good at failing weak candidates and the principal does not really

need to postpone the test by one period to prevent weak candidates from taking it. They will fail the test anyway. Then a surge in p_G increases the opportunity cost of postponing the examination. If δ_P is too small, then the principal is eager to hire as soon as possible and a rise in p_G makes him more willing to start administering the test in period 1. When $\delta_P p_G$ is above $\frac{4}{9}$, that means that both δ_P and p_G are high. Thus, the test is very good at passing strong candidates and the principal is less concerned about leaving them with just one attempt and more willing to allow for the growth of human capital. Therefore, an increase in p_G makes her prefer mandatory training even for lower values of p_B .

The comparative statics results for $\bar{\delta}_A$ are consistent with the above findings for p_{AL} and p_{MT} . When p_G increases, threshold $\bar{\delta}_A$ decreases because: on the one hand, the attempt limit becomes incentive compatible even for less patient candidates; on the other hand, the principal herself becomes less concerned by the loss of human capital because the loss itself gets less likely. Finally, threshold $\bar{\delta}_A$ is increasing in δ_P : since the principal becomes more averse to the loss of human capital, he's less willing to implement the attempt limit unless p_B is very high; however, a high p_B requires a considerably patient agent so that the attempt limit remains incentive compatible.

5 Optimal passing threshold

In the previous section I analyzed the optimal test restriction choice while keeping the other test settings fixed, which resulted in considering p_B and p_G exogenous. In this section I consider the joint design of the test restriction and the minimum passing threshold, endogenizing the probability of passing the test for each type of agent.

Before delving into depth with the analysis of the optimal passing threshold, I need to introduce the test as a signal. Indeed any test can be thought of as a signal consisting of a realization space $S = [0, 1]$ and conditional distributions $\{\pi(s|\theta)\}_{\theta \in \{B, G\}}$. That means that a test score s is between 0 and 1 and the probability distribution of the test score is type dependent. When agent's type is B , he can get a score that is uniformly distributed between 0 and β , that is, $s_B \sim u([0, \beta])$. When agent's type is G , he can get a score that is uniformly distributed between γ and 1, that is, $s_G \sim u([\gamma, 1])$. Assume that $0 \leq \gamma \leq \frac{1}{2} \leq \beta \leq 1$, so that the two conditional distributions overlap. Figure 11 provides an example of two conditional test score distributions. The left panel shows the score distribution of

the bad candidate, who cannot get a score beyond 0.7: the probability density function is $\pi_B(s) = \frac{10}{7}$ for every $s \in [0, 0.7]$ and 0 elsewhere. The right panel displays the test score distribution of the good agent, who gets at least 0.3; thus the probability density function is $\pi_G(s) = \frac{10}{7}$ for every $s \in [0.3, 1]$ and 0 elsewhere. Given test conditional distributions, setting passing threshold $\sigma \in [0, 1]$ determines

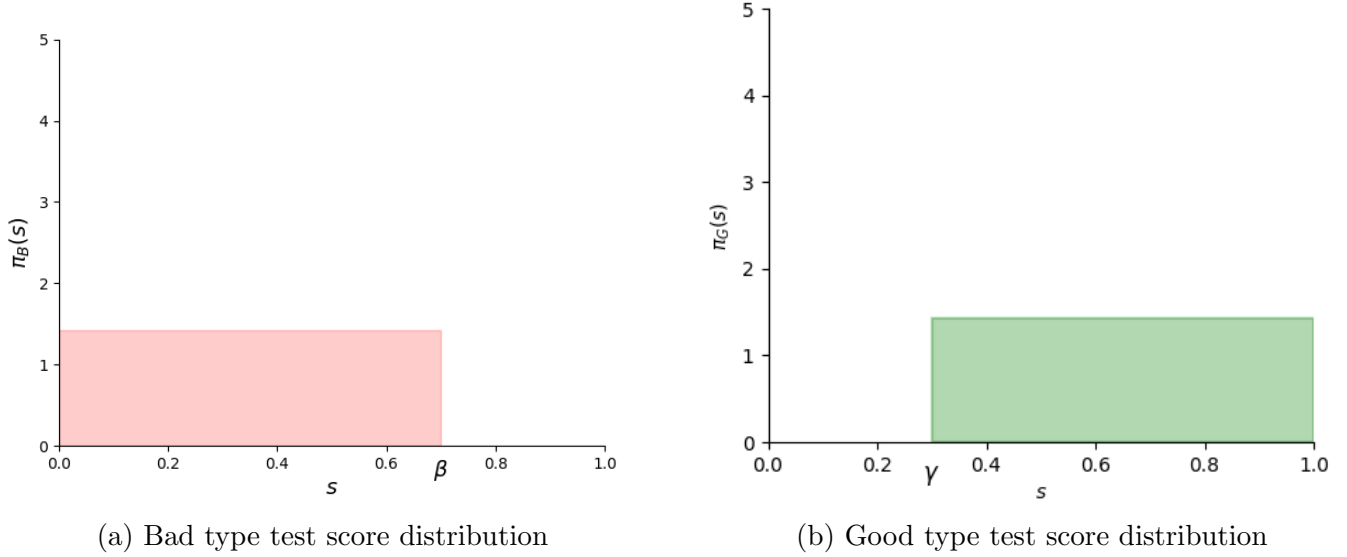


Figure 11: Conditional test score distributions

good and bad type probabilities to pass the test. The minimum passing threshold σ is the least score required to pass the test: everyone with a score above it passes the test. It's decided by the principal at the beginning of the game and it's applied to every test taker, independently of the type and the test session. Given that $\gamma \leq \beta$, the conditional distribution overlap and share a region of their score range. For every $\sigma \in [\gamma, \beta]$, the probabilities to pass the test for the bad and the good agents are respectively:

$$p_B(\sigma) = 1 - \frac{\sigma}{\beta},$$

$$p_G(\sigma) = \frac{1 - \sigma}{1 - \gamma}.$$

Lemma 1. *The optimal passing threshold $\hat{\sigma}$ must live within the overlapping interval $[\gamma, \beta]$. For every $\sigma \in [\gamma, \beta]$, $\Delta p_\theta(\sigma) \geq 0$.*

Proof. I start proving that $\hat{\sigma} \in [\gamma, \beta]$. Without loss of generality, assume that $0 < \gamma \leq \beta < 1$. Suppose

that $\gamma \leq \beta < \sigma$: then $p_B(\sigma) = 0 < p_G(\sigma) < 1$. By reducing σ to β , the principal would increase $p_G(\sigma)$ without affecting $p_B(\sigma)$. Then $\sigma > \beta$ cannot be optimal. Likewise, suppose $\sigma < \gamma \leq \beta$: then $0 < p_B(\sigma) < 1 = p_G(\sigma)$. By increasing σ to γ , the principal would reduce $p_B(\sigma)$ without affecting $p_G(\sigma)$. Thus, $\sigma < \gamma$ is not optimal. Therefore $\hat{\sigma} \in [\gamma, \beta]$. Now I prove that $\sigma \in [\gamma, \beta]$ implies $\Delta p_\theta(\sigma) \geq 0$. I need to show that for every $\sigma \in [\gamma, \beta]$

$$p_G(\sigma) = \frac{1 - \sigma}{1 - \gamma} \geq 1 - \frac{\sigma}{\beta} = p_B(\sigma)$$

By rearranging terms, the above inequality is equivalent to $\gamma\beta + (1 - \gamma)\sigma \geq \sigma\beta$. Because $\sigma \in [\gamma, \beta]$ and $\gamma \in [0, 1]$, $\gamma\beta + (1 - \gamma)\sigma \geq \sigma \geq \sigma\beta$. Therefore $\Delta p_\theta(\sigma) \geq 0$ for every $\sigma \in [\gamma, \beta]$. □

The null hypothesis of the test is that the test-taker is a weak candidate. By scoring beyond the passing threshold, he provides enough evidence to reject the null hypothesis. Because of the overlapping of the two conditional distributions, there must be at least one type of error occurring for any passing threshold set by the principal. False positive error happens when the bad agent obtains a score above the threshold, $s_B \geq \sigma$, while the false negative error occurs when the good type agent's score is below the minimum passing score, $s_G \leq \sigma$. Figure 12 shows two conditional distributions sharing the brown area between γ and β : the dark brown shaded area to the right of the passing threshold σ represents the false positive error, while the dark brown shaded area to the left of σ displays the false negative error.

Before analyzing the joint design of the test restriction and the passing threshold, I focus on the problem of the threshold setting for each test restriction. In doing so, I mainly focus on three alternative types of test signals and I define them by using the likelihood ratio $\mathcal{L}(s) = \frac{\pi_G(s)}{\pi_B(s)}$.

Definition 2. Consider a test signal such that $\beta - \gamma < 1$.

- A *Bad-Certifying-Test (BCT)* is a signal in which there is a score \underline{s} such that $\mathcal{L}(s) = 0$ for $s < \underline{s}$ and $0 < \mathcal{L}(s) < \infty$ for $s \geq \underline{s}$.
- A *Good-Certifying-Test (GCT)* is a signal in which there is a score \bar{s} such that $0 < \mathcal{L}(s) < \infty$ for $s \leq \bar{s}$ and $\mathcal{L}(s) = \infty$ for $s > \bar{s}$;

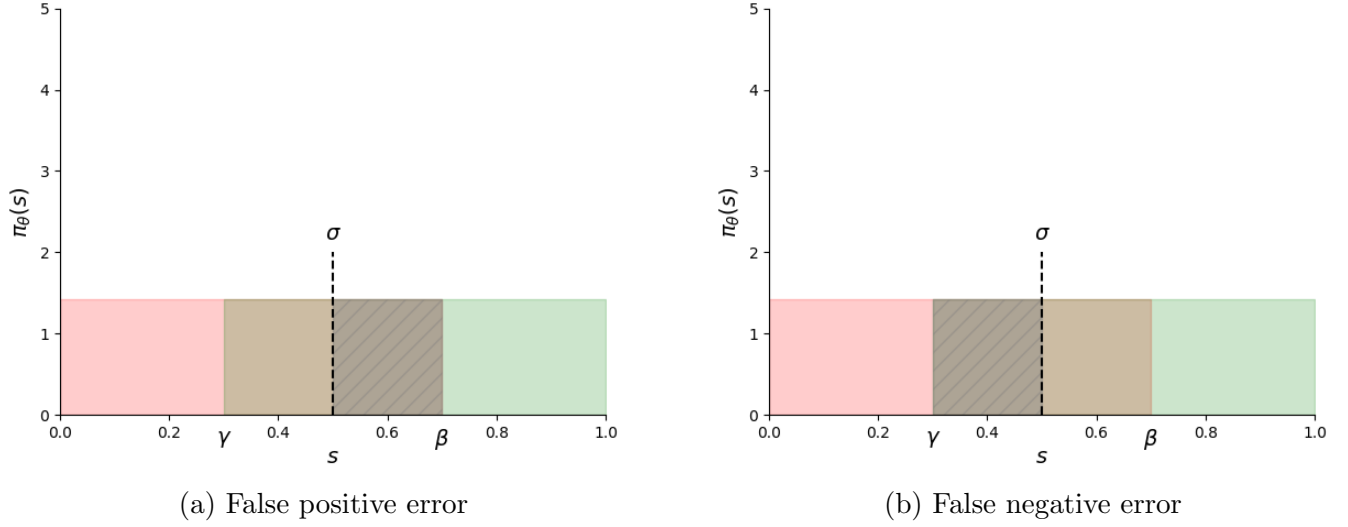


Figure 12: False positive *vs* False negative

► A *Two-Sided-Certifying-Test (TSCT)* is a signal in which there are \underline{s} and \bar{s} such that $\mathcal{L}(s) = 0$ for $s < \underline{s}$ and $\mathcal{L}(s) = \infty$ for $s > \bar{s}$.

Figure 13 gives an example of a bad-certifying test. The right panel of Figure 13 shows the likelihood ratio $\mathcal{L}(s)$ constant at 0 for $s < \gamma = \frac{1}{2}$ and $0 < \mathcal{L}(s) < \infty$ for $\gamma \leq s$. Even though the test is not able to certify if an agent is competent, a *BCT* signal provides conclusive evidence about the bad type. By observing a score $s < \gamma$, the principal is able to conclude that the agent is a weak candidate with probability 1. Likewise, a *GCT* signal provides conclusive evidence about the good type. Whoever gets

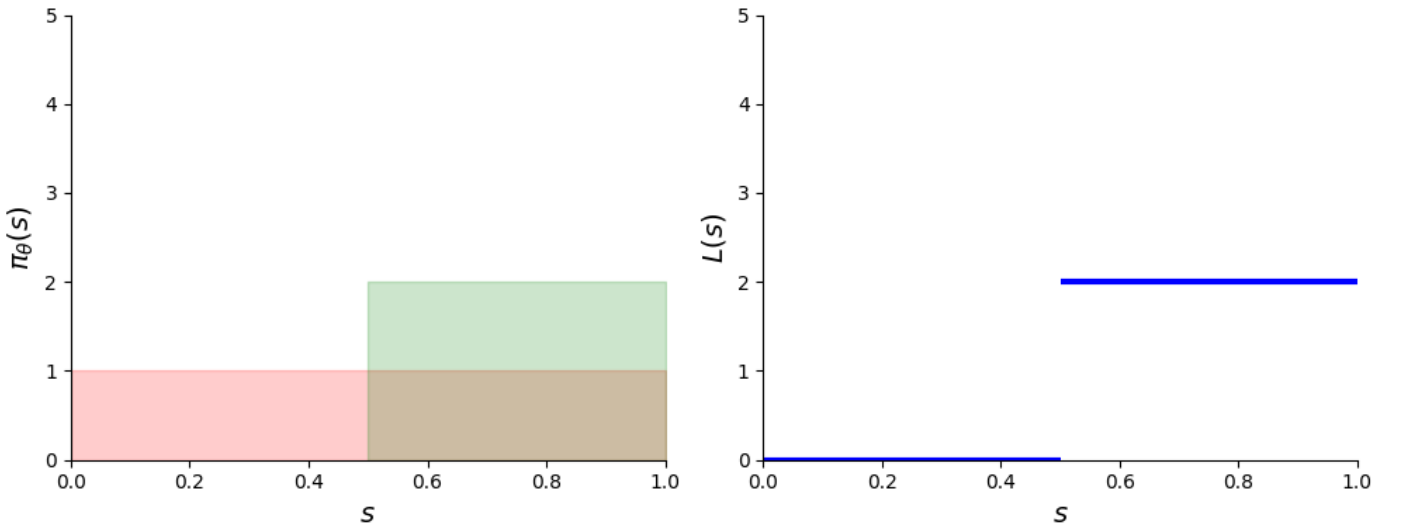


Figure 13: Bad-Certifying-Test

a score above γ can be certified as competent. Figure 14 shows an example of a test providing conclusive evidence about the good type, but not about the bad one. Indeed the likelihood ratio is $\mathcal{L}(s) = 2$ for $s \leq \beta = \frac{1}{2}$ and $\mathcal{L}(s) = \infty$ for $s > \beta$. Finally Figure 15 provides an example of a Two-Sided Certifying

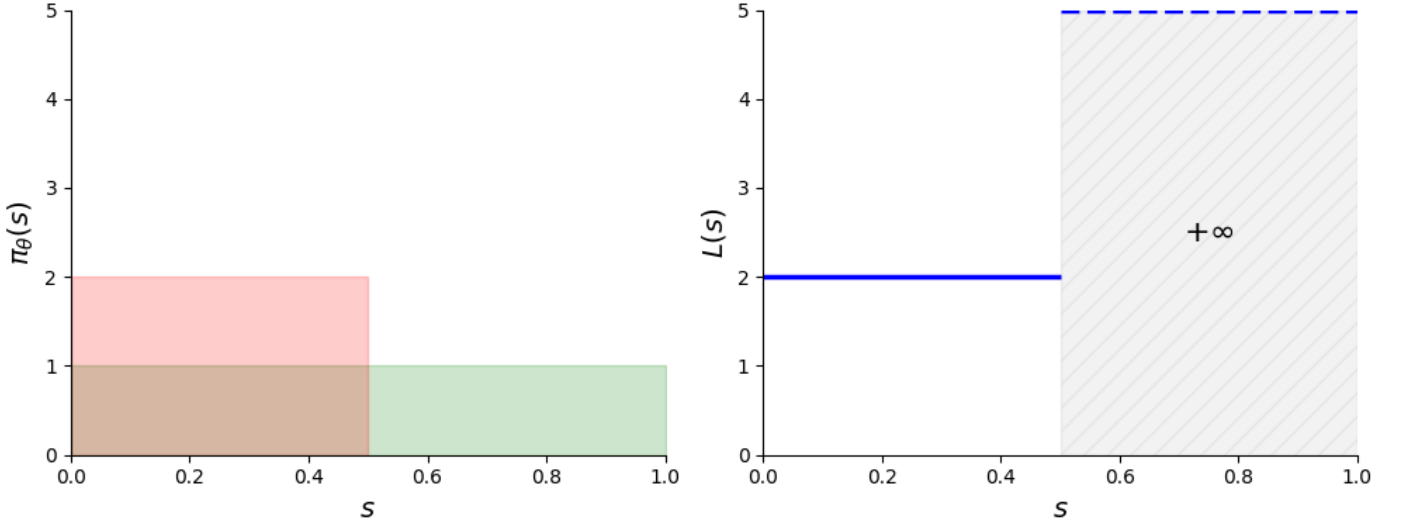


Figure 14: Good-Certifying-Test

Test. Here the test is able to give conclusive evidence about both types of candidates. Indeed the likelihood ratio is $\mathcal{L}(s) = 0$ for $s \leq \gamma = \frac{1}{4}$ and $\mathcal{L}(s) = \infty$ for $s \geq \beta = \frac{3}{4}$, while $\mathcal{L}(s)$ is 1 for $s \in [\gamma, \beta]$, as both types are equally likely within that interval.

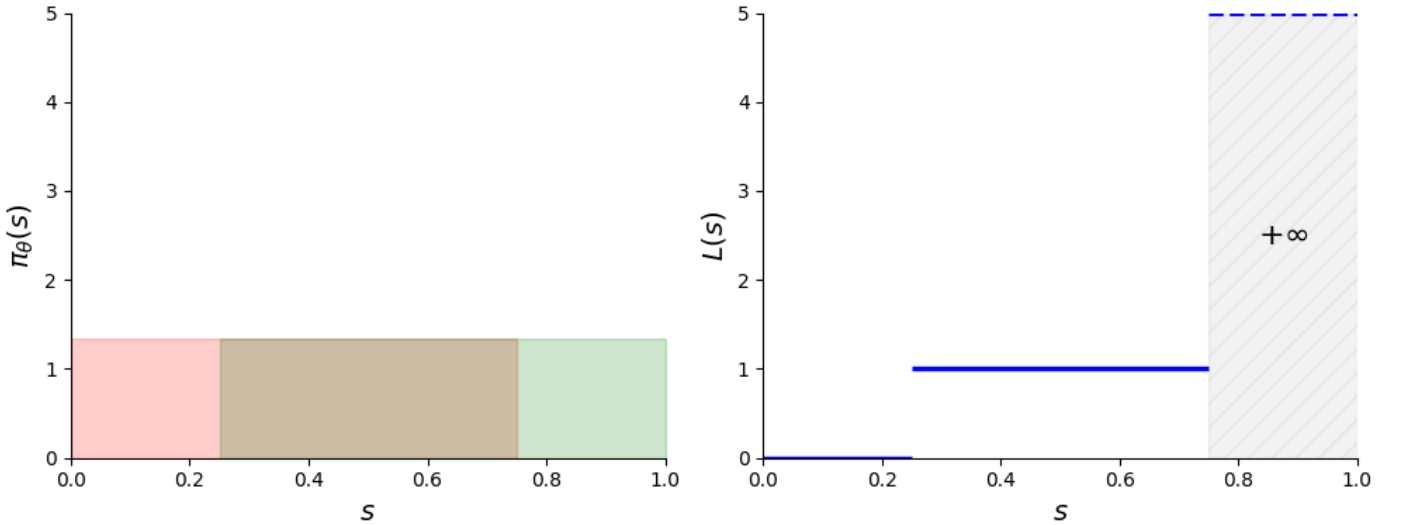


Figure 15: Two-Sided-Certifying-Test

It's easy to notice from the graphs here represented that, consistent with $\Delta p_\theta(\sigma) \geq 0$ for every

$\sigma \in [\gamma, \beta]$, any test here considered respects the *Monotone Likelihood Ratio Property (MLRP)*: the higher is the score, the more likely is that the agent is competent. For the analysis that follows I assume three specific test signals, one for each type of tests defined above:

1. *Bad-certifying test signal:*

$$\pi_{BCT}(s|\theta) = \begin{cases} u([0, 1]), & \text{if } \theta = B, \\ u([\frac{1}{2}, 1]), & \text{if } \theta = G; \end{cases}$$

2. *Good-certifying test signal:*

$$\pi_{GCT}(s|\theta) = \begin{cases} u([\frac{1}{2}, 1]), & \text{if } \theta = B, \\ u([0, 1]), & \text{if } \theta = G; \end{cases}$$

3. *Two-sided-certifying test signal:*

$$\pi_{TSCT}(s|\theta) = \begin{cases} u([0, \frac{3}{4}]), & \text{if } \theta = B, \\ u([\frac{1}{4}, 1]), & \text{if } \theta = G; \end{cases}$$

All the three test signals have the same overlapping area size $\beta - \gamma = \frac{1}{2}$, even if its location within the score range changes.

5.1 Exogenous restriction

In this part of the section I assume that the test restriction is exogenous and I discuss what is the optimal threshold for each type of contract menu $\xi \in \{AL, UT, MT\}$ and test signal. Thus, the timing of the game is:

1. Principal establishes threshold $\hat{\sigma} \in [\gamma, \beta]$;
2. Competence level $\theta \in \{B, G\}$ is privately observed by the agent;
3. Agent chooses whether to take the test;

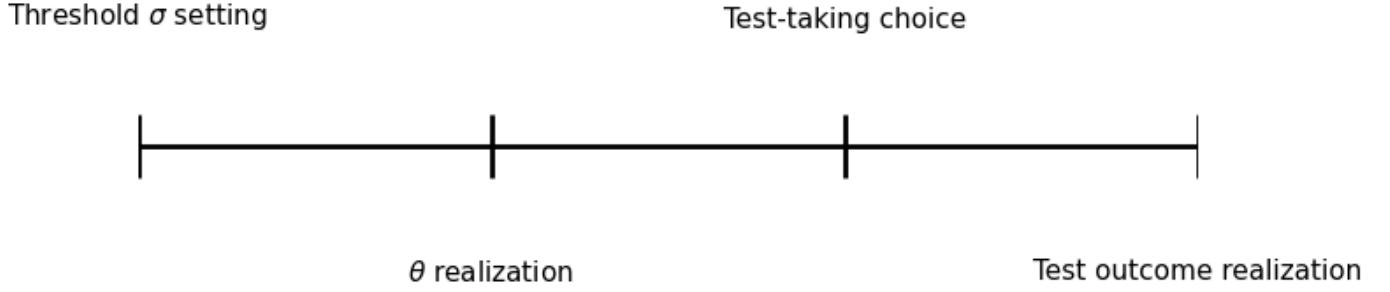


Figure 16: Timing of the events

4. Test outcome is realized, $\omega \in \{P, F\}$. If $\omega = P$, the game ends; if $\omega = F$, a new test session might take place, depending on test settings.

5.1.1 Attempt Limit

I start illustrating a general result, which allows me to briefly discuss two extreme cases and then I discuss in detail the implications for each of the three types of test signals.

Proposition 5. *Assume that $\xi = AL$. The optimal incentive compatible passing threshold $\hat{\sigma}_{AL}$ is*

$$\hat{\sigma}_{AL} = \begin{cases} \gamma, & \text{if } \frac{\gamma}{\beta} \geq 2^{\frac{1-\delta_A}{2-\delta_A}} \\ \beta \frac{(1-\gamma)(2-\delta_A)-\delta_A}{(1-\gamma)(2-\delta_A)-\beta\delta_A}, & \text{if } \frac{\gamma}{\beta} < 2^{\frac{1-\delta_A}{2-\delta_A}} \end{cases} \quad (3)$$

Here there are two issues to discuss separately: on the one hand I need to find which passing thresholds are incentive compatible in a test with attempt limit and, on the other hand, I need to look for the utility maximizing incentive compatible passing threshold. As far as the first one is concerned, I have to prove that the set of incentive compatible passing thresholds is a compact space. In particular, if $\frac{\gamma}{\beta} \geq 2^{\frac{1-\delta_A}{2-\delta_A}}$, this space is $[\gamma, \beta]$ itself. Otherwise, there must be a threshold $\bar{\sigma} \in (\gamma, \beta]$ such that the compact space is given by $[\bar{\sigma}, \beta]$ and this value is exactly $\bar{\sigma} = \beta \frac{(1-\gamma)(2-\delta_A)-\delta_A}{(1-\gamma)(2-\delta_A)-\beta\delta_A}$.

After finding the set of incentive compatible thresholds, I need to select the optimal one from this set. To do that, it suffices to show that principal's utility is decreasing over the space of incentive compatible passing thresholds and she ends up choosing the lower bound of such a space. Intuitively, the principal

knows that, by implementing the attempt limit, only the good agents show up in the first period and in the second period competence is equally likely. Thus, she prefers to minimize the probability of a false negative by selecting the lowest incentive compatible passing threshold.

Notice that the this threshold is not-increasing in δ_A . As long as δ_A is too small, the passing threshold strictly decreases in δ_A : the smaller δ_A is, the higher the passing threshold must be to deter the bad agent from taking the test. Therefore the optimal passing threshold decreases in δ_A till hitting the lower bound of the good agent's test score distribution. From that point on, $\hat{\sigma}_{AL}$ is constant in δ_A . Intuitively, if the agent is patient enough, the principal can easily convince him to wait one period by imposing a low passing threshold; as the agent becomes more impatient, the principal needs to make it harder for the candidate to pass the test by increasing the passing score.

Another important result is that, when the ratio $\frac{\gamma}{\beta}$ is high enough, the attempt limit can be easily implemented by setting $\hat{\sigma}_{AL} = \gamma$ and the probability of a false negative error is zero. This is because γ and β are so close to each other that the probability of the bad type to pass the test is very small for any $\sigma \geq \gamma$. Therefore, it's not convenient for him to take the test in period 1. In other words, the test is highly informative despite the overlapping between the two conditional distributions making the no-retakes threat compelling. On the other hand, when the ratio $\frac{\gamma}{\beta}$ is small, the optimal passing threshold gets closer to β : $\gamma < \hat{\sigma}_{AL} \leq \beta$. This allows me to briefly discuss two extreme cases: one is the full overlapping case whereby $\beta - \gamma = 1$, the other is the full information case in which $\beta - \gamma = 0$. In the former case, the test is not able to distinguish the good from the bad agent at all. Thus, the only incentive compatible passing threshold is $\sigma = 1$, that is, failing everyone. In fact, there is no incentive compatible passing threshold. In the latter case, the test is fully informative: the two conditional distributions share a zero measure overlapping interval and the optimal incentive compatible passing threshold is $\hat{\sigma} = \gamma = \beta$. This way the test is able to pass every competent agent and fail every weak candidate.

There are intermediate cases in which $0 < \beta - \gamma < 1$ and, depending on the location of the overlapping interval, the optimal passing threshold could change. With regard to this aspect, consider the three alternative test signals: π_{BCT} , π_{GCT} and π_{TSCCT} . All of them have the same overlapping area size, $\beta - \gamma = \frac{1}{2}$, but the score interval in which the two conditional distribution overlap differs by the type of

test.

Corollary 1. *Assume $\xi = AL$. The optimal policy is test-signal dependent:*

- *If $\pi = \pi_{BCT}$, the optimal IC passing threshold is $\hat{\sigma}_{AL} = \gamma$ if $\delta_A \geq \frac{2}{3}$ and $\hat{\sigma}_{AL} = \beta$ otherwise;*
- *If $\pi = \pi_{GCT}$, the optimal IC passing threshold is $\hat{\sigma}_{AL} = 2\beta \frac{1-\delta_A}{2-(1+\beta)\delta_A}$ for every $\delta_A \in (0, 1)$;*
- *If $\pi = \pi_{TSCT}$, the optimal IC passing threshold is $\hat{\sigma}_{AL} = \gamma$ if $\delta_A \geq \frac{4}{5}$ and $\hat{\sigma}_{AL} = \frac{2\beta-\delta_A(1+\beta)}{2(1-\delta_A)}$ otherwise.*

When the principal is running a bad-certifying test, the optimal incentive compatible passing threshold is a cut-off function of δ_A . If candidate's sensitivity to attempt limit is high enough, $\delta_A \geq \frac{2}{3}$, the optimal minimum passing score coincides with the lower bound of the good agent's test score distribution, which means that the principal is able to pass every competent agent while avoiding to select bad candidates in the first test period. If $\delta_A \leq \frac{2}{3}$, the optimal passing score is the upper bound of the bad agent's test score distribution and, in the context of the bad-certifying test, the principal is failing everyone. When the principal is running a good-certifying test, the optimal incentive compatible passing threshold is strictly between γ and β and it's a strictly decreasing function of δ_A . Therefore - unlike the bad-certifying test - both types of errors are tolerated over the two periods and the passing threshold is a continuous function of δ_A . The two-sided-certifying test combines the features of the other two: it's constant at γ for δ_A high enough and it's a strictly decreasing function of δ_A for $\delta_A < \frac{4}{5}$.

5.1.2 Unrestricted testing

I'll briefly discuss the two extreme cases the same way I did for the attempt limit restriction. For the full information case, the passing threshold is always $\hat{\sigma}_{UT} = \gamma = \beta = \frac{1}{2}$. Unlike the attempt limit restriction, the optimal passing threshold for the full overlapping case is $\hat{\sigma}_{UT} = \frac{1}{2}$. As far as the intermediate case is concerned - when $\beta - \gamma = \frac{1}{2}$ - the optimal passing threshold is still signal dependent.

Proposition 6. *Assume that $\xi = UT$. The optimal policy is test-signal dependent:*

- *If $\pi = \pi_{BCT}$, the optimal passing threshold is $\hat{\sigma}_{UT} = \gamma$ for $\delta_P \leq 1/2$ and $\hat{\sigma}_{UT} \in (\gamma, \beta)$ otherwise;*

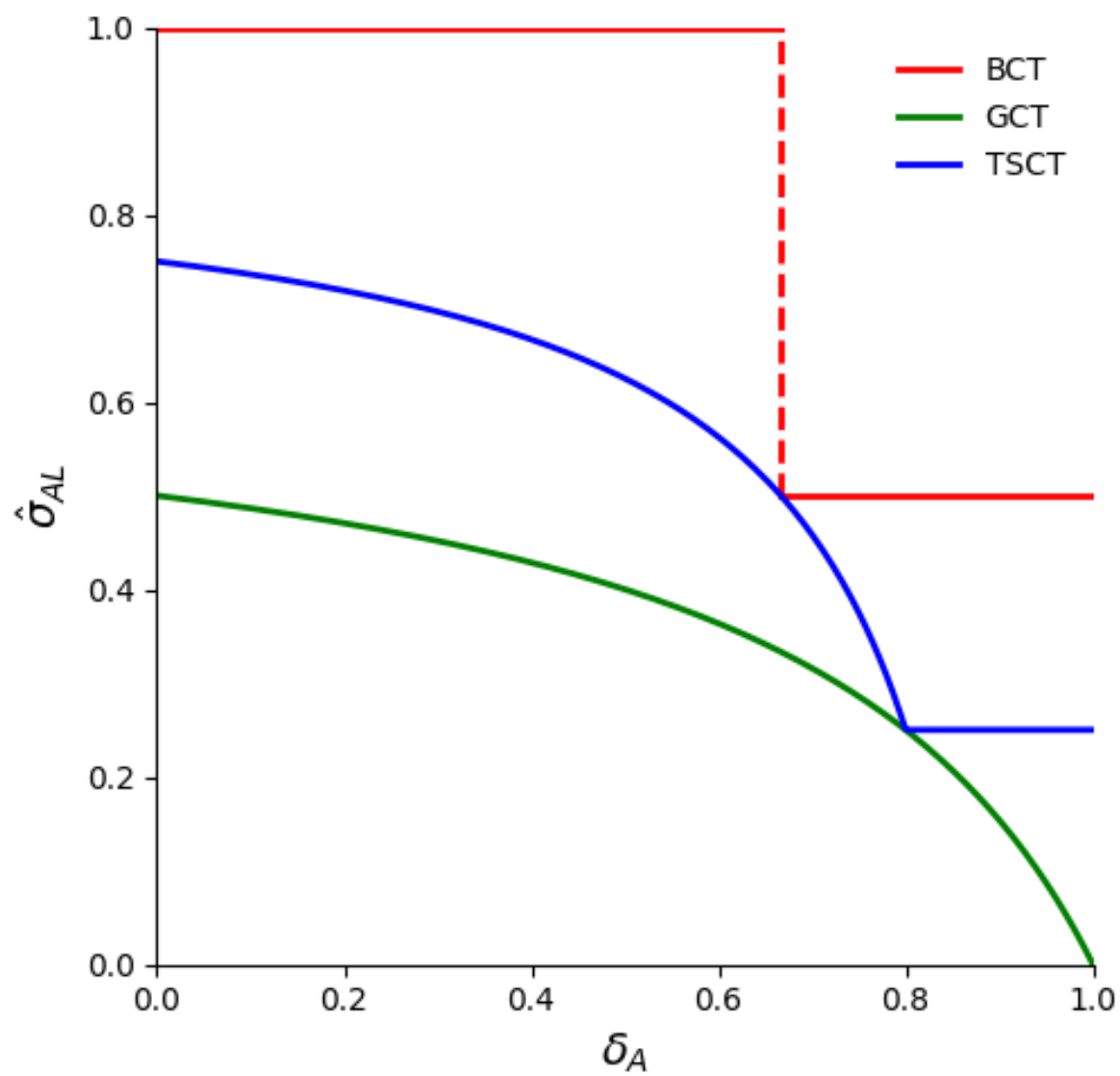


Figure 17: Attempt limit passing thresholds

- If $\pi = \pi_{GCT}$, the optimal passing threshold is $\hat{\sigma}_{UT} = \beta$;
- If $\pi = \pi_{T SCT}$, the optimal passing threshold is $\hat{\sigma}_{UT} \in (\gamma, \beta)$;

When the principal is running a good-certifying test, she does not tolerate a false positive error and she ends up setting a high passing threshold. By contrast, administering a bad-certifying test leads the principal to select $\hat{\sigma}_{UT} = \gamma$ unless δ_P is too high: in that case, the principal is willing to tolerate both types of errors because he prefers to allow for the growth of human capital by being more selective. Notice that, when $\delta < \frac{1}{2}$, the principal chooses the same threshold for both types of test, $\hat{\sigma}_{UT} = \frac{1}{2}$, even though the type of error tolerated is different. Consistent with the results for *BCT* and *GCT*, the two-sided certifying tests tolerates both the false positive and the negative errors by setting $\hat{\sigma} \in (\gamma, \beta)$.

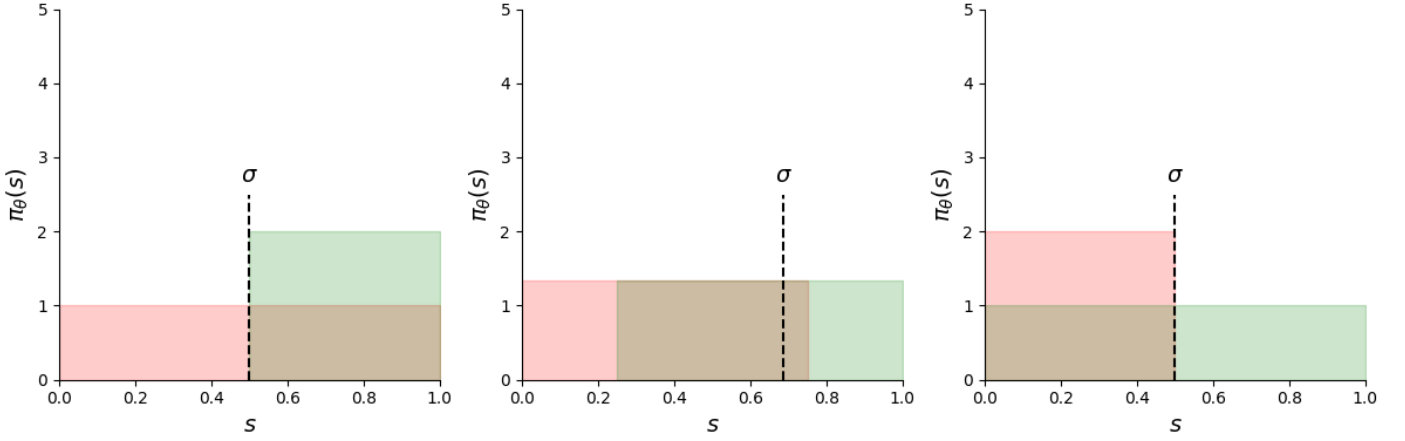


Figure 18: Unrestricted testing passing thresholds

5.1.3 Mandatory training

As far as the extreme cases are concerned, the full information environment has always the same passing threshold $\hat{\sigma}_{MT} = \gamma = \beta$. In the full overlapping case, the optimal threshold is $\hat{\sigma}_{MT} = 0$. This is because the test is completely unreliable as a selection device, but at the same time most candidates are competent because selection has been postponed allowing for the growth of human capital. Without any informative tool to select personnel, the principal is better off by admitting the candidate given that he's more likely to be competent. The same principle drives the tester through the selection of the optimal passing threshold for the intermediate cases in which $\beta - \gamma = \frac{1}{2}$.

Proposition 7. Assume that $\xi = MT$. The optimal passing threshold is $\hat{\sigma} = \gamma$.

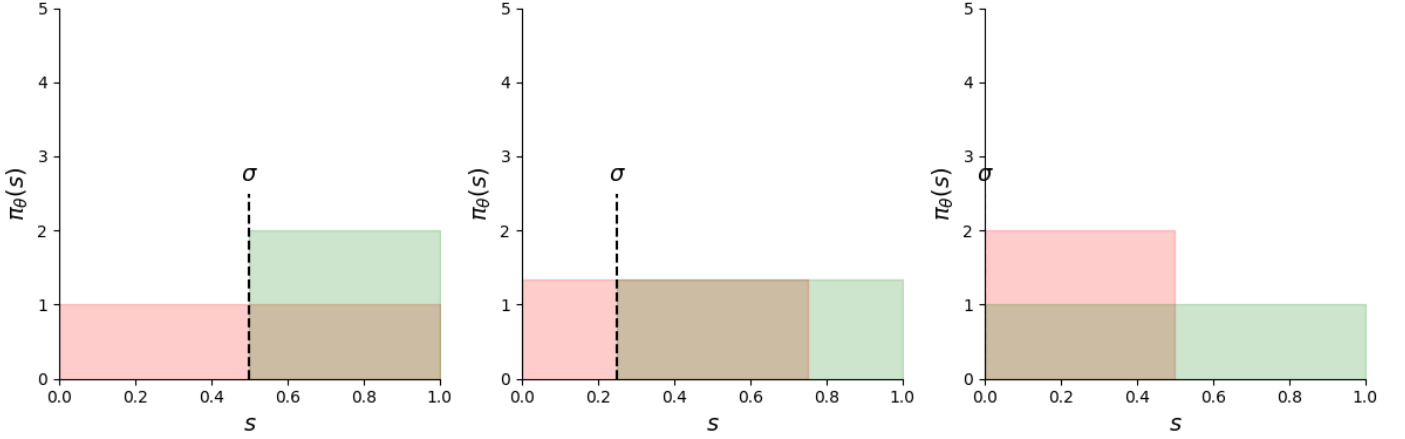


Figure 19: Mandatory training passing thresholds

5.2 Joint design of test restriction and passing threshold

In this section I analyze the combined choice of (ξ, σ) for every type of test signal. The timing of the game is

1. Principal sets test settings (ξ, σ) ;
2. Competence level $\theta \in \{B, G\}$ is privately observed by the agent;
3. Agent decides whether or not to take the test;
4. Test outcome is realized, $\omega \in \{P, F\}$. If $\omega = P$, the game ends; if $\omega = F$, a new test session might take place, depending on test settings.

Before discussing the results for the three types of test signals, I'm going to briefly illustrate the equilibrium policy for the two extreme cases so far considered. As far as the full information case is concerned, the principal is indifferent between implementing the attempt limit restriction and the unrestricted testing: the incompetent agent always fail the test, while the strong candidate always pass it. In the full overlapping scenario, the principal ends up choosing the mandatory training and letting everyone pass the test, that is, $(\xi, \hat{\sigma}_{MT}) = (MT, 0)$

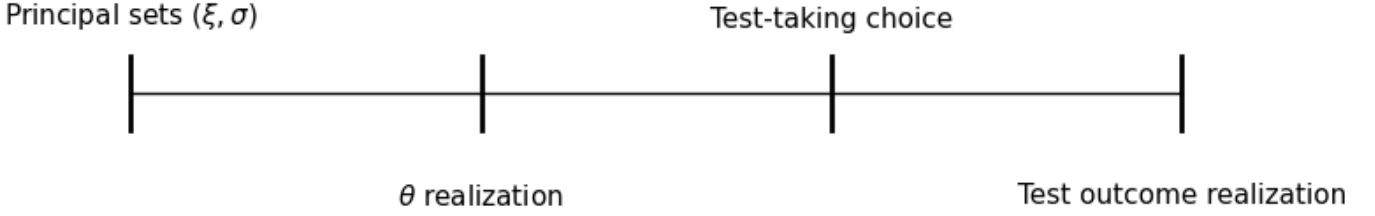


Figure 20: Timing of the events

5.2.1 Bad-certifying-test

I'm going to compare the results of the first part of this section to obtain the best combination (ξ, σ) for a bad-certifying test and I'll show that the best policy is a function of two parameters, δ_A and δ_P . In particular, the optimality of the attempt limit is a cut-off function of δ_A and the optimal passing threshold does not depend on δ_A at all. On the other hand, when the attempt limit is not optimal, the optimal restriction depends only on the value of δ_P .

In the first part of the section I've shown that the attempt limit threshold for a bad-certifying test is a cut-off function that changes value depending on the level of δ_A : if $\delta_A \geq \frac{2}{3}$, then $\hat{\sigma}_{AL} = \gamma$; otherwise, $\hat{\sigma}_{AL} = 1$. In fact, when candidate is sufficiently sensitive to the attempt limit, the tester is able to set a passing threshold that can pass every competent agent while deterring weak candidates from taking the test in the first period. If instead the weak candidate is not patient enough, the principal has to raise the passing threshold so high that everyone ends up failing the test. In contrast, the passing threshold does not need to be extremely high when the principal allows the agent to take the test at any time - $\hat{\sigma}_{UT} < \beta$ - and, if δ_P is sufficiently small, she prefers $\hat{\sigma}_{UT} = \gamma$. Unlike the other two restrictions, the optimal mandatory training passing threshold does not depend on either δ_A or δ_P and is constant at γ . Now that I've reviewed the optimal passing threshold for each test restriction, I can illustrate the optimal combination $(\xi^{BCT}, \sigma_{BCT})$.

Proposition 8. *Assume a bad-certifying test, $\pi = \pi_{BCT}$:*

1. *The optimal passing threshold is restriction independent, $\sigma_{BCT} = \gamma$;*
2. *The attempt limit is the optimal test restriction if and only if $\delta_A \geq \frac{2}{3}$;*

3. If $\delta_A < \frac{2}{3}$, the optimal bad-certifying test restriction is

$$\xi^{BCT} = \begin{cases} UT, & \text{if } \delta_P \leq \frac{4}{9} \\ MT, & \text{if } \delta_P > \frac{4}{9} \end{cases}$$

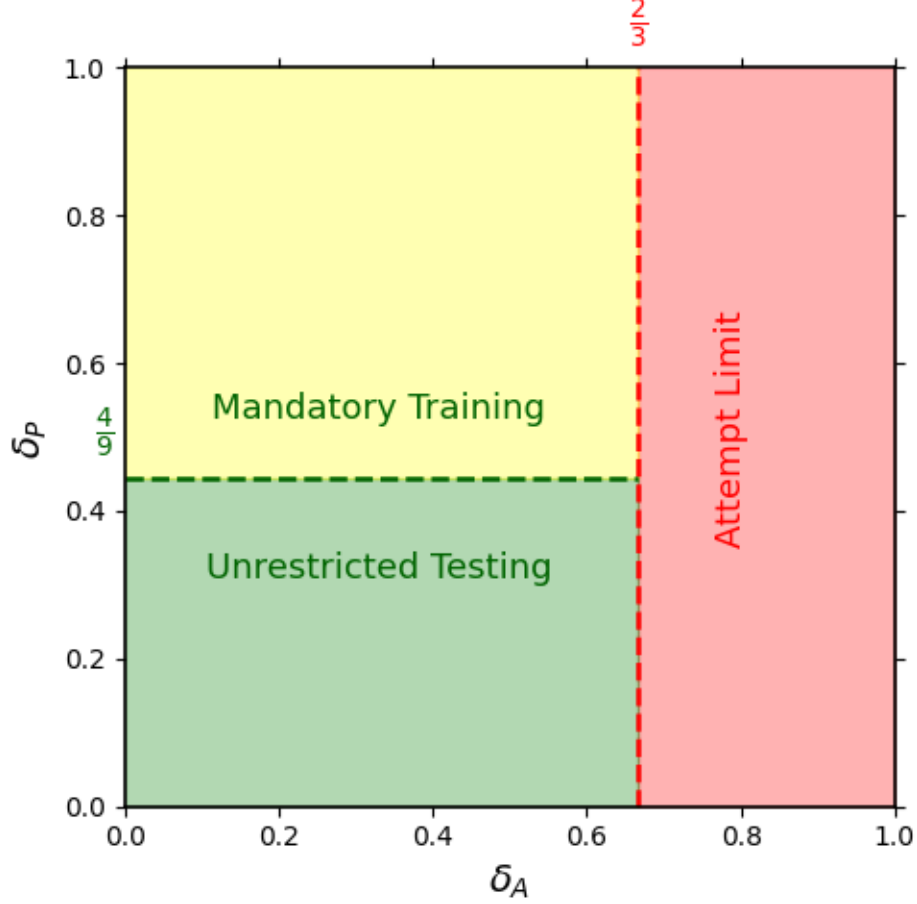


Figure 21: Optimal bad-certifying test setting

The first important conclusion regarding the best bad-certifying test setting is that the principal never tolerates a false negative error. I temporarily narrow the attention down to the choice between the mandatory training restriction and the unrestricted testing regime. In the first part of the section I've shown that the choice of the threshold depends on the size of δ_P : the larger δ_P is, the higher $\hat{\sigma}_{UT}$ is. Here I show that if the size of δ_P justifies $\hat{\sigma}_{UT} > \gamma$, then the same value of δ_P motivates the principal to choose *MT* over *UT* itself. Therefore the optimal passing threshold is always γ , regardless of the test restriction applied, meaning that the equilibrium passing probabilities for the bad and the good

candidates, $p_B(\sigma)$ and $p_G(\sigma)$, are always the same. Related to this conclusion, the other important result is that, if $\delta_A \geq \frac{2}{3}$, there is no other policy better than the attempt limit. The intuition builds exactly upon the fact that all the three restrictions show the same probabilities for the bad and the good candidates to pass the test, which leads the discussion back to the previous section - when I compared the three restrictions while keeping fixed the passing probabilities. Because there is no false negative error, the principal is not concerned by the loss of human capital caused by no-retakes policy. Therefore she always prefers the attempt limit. If instead $\delta_A < \frac{2}{3}$, the attempt limit is in fact not implementable because its incentive compatible threshold would be $\sigma = 1$ - failing everyone - and the optimal restriction choice boils down to selecting either UT or MT , which depends on δ_P .

5.2.2 Good-certifying test

Now I analyze what is the optimal setting combination when the principal is running a good-certifying test. If the attempt limit gets implemented, the principal needs to set a threshold that is strictly between γ and β , which means allowing for both the false negative and the false positive. If the principal chooses to let the candidate free to take the test anytime, then she would set a high passing standard, $\hat{\sigma} = \beta_P$, implying no tolerance for a false positive error. In contrast, the mandatory training would imply no false negative error, given that $\hat{\sigma}_{MT} = \gamma$. If the equilibrium policy in the bad-certifying test implied only one type of error, here it's guaranteed that every equilibrium test restriction determines a different degree of tolerance towards the false negative and the false positive errors.

Proposition 9. *Assume a good-certifying test, $\pi = \pi_{GCT}$. The optimal setting is either $(AL, \beta \frac{2(1-\delta_A)}{2-\delta_A(1+\beta)})$ or (UT, β) . Attempt limit setting is the optimal policy if and only if $\delta_P \leq \frac{\delta_A}{2-\delta_A}$.*

It's immediate to notice that mandatory training is never an equilibrium test restriction. The reason is the optimal passing threshold of the unrestricted testing regime with a high passing threshold working as a double filter: first, it allows for the growth of human capital because bad candidates are prevented from passing the test in the first period; second, weak candidates remaining incompetent in the second period fail the test again - unlike the mandatory training test passing everyone in the second period. Finally, compared to mandatory training, unrestricted testing has the advantage of giving competent people one more attempt at the outset of the game. So it can be shown that unrestricted testing

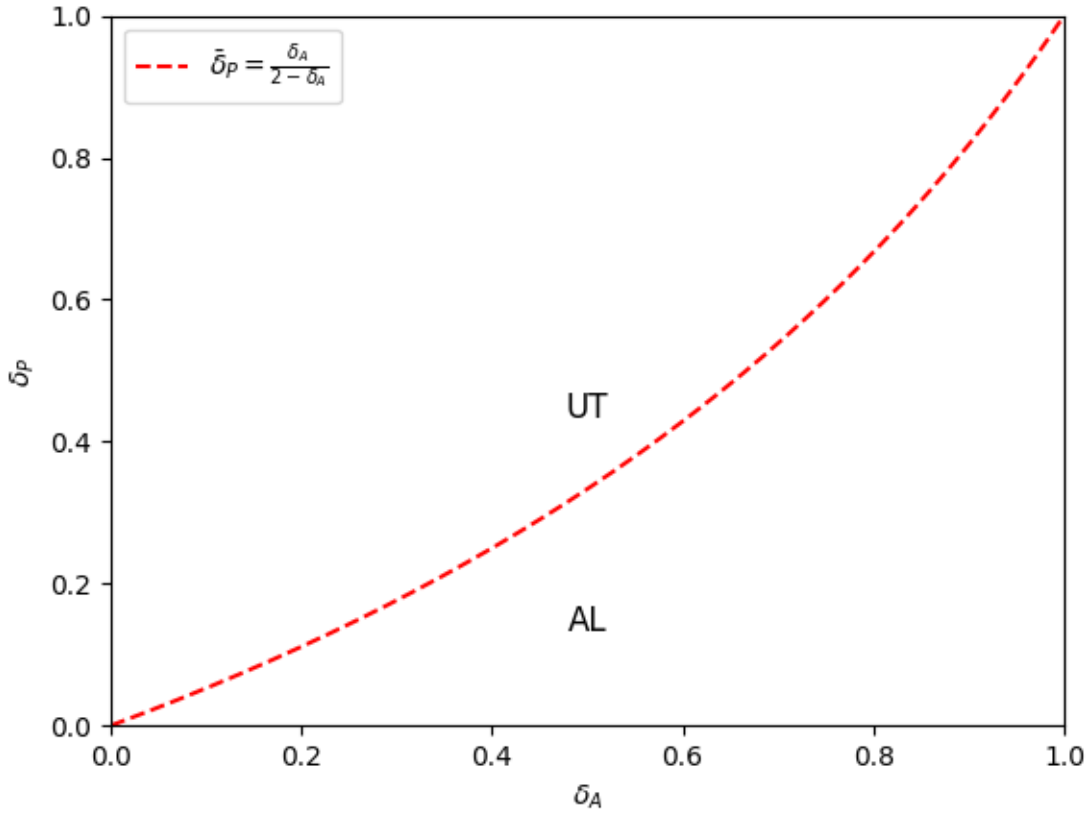


Figure 22: Optimal good-certifying test setting

dominates mandatory training and the principal is left to compare attempt limit with unrestricted testing. Both are able to avoid hiring bad candidates in the first period: on the one hand, attempt limit has a lower passing threshold; on the other hand, it leaves period 1 good candidates with just one attempt. Principal's aversion to the loss of human capital combined with candidate's sensitivity to attempt limit makes the difference. For a sufficiently high aversion to human capital loss, the principal ends up choosing unrestricted testing. The threshold value that makes her prefer unrestricted testing over attempt limit depends on candidate's sensitivity to attempt limit: if the agent is extremely impatient, the principal needs to raise the passing threshold to deter the agent from taking the test, increasing the probability of a false negative. So the lower δ_A is, the lower is threshold.

5.2.3 Two-sided-certifying test

The optimal two-sided-certifying test settings combines the features of the other two types of tests. Mandatory training imposes $\hat{\sigma}_{MT} = \gamma$, while the optimal unrestricted testing threshold is interior, $\hat{\sigma}_{UT} = \frac{11}{36} \in (\gamma, \beta)$. As far as the attempt limit is concerned, its optimal passing threshold is γ for $\delta_A \geq \frac{4}{5}$ and, otherwise, it's equal to $\frac{2\beta - \delta_A(1+\beta)}{2(1-\delta_A)}$.

Proposition 10. *Assume a two-sided-certifying test, $\pi = \pi_{TSC}$. There exist $\bar{\delta}_A$ and $\underline{\delta}_A$, $\underline{\delta}_A < \bar{\delta}_A$, such that test setting (AL, γ) is the optimal policy if and only if $\delta_A \geq \bar{\delta}_A$. If $\delta_A < \bar{\delta}_A$:*

1. *The optimal test setting is $(AL, \frac{2\beta - \delta_A(1+\beta)}{2(1-\delta_A)})$ if and only if*

$$\delta_P \leq \min \left\{ \frac{8}{11} \frac{\delta_A}{1 - \delta_A}, \frac{2 - \delta_A}{1 - \delta_A} \frac{1}{6} \right\}$$

2. *The optimal test setting is (MT, γ) if and only if*

$$\delta_P > \max \left\{ \frac{2 - \delta_A}{1 - \delta_A} \frac{1}{6}, \frac{16}{37} \right\}$$

3. *The optimal test setting is $(UT, \frac{11}{16})$ if and only if $\delta_A < \underline{\delta}_A$ and $\delta_P \in (\frac{8}{11} \frac{\delta_A}{1 - \delta_A}, \frac{16}{37}]$.*

To explain the intuition behind the result, it's convenient to split the space of $\delta_A \in (0, 1)$ into three parts defined by two thresholds $\underline{\delta}_A = \frac{22}{59}$ and $\bar{\delta}_A = \frac{4}{5}$, like I did in Figure 23. As long as the incompetent agent is highly reluctant to one period wait - $\delta_A < \underline{\delta}_A$ - the principal needs to set an extremely high attempt limit passing score $\hat{\sigma}_{AL}$, implying a large probability of a false negative error without any retake allowed. Therefore the principal chooses the attempt limit restriction for extremely low values of δ_P , $\delta_P \leq \frac{8}{18} \frac{\delta_A}{1 - \delta_A}$. when δ_P takes intermediate values, the principal chooses unconstrained testing and this region is represented by the yellow shaded area in panel (b) of Figure 23. For high values of δ_P , $\delta_P > \frac{16}{37}$, she ends up selecting mandatory training over unrestricted testing.

As the agent becomes less reluctant to the one-period wait, the principal can lower the attempt limit passing score $\hat{\sigma}_{AL}$ and the interval of δ_P values in which the principal prefers unrestricted testing over attempt limit shrinks. As soon as $\delta_A > \underline{\delta}_A$, the optimal restriction is either attempt limit or mandatory training: it can be shown that for any value of δ_P such that $UT \succsim_P AL$, then $MT \succsim_P UT$. This is

represented by the area between the two red dashed vertical lines in panel (a) of Figure 23. Thus, when

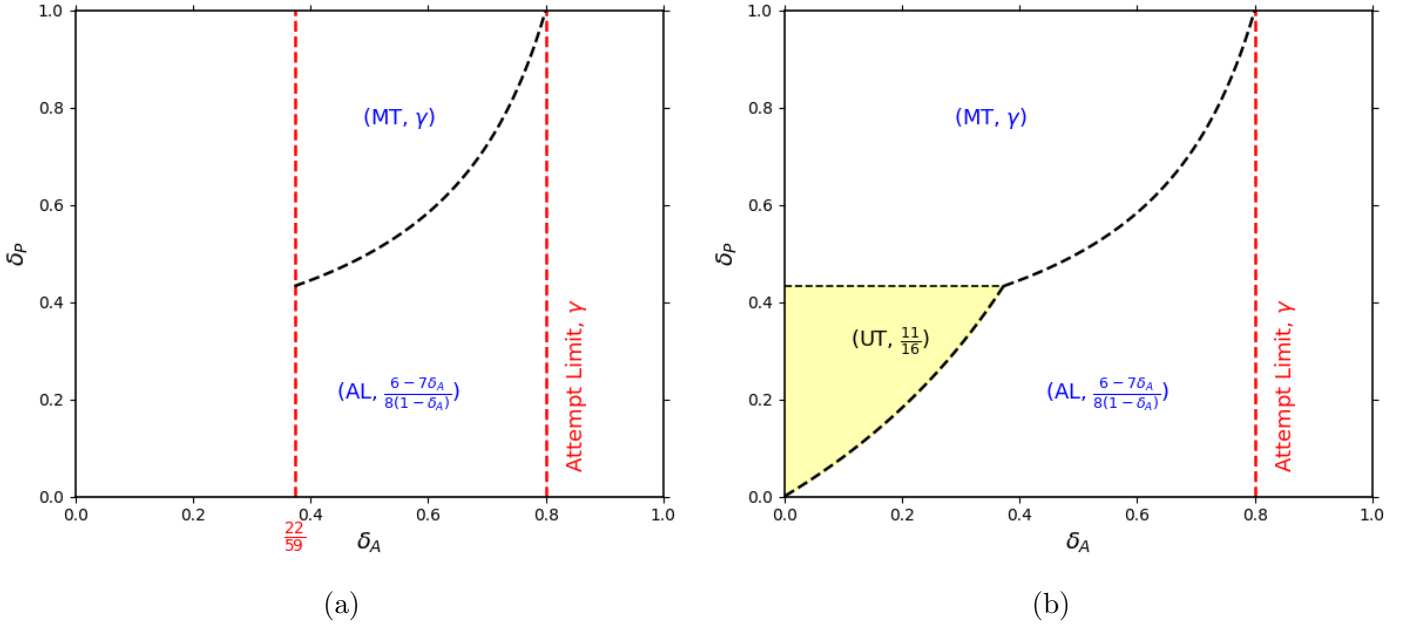


Figure 23: False positive *vs* False negative

the agent is moderately reluctant to wait one period, $\delta_A \in (\underline{\delta}_A, \bar{\delta}_A)$, the principal compares mandatory training with attempt limit. Both restrictions limit the number of attempts of competent people at one, but the attempt limit lets strong candidates take the test as soon as possible. On the other hand, the passing rate for the good agents is higher under mandatory training regime because $\hat{\sigma}_{MT} = \gamma < \hat{\sigma}_{AL}$. Hence, for $\delta_P > \frac{2-\delta_A}{1-\delta_A} \frac{1}{6}$, the principal prefers mandatory training over attempt limit and, the higher δ_A is, the higher δ_P must be for the principal to select mandatory training. Finally, when $\delta_A \geq \bar{\delta}_A$, the principal can set $\hat{\sigma}_{AL} = \gamma$ and attempt limit restriction dominates mandatory training.

6 Endogenous test signal

So far test signals have been treated as exogenous and I've analyzed principal's choice of test settings for each type of signal. As I said, there can be several types of signals depending on the kind of evidence that is disclosed. There are tests that can certify the incompetence of a candidate (*BCT*), other examinations certify the competence of test-takers (*GCT*) and finally there are tests that can do both (*TSCT*). However, it's not hard to think that sometimes, if the principal can introduce the attempt limit and select the passing score, they can choose the type of evidence to disclose too. In other

words, they can design a test with the intent of disclosing some specific candidate's type. Therefore, I suppose that at the beginning of the game the principal can choose not only the test restriction and the passing threshold, but also the test signal $\pi \in \{\pi_{BCT}, \pi_{GCT}, \pi_{TSCT}\}$. The timing of the game is:

1. Principal sets test settings (ξ, σ, π) ;
2. Competence level $\theta \in \{B, G\}$ is privately observed by the agent;
3. Agent decides whether or not to take the test;
4. Test outcome is realized, $\omega \in \{P, F\}$. If $\omega = P$, the game ends; if $\omega = F$, a new test session might take place, depending on test settings.

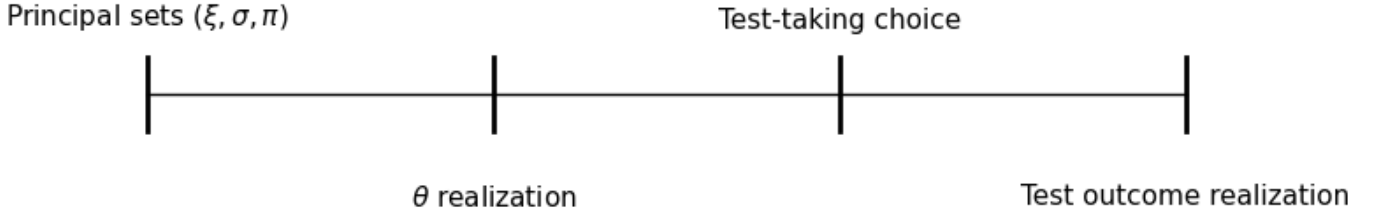


Figure 24: Timing of the events

Throughout this section I'm going to use γ and β to denote respectively the lower bound of the good type conditional distribution and the upper bound of the bad type conditional distribution. When the test signal changes, both γ and β change too. A passing threshold equal to γ means $\sigma = 0$ if the principal is using a good-certifying test and it means $\sigma = \frac{1}{2}$ if the principal is using a bad-certifying test. Also, to make notation easy, I'll denote passing threshold $\beta \frac{2(1-\delta_A)}{2-\delta_A(1+\beta)}$ by $\hat{\sigma}_{AL}^{GCT}$.

Proposition 11. *Assume that the principal can choose (ξ, σ, π) . There exist $\bar{\delta}_A$ and $\bar{\delta}_P$ such that*

- *If $\delta_A < \bar{\delta}_A$ and $\delta_P < \bar{\delta}_P$, the optimal test settings has a good-certifying test, $\hat{\pi} = \pi_{GCT}$, and*

$$(\hat{\xi}, \hat{\sigma}) = \begin{cases} (AL, \hat{\sigma}_{AL}^{GCT}), & \text{if } \delta_P \leq \frac{\delta_A}{2-\delta_A} \\ (UT, \beta) & \text{if } \delta_P > \frac{\delta_A}{2-\delta_A} \end{cases}$$

- If $\delta_A < \bar{\delta}_A$ and $\delta_P \geq \bar{\delta}_P$, the optimal policy is (MT, γ, π_{BCT}) .
- If $\delta_A \geq \bar{\delta}_A$, the optimal policy is (AL, γ, π_{BCT}) .

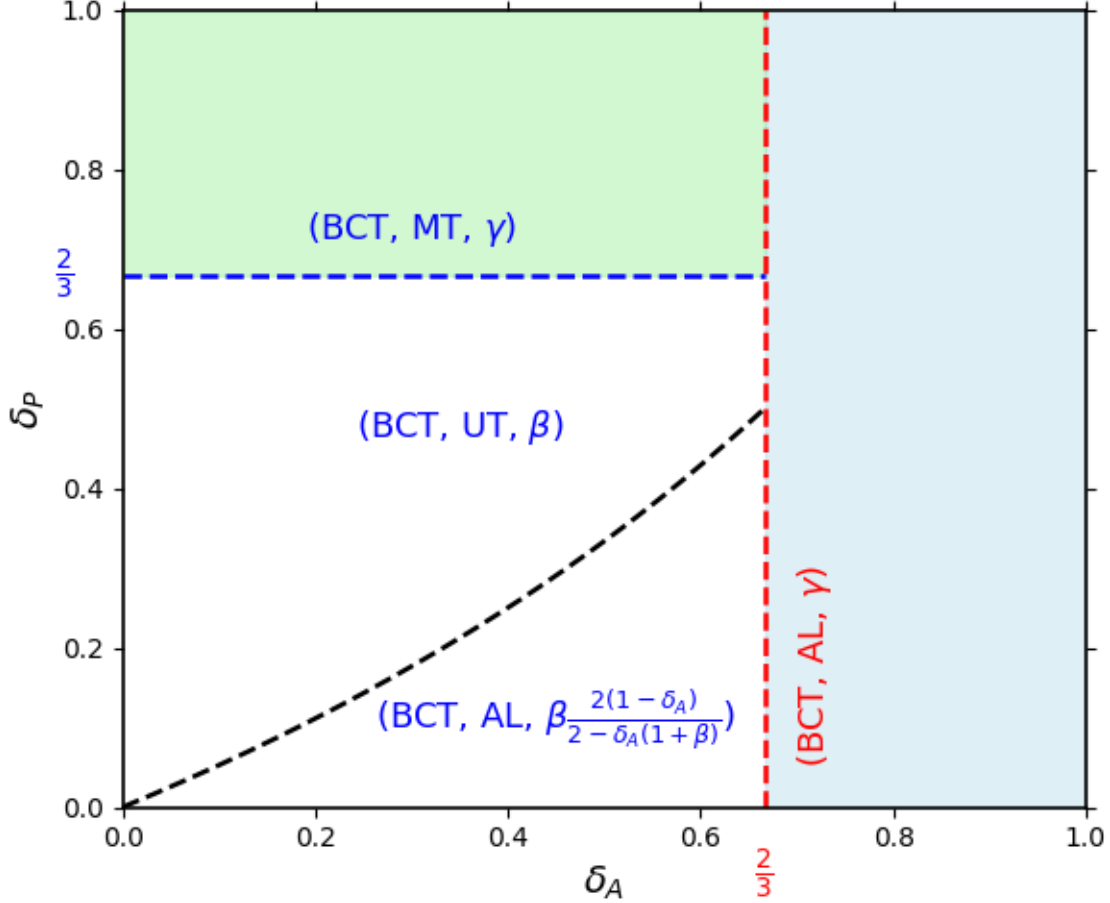
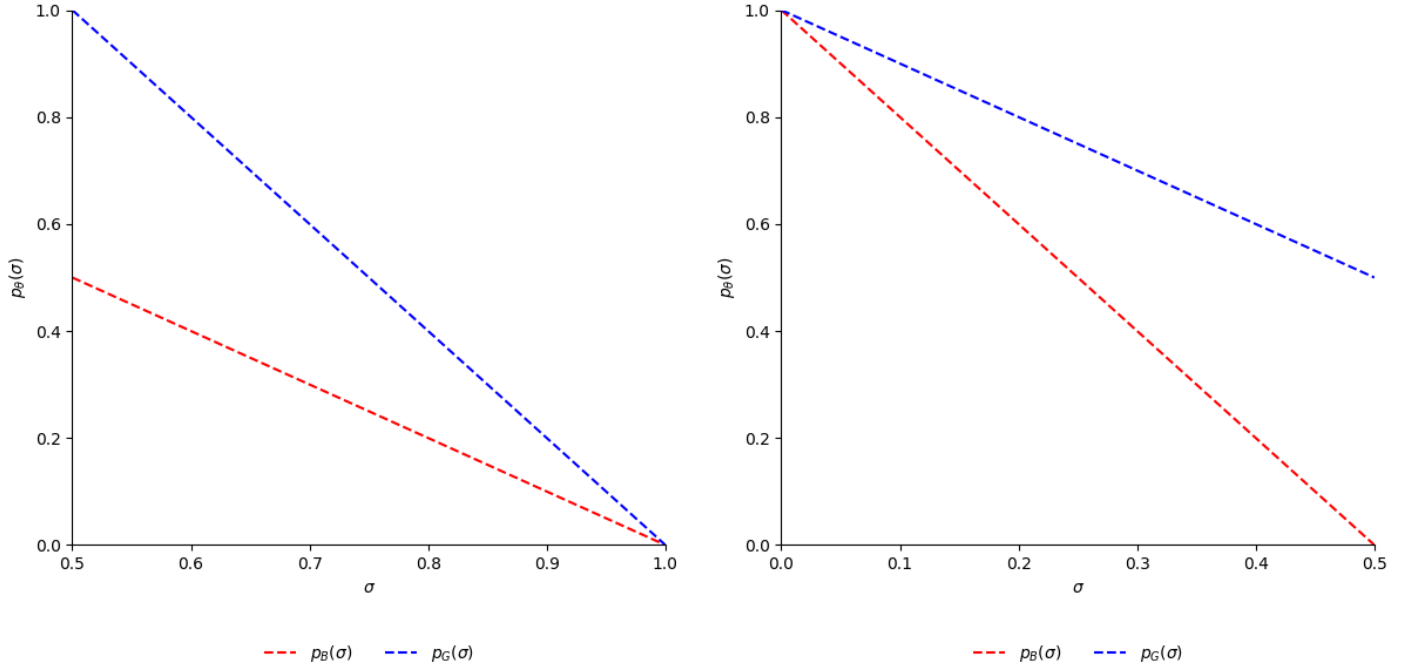


Figure 25: Endogenous test signal

From Figure 25 it's possible to see that the two-sided certifying signal is never part of the optimal test setting: principal chooses either a good or a bad-certifying test. Furthermore, when either δ_A or δ_P is sufficiently high, the principal tailors a test that is able to provide conclusive evidence about the incompetent agent; when instead both δ_P and δ_A are sufficiently small, the principal chooses to design a good-certifying test. I begin by illustrating the second part and later I'll explain why the two-sided certifying test is never chosen.

Before analyzing the choice between a bad and a good-certifying test, it's convenient to show that the trade-off between $p_B(\sigma)$ and $p_G(\sigma)$ differs by the type of test considered. When a good-certifying test is administered and the passing threshold σ is raised, $p_G(\sigma)$ diminishes less than $p_B(\sigma)$. On the other

hand, when a bad-certifying test is run and the passing threshold σ is lowered, $p_G(\sigma)$ grows more than $p_B(\sigma)$. Therefore, in a good-certifying test, the principal obtains an extremely low $p_B(\sigma)$ in exchange for a sufficiently high $p_G(\sigma)$ whereas, in bad-certifying test, she achieves an extremely high $p_G(\sigma)$ in exchange for a limited $p_B(\sigma)$.



(a) Bad-certifying test passing probabilities

(b) Good-certifying test passing probabilities

Figure 26: Passing probabilities

When $\delta_A \geq \frac{2}{3}$, the principal can conveniently choose (AL, γ, π_{BCT}) , which has the dual advantage of preventing the hiring of bad agents in the first period and, at the same time, avoiding false negatives in subsequent periods—at the cost of a moderate probability of false positives. By contrast, if the principal selects $(AL, \sigma(\beta), \pi_{GCT})$, she prevents the hiring of bad agents in the first period only by accepting a high probability of false positives in the next. Alternatively, under (UT, β, π_{GCT}) , the principal avoids selecting weak candidates but must tolerate a moderate probability of false negatives. For these reasons, the principal ultimately opts for (AL, γ, π_{BCT}) .

When $\delta_A < \frac{2}{3}$, the principal is not longer able to implement (AL, γ, π_{BCT}) . However she can still use a bad-certifying test and avoid hiring bad type agents by choosing (MT, γ, π_{BCT}) . Postponing the selection by one period helps her maximize the probability to test someone competent and setting $\sigma = \gamma$ allows her to fully capitalize the human capital growth from the one-period wait. For this setting to be

relatively more attractive, principal discount rate δ_P must be sufficiently high, that is, $\delta_P \geq \frac{2}{3}$.

If δ_P and δ_A are below $\frac{2}{3}$, the principal is unable to implement (AL, γ, π_{BCT}) and she does not find convenient postponing the test by using setting (MT, γ, π_{BCT}) . Therefore, she can implement either (UT, γ, π_{BCT}) or use any optimal good-certifying test. If the principal chooses (UT, γ, π_{BCT}) , she minimizes the probability of a false negative error but she doesn't allow for human capital growth because the bad type agent has high chances to pass the test in the first period. When the principal is sufficiently patient, she chooses (UT, β, π_{GCT}) : this way, she allows for human capital growth by setting a high passing threshold and, at the same time, she lets strong candidates take the test as many attempts as possible. When the individual is not patient enough, she implements $(AL, \sigma(\beta), \pi_{GCT})$: she can hire a strong candidate with a sufficiently high probability and avoid to select a bad type agent in the first period.

Now I illustrate why the two-sided-certifying test is never chosen. The main reason roots in the trade-off between $p_B(\sigma)$ and $p_G(\sigma)$ I talked about early this section. It's easy to see $\pi_{TSCT}(s|G)$ and $\pi_{TSCT}(s|B)$ have identical probability density functions. Therefore, unlike the bad and the good-certifying tests, $p_B(\sigma)$ and $p_G(\sigma)$ change the same way when the score is either raised or lowered in a two-sided certifying test. This can be easily illustrated by comparing the marginal probability rates of the three test signals. Given $MPR(\pi) = \frac{p'_G(\sigma; \pi)}{p'_B(\sigma; \pi)}$, the relationship between the marginal probability rates of the three test signals is represented by $MPR(\pi_{BCT}) > MPR(\pi_{TSCT}) > MPR(\pi_{GCT})$. Therefore, when $\delta_A \geq \frac{4}{5}$, both (AL, γ, π_{BCT}) and (AL, γ, π_{TSCT}) let the principal avoid hiring bad agents in the first period and select strong candidates with probability one in each period. However, she chooses π_{BCT} because $MPR(\pi_{BCT}) > MPR(\pi_{TSCT})$ implies that $p_G(\gamma) = 1$ is obtained in exchange for a smaller probability of a false positive error, that is, $p_B(\gamma; \pi_{BCT}) < p_B(\gamma; \pi_{TSCT})$. So $\pi_{BCT} \succsim \pi_{TSCT}$ for every $\delta_A \geq \frac{2}{3}$. Likewise, if $\delta_A < \frac{2}{3} \leq \delta_P$, the principal ends up choosing π_{BCT} over π_{TSCT} : both test settings (MT, γ, π_{BCT}) and (MT, γ, π_{TSCT}) - which are the best test settings for each type of test signal - let the principal avoid hiring bad agents in the first period and select strong candidates with probability one in each period, but $p_B(\gamma; \pi_{BCT}) < p_B(\gamma; \pi_{TSCT})$. When $\delta_A < \frac{2}{3}$ and $\delta_P < \frac{2}{3}$, it's not anymore convenient to avoid hiring the weak candidate in the first test session and, at the same time, keep at zero the probability of a false negative error. Since $MPR(\pi_{TSCT}) > MPR(\pi_{GCT})$, the principal finds optimal to use a good-certifying

test: this way, she can avoid hiring a bad candidate in the first period while keeping low the probability of a false negative error.

7 Conclusion

This paper has examined the rationale and welfare implications of limiting the number of attempts allowed in standardized testing. Motivated by evidence from the U.S. Bar Exam, I have shown that jurisdictions imposing an attempt limit achieve systematically higher passing rates among repeat test takers. Using cross-state variation generated by institutional persistence in testing rules following the adoption of the Uniform Bar Exam, I identified a possible causal relationship between the imposition of attempt limits and improved test performance among repeaters. This empirical pattern is consistent with the view that attempt limits act as a sorting mechanism, discouraging weak candidates from repeated participation while inducing stronger candidates to take the test only when sufficiently prepared.

The theoretical analysis provides a unified framework for understanding when such restrictions are welfare improving. In a two-period principal–agent model, the attempt limit operates as a screening device that affects candidates’ intertemporal incentives. The key insight is that the desirability of an attempt limit depends on the informativeness of the test, the candidate’s sensitivity to the attempt limit, and the principal’s aversion to human-capital loss. When the test is sufficiently informative—so that weak candidates are moderately sensitive to the limit while strong candidates retain a clear advantage—the attempt limit effectively deters premature participation and enhances the quality of selected candidates. If the test is either too noisy or too stringent, however, the policy ceases to be optimal: in the former case, the limit cannot be enforced; in the latter, it yields limited benefits relative to the cost of excluding competent individuals.

The analysis further shows that the optimal testing policy depends on the type of information the test conveys. When the test produces conclusive evidence of competence (a good-certifying test), attempt limits are desirable only if the principal places little weight on avoiding false negatives. Conversely, when the test produces conclusive evidence of incompetence (a bad-certifying test), the attempt limit becomes optimal when either the principal or the agent is sufficiently patient to delay testing. Extending the framework to allow for endogenous information design, I show that the two-sided-certifying test,

in which both high and low scores are informative, is always dominated by specialized designs. In equilibrium, the principal optimally commits to either a good- or bad-certifying signal, depending on the intertemporal preferences of both sides.

Taken together, the empirical and theoretical results demonstrate that attempt limits are not arbitrary constraints, but rather integral components of an optimal test design. By shaping the timing and preparedness of candidates, they help institutions manage the trade-off between false positives and false negatives, improving candidate selection.

References

- Abernethy, Jacob, and Rafael Frongillo (2012). “A Characterization of Scoring Rules for Linear Properties.” *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 27(1): 27–51.
- Angrist, Joshua D., Jonathan Guryan, and Parag Pathak (2023). “Accountability and Admissions: Evidence from Test-Based Reforms.” *American Economic Review*, 113(2): 485–522.
- Armstrong, Mark (1996). “Multiproduct Nonlinear Pricing.” *Econometrica*, 64(1): 51–75.
- Battaglini, Marco (2005). “Long-Term Contracting with Markovian Consumers.” *American Economic Review*, 95(3): 637–658.
- Banerjee, A., Dubey, A., and Tadelis, S. (2023). ”Testing and grading in the presence of effort and risk”, Working Paper
- Ben-Porath, Elchanan, Eddie Dekel, and Barton L. Lipman (2014). “Optimal Allocation with Costly Verification.” *American Economic Review*, 104(12): 3779–3813.
- Bergemann, D., Morris, S. (2016). ”Information design, Bayesian persuasion, and Bayes correlated equilibrium”. *American Economic Review*, 106(5), 586–591.
- Bizzotto, Jean, Jan Rüdiger, and Adrien Vigier (2020). “Testing, Disclosure and Approval.” *Journal of Economic Theory*, 187: 105002.
- Bull, Jesse, and Joel Watson (2007). “Hard Evidence and Mechanism Design.” *Games and Economic Behavior*, 58(1): 75–93.
- Carroll, Gabriel (2017). “Robustness and Separation in Multidimensional Screening.” *Econometrica*, 85(2): 453–488.
- Carroll, Gabriel, and Georgy Egorov (2019). “Strategic Communication with Minimal Verification.” *Econometrica*, 87(6): 1867–1892.
- Egorov, G. (2021). ”Dynamic mechanisms with evidence”. *Theoretical Economics*, 16(4), 1235–1280.
- Chade, H., Lewis, G., and Smith, L. (2011). Student portfolios and the college admissions problem. *Review of Economic Studies*, 78(3), 911–943.
- Dasgupta, Sulagna (2024). *Optimal Test Design for Knowledge-Based Screening*. Working Paper, University of Chicago.

- Dasgupta, Sulagna, and Zizhe Xia (2024). *Screening Knowledge with Verifiable Evidence*. Working Paper, University of Chicago and University of Bonn.
- Dee, Thomas, and Brian Jacob (2006). “The Impact of No Child Left Behind on Student Achievement.” *Journal of Policy Analysis and Management*, 30(3): 418–446.
- Deneckere, Raymond, and Sergei Severinov (2008). “Mechanism Design with Partially Verifiable Information.” *Games and Economic Behavior*, 64(2): 487–513.
- Dessein, W., Frankel, A., and Kartik, N. (2024). Communication, delegation, and testing. *American Economic Review*, forthcoming.
- Figlio, David, and Susanna Loeb (2011). “School Accountability.” *Handbook of the Economics of Education*, 3: 383–421.
- Forges, Françoise, and Frédéric Koessler (2005). “Communication Equilibria with Partially Verifiable Types.” *Journal of Mathematical Economics*, 41(7): 793–811.
- Fryer, R. G., and Loury, G. C. (2013). Valuing diversity. *Journal of Political Economy*, 121(4), 747–774.
- Gale, Douglas, and Martin Hellwig (1985). “Incentive-Compatible Debt Contracts: The One-Period Problem.” *Review of Economic Studies*, 52(4): 647–663.
- Goodman, Joshua (2021). “The Labor-Market Effects of Standardized Testing Policies.” *American Economic Journal: Applied Economics*, 13(4): 210–245.
- Green, Jerry, and Jean-Jacques Laffont (1986). “Partially Verifiable Information and Mechanism Design.” *Review of Economic Studies*, 53(3): 447–456.
- Haghpanah, Nima, and Jason Hartline (2021). “When Is Pure Bundling Optimal?” *Review of Economic Studies*, 88(3): 1127–1156.
- Hancart, Nathan (2022). *Designing the Optimal Menu of Tests*. Working Paper.
- Harbaugh, Rick, and Eric Rasmusen (2018). “Coarse Grades: Informing the Public by Withholding Information.” *American Economic Journal: Microeconomics*, 10(1): 210–235.
- Kamenica, E., Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6), 2590–2615.
- Kane, Thomas J., and Douglas O. Staiger (2002). “The Promise and Pitfalls of Using Imprecise School

- Accountability Measures.” *Journal of Economic Perspectives*, 16(4): 91–114.
- Kolotilin, A., Mylovanov, T., Zapechelnyuk, A., Li, M. (2017). Persuasion of a rationally inattentive agent. *Econometrica*, 85(6), 1949–1964.
- Lambert, Nicolas (2011). *Elicitation and Evaluation of Statistical Forecasts*. Preprint.
- Li, Yingkai, Jason D. Hartline, Liren Shan, and Yifan Wu (2022). “Optimization of Scoring Rules.” *Proceedings of the ACM Conference on Economics and Computation*, 988–989.
- Li, Yingkai, and Jonathan Libgober (2023). “Optimal Scoring for Dynamic Information Acquisition.” *arXiv preprint arXiv:2310.19147*.
- Manelli, Alejandro, and Daniel Vincent (2006). “Bundling as an Optimal Selling Mechanism for a Multiple-Good Monopolist.” *Journal of Economic Theory*, 127(1): 1–35.
- Osband, Kent, and Stefan Reichelstein (1985). “Information-Eliciting Compensation Schemes.” *Journal of Public Economics*, 27(1): 107–115.
- Pavan, Alessandro, Ilya Segal, and Juuso Toikka (2014). “Dynamic Mechanism Design: A Myersonian Approach.” *Econometrica*, 82(2): 601–653.
- Rosar, Frank (2017). “Test Design under Voluntary Participation.” *Games and Economic Behavior*, 104: 632–655.
- Rochet, Jean-Charles, and Philippe Choné (1998). “Ironing, Sweeping, and Multidimensional Screening.” *Econometrica*, 66(4): 783–826.
- Sher, Itai, and Rakesh Vohra (2015). “Price Discrimination through Communication.” *Theoretical Economics*, 10(2): 597–648.
- Spence, Michael (1973). “Job Market Signaling.” *Quarterly Journal of Economics*, 87(3): 355–374.
- Stiglitz, Joseph E. (1975). “The Theory of Screening, Education, and the Distribution of Income.” *American Economic Review*, 65(3): 283–300.
- Townsend, Robert M. (1979). “Optimal Contracts and Competitive Markets with Costly State Verification.” *Journal of Economic Theory*, 21(2): 265–293.
- Tyler, John H. (2020). “Testing, Accountability, and Human Capital.” *Journal of Human Resources*, 55(3): 789–826.
- Vrastosinos, Orestis (2025). *Multidimensional Screening of Strategic Candidates*. Working Paper, New

York University.

Weksler, Ran, and Boaz Zik (2022). “Informative Tests in Signaling Environments.” *Theoretical Economics*, 17(3): 977–1006.

Yang, Frank (2025a). “Costly Multidimensional Screening.” *Review of Economic Studies*, forthcoming.

Yang, Frank (2025b). “Nested Bundling.” *American Economic Review*, 115(9): 2970–3013.

8 Appendix

8.1 Proof of Proposition 1

To begin, I illustrate which contract menus are incentive compatible. Later, I determine which incentive compatible contract menu can be ruled out as an optimal restriction.

Lemma 2. *If an incentive compatible contract menu is separating, then it must be the attempt limit restriction.*

Proof. Any separating incentive compatible contract menu must have ξ^1 strictly increasing in the level of competence θ . Suppose that it's not true: then either $\xi_B^1 = \xi_G^1$ or $\xi_B^1 > \xi_G^1$. If $\xi_B^1 = \xi_G^1$, then $\xi_B^2 = \xi_G^2$; otherwise, ξ would not be incentive compatible. But $\xi_B^1 = \xi_G^1$ and $\xi_B^2 = \xi_G^2$ would contradict ξ being a separating contract menu. If $\xi_B^1 > \xi_G^1$, then $\xi_B^2 < \xi_G^2$; otherwise, the contract menu would not be incentive compatible. Because strong candidates remain strong throughout the game, type G would take the test as soon as possible. Hence, if $\xi_B^1 > \xi_G^1$, then ξ is not an incentive compatible contract menu. Since any separating contract menu must have $\xi_B^1 < \xi_G^1$, then it needs to have $\xi_B^2 > \xi_G^2$ because otherwise ξ is not incentive compatible. Then it's possible to conclude that if an incentive compatible contract menu is separating, then it must be the attempt limit restriction. \square

Therefore if a restriction is not the attempt limit contract menu and it's incentive compatible, it must be a pooling contract menu and, since every pooling contract menu is incentive compatible, any incentive compatible contract menu is either the attempt limit restriction or one of the following pooling contract menu:

$$\xi = (0, 0, 0, 0)$$

$$\xi = (0, 1, 0, 1)$$

$$\xi = (1, 0, 1, 0)$$

$$\xi = (1, 1, 1, 1)$$

Given the assumptions regarding type distribution, contract menus $\xi = (0, 0, 0, 0)$ and $\xi = (1, 0, 1, 0)$ are never optimal because unrestricted testing $\xi = (1, 1, 1, 1)$ makes the principal strictly better off. Thus, any optimal incentive compatible contract menu can be one of the following test restrictions: attempt limit, mandatory training and unrestricted testing.

8.2 Proof of Proposition 3

I split the proof into four steps: the first one is devoted to see when either attempt limit or mandatory training can be ruled out as optimal incentive compatible policies; the second step and the third steps are used to illustrate the sufficient and necessary conditions for the attempt limit to be the optimal incentive compatible test restriction; the last step illustrates when mandatory training is chosen by the principal. The optimality of unrestricted testing follows for the three steps.

Step 1. *If $p_B > p_G \frac{\delta_A}{2-\delta_A}$, attempt limit is not incentive compatible. If $p_B \leq \frac{\delta_A}{2-\delta_A}$, mandatory training is suboptimal.*

Proof. From Proposition 2, the attempt limit is incentive compatible if and only if $\delta_A \geq 2 \frac{p_B}{p_B + p_G}$ which can be rearranged as $p_B \leq p_G \frac{\delta_A}{2-\delta_A}$: thus, condition $p_B > p_G \frac{\delta_A}{2-\delta_A}$ implies that the attempt limit is not incentive compatible. If $p_B \leq p_G \frac{\delta_A}{2-\delta_A}$, the attempt limit is incentive compatible and mandatory training is always suboptimal because the attempt limit can postpone testing the bad agent as much as the mandatory training does and start testing strong candidates in the first period, sooner than the mandatory training. From now on, I denote $p_G \frac{\delta_A}{2-\delta_A}$ by p_{IC} . \square

Step 2. *Assume that the attempt limit is incentive compatible. There exists a p_{AL} such that the attempt limit is the optimal incentive compatible restriction if and only if $p_B \geq p_{AL}$.*

Proof. Given that the attempt limit is incentive compatible, Step 1 implies that mandatory is suboptimal. Principal compares the optimal policy by comparing her expected payoff from attempt limit and her expected payoff from unrestricted testing. She prefers attempt limit over unrestricted testing if and

only if

$$\frac{1}{1-\delta_P} \frac{p_G}{2} + \frac{\delta_P}{1-\delta_P} \frac{p_G - p_B}{4} \geq \frac{1}{1-\delta_P} \frac{p_G}{2} + \frac{\delta_P}{1-\delta_P} \frac{p_G(1-p_G)}{2} - \frac{1}{1-\delta_P} \frac{p_B}{2} + (1-p_B) \frac{\delta_P}{1-\delta_P} \frac{p_G - p_B}{4}, \quad (4)$$

that boils down to $\delta_P p_B^2 - (\delta_P p_G + 2)p_B + 2\delta_P p_G(1-p_G) \leq 0$, that is quadratic in p_B . The inequality is true for $p_B \in (p_B^-, p_B^+)$, where $p_B^+ = \frac{B+\sqrt{D}}{2\delta_P}$ and $p_B^- = \frac{B-\sqrt{D}}{2\delta_P}$, with $B = 2 + p_G\delta_P$ and $D = B^2 - 8\delta_P^2 p_G(1-p_G)$. It is worth to note that $p_B^+ = \frac{B+\sqrt{D}}{2\delta_P} > p_G$, because $B + \sqrt{D} \geq B = p_G\delta_P + 2 > 2\delta_P p_G$. At the same time, it can be noted that $p_B^- \in (0, p_G)$: indeed $f(p_B) = \delta_P p_B^2 - (\delta_P p_G + 2)p_B + 2\delta_P p_G(1-p_G)$ is positive at $p_B = 0$ and negative at $p_B = p_G$. Because $p_B \in [0, p_G]$ and $0 < p_B^- < p_G < p_B^+$, it's possible to conclude that, if the attempt limit is incentive compatible, then it's optimal if and only if $p_B \geq p_B^-$. From now on, root p_B^- is denoted by p_{AL} \square

Step 3. *There exists a $\bar{\delta}_A(p_G, \delta_P)$ such that the attempt limit is the optimal incentive compatible restriction if and only if $\delta_A \geq \bar{\delta}_A(p_G, \delta_P)$ and $p_B \in [p_{AL}, p_{IC}]$*

Proof. From Step 1, the attempt limit is incentive compatible if and only if $p_B \leq p_{IC}$. By Step 2, if the attempt limit is incentive compatible, then it is optimal if and only if $p_B \geq p_{AL}$. Therefore, the set of p_B values making the attempt limit incentive compatible needs to be larger than the set of p_B values for which the attempt limit is the best incentive compatible restriction, that is, $p_{AL} \leq p_{IC}$. This condition is true if and only if δ_A is greater than

$$\begin{aligned} \bar{\delta}_A(p_B, \delta_P) &= 2 \frac{p_{AL}}{p_G + p_{AL}} \\ &= 2 \frac{\delta_P p_G + 2 - \sqrt{(\delta_P p_G + 2)^2 - 8\delta_P^2 p_G(1-p_G)}}{3\delta_P p_G + 2 - \sqrt{(\delta_P p_G + 2)^2 - 8\delta_P^2 p_G(1-p_G)}} \end{aligned}$$

Thus, if $\delta_A \geq \bar{\delta}_A(p_B, \delta_P)$ and $p_B \in [p_{AL}, p_{IC}]$, the attempt limit is the best incentive compatible restriction. If $\delta_A < \bar{\delta}_A(p_B, \delta_P)$ or $p_B \notin [p_{AL}, p_{IC}]$, the attempt limit is not either incentive compatible or optimal. Hence, the attempt limit is the optimal incentive compatible restriction if and only if $\delta_A \geq \bar{\delta}_A(p_B, \delta_P)$ and $p_B \in [p_{AL}, p_{IC}]$. \square

Step 4. *There is a p_{MT} such that mandatory training is the optimal incentive compatible policy if and only if $p_B > \max\{p_{MT}, p_{IC}\}$.*

Proof. Suppose that $p_B > p_B^{IC}$. Step 1 implies that attempt limit is not incentive compatible. Hence, the principal chooses the optimal policy between mandatory and unrestricted testing. Mandatory training is chosen if and only if

$$\frac{\delta_P}{1-\delta_P} \left[\frac{3}{4}p_G - \frac{1}{4}p_B \right] > \frac{1}{1-\delta_P} \frac{p_G}{2} + \frac{\delta_P}{1-\delta_P} \frac{p_G(1-p_G)}{2} - \frac{1}{1-\delta_P} \frac{p_B}{2} + (1-p_B) \frac{\delta_P}{1-\delta_P} \frac{p_G - p_B}{4},$$

which is equivalent to $\delta_P p_B^2 - (\delta_P p_G + 2)p_B + 2p_G(1 - \delta_P p_G) < 0$. Since the left hand side of the inequality is quadratic in p_B , the inequality is true for $p_B \in (p_B^-, p_B^+)$, where $p_B^+ = \frac{B+\sqrt{D}}{2\delta_P}$ and $p_B^- = \frac{B-\sqrt{D}}{2\delta_P}$, with $B = 2 + p_G\delta_P$ and $D = 9\delta_P^2 p_G - 4\delta_P p_G + 4$. It is worth to note that $p_B^+ = \frac{B+\sqrt{D}}{2\delta_P} > p_G$, because $B + \sqrt{D} \geq B = p_G\delta_P + 2 > 2\delta_P p_G$. At the same time, it can be noted that $p_B^- \in (0, p_G)$, because $f(p_B) = \delta_P p_B^2 - (\delta_P p_G + 2)p_B + 2p_G(1 - \delta_P p_G)$ is positive at $p_B = 0$ and negative at $p_B = p_G$. Because $p_B \in [0, p_G]$ and $0 < p_B^- < p_G < p_B^+$, it's possible to conclude that, if $p_B > p_{IC}$, then mandatory training is the optimal policy if and only if $p_B \geq p_B^-$. From now on, root p_B^- is denoted by p_{MT} . Hence, if $p_B > \max\{p_{MT}, p_{IC}\}$, mandatory training is the optimal policy. If $p_B \leq \max\{p_{MT}, p_{IC}\}$, mandatory training is suboptimal: if $p_{MT} < p_{IC}$, then attempt limit is preferred to mandatory training by Step 1; if $p_{MT} > p_{IC}$, I've just shown that unrestricted testing is preferred to mandatory training. To conclude, mandatory training is the optimal policy if and only if $p_B > \max\{p_{MT}, p_{IC}\}$. \square

8.3 Proof of Proposition 4

I divide the proof into several claims, one for each comparative statics claim.

Claim 1. *Threshold p_{AL} is increasing in δ_P .*

Proof. Recall that threshold p_{AL} is the value of p_B such that $f(p_B, p_G) = \delta_P p_B^2 - (\delta_P p_G + 2)p_B + 2\delta_P p_G(1 - p_G) = 0$. The implicit function theorem implies $p'_B(\delta_P) = -\frac{f_{\delta_P}}{f_{p_B}}$. Given that $f_{p_B}(p_B, \delta_P) = 2\delta_P p_G - (2 + \delta_P p_G) < 0$, then $p'_B(\delta_P)$ is negative if and only if $f_{\delta_P}(p_B, \delta_P) < 0$. Thus, consider

$f_{\delta_P}(p_B, \delta_P) = p_B^2 - p_G \delta_P + 2p_G(1 - p_G)$ and evaluate it at $p_B = p_{AL}$:

$$\begin{aligned}
f_{\delta_P}(p_{AL}, \delta_P) &= p_{AL}^2 + 2p_G(1 - p_G) - p_G p_{AL} \\
&= \frac{\delta_P p_G + 2}{\delta_P} p_{AL} - p_G p_{AL} \\
&= p_{AL} \left[\frac{\delta_P p_G + 2}{\delta_P} - p_G \right] \\
&= p_{AL} \frac{2}{\delta_P} > 0
\end{aligned}$$

where the second equality is due to $f(p_{AL}, \delta_P) = 0$ implying $\delta_P p_{AL}^2 + 2\delta_P p_G(1 - p_G) = p_{AL}(\delta_P p_G + 2)$. \square

Claim 2. *Threshold p_{AL} is increasing in p_G if and only if $p_G < \bar{p}_G$.*

Proof. Recover the compact form of p_{AL} and differentiate it with respect to p_G : $\frac{\partial p_{AL}}{\partial p_G} = \frac{1}{2} \left[1 - \frac{9\delta_P p_G + 2 - 4\delta_P}{\sqrt{S}} \right]$, where $S = (2 + p_G \delta_P)^2 - 8\delta_P^2 p_G(1 - p_G)$. Hence, $\frac{\partial p_{AL}}{\partial p_G} > 0$ if and only if $\sqrt{S} \geq 9\delta_P p_G + 2 - 4\delta_P$. which is equivalent to $f(p_G) = 9\delta_P p_G^2 + 4(1 - 2\delta_P)p_G - 2(1 - \delta_P) < 0$. Because $f(0) < 0 < f(1)$, there exists a \bar{p}_G such that p_{AL} is increasing in p_G for $p_G \leq \bar{p}_G$ and decreasing otherwise. \square

Claim 3. *Threshold p_{MT} is increasing in δ_P .*

Proof. Recall that p_{MT} is the value of p_B such that $f(p_B, \delta_P) = \delta_P p_B^2 - (\delta_P p_G + 2)p_B + 2p_G(1 - \delta_P p_G) = 0$. The implicit function theorem implies that $p'_{MT}(\delta_P) = -\frac{f_{\delta_P}}{f_{p_B}}$. Because $f_{p_B} = 2\delta_P p_B - (\delta_P p_G + 2) < 0$, $p'_{MT}(\delta_P)$ is negative if and only if $f_{\delta_P}(p_B, \delta_P) < 0$. Since $f_{\delta_P}(p_B, \delta_P) = p_B^2 - p_G p_B - 2p_G^2 = p_G^2(p^2 - p - 2) = p_G^2(p - 2)(p + 1) < 0$ - with $p = \frac{p_B}{p_G}$ - it's possible to conclude that $p'_{MT}(\delta_P) < 0$. \square

Claim 4. *Threshold p_{MT} is increasing in p_G if and only if $\delta_P p_G \leq \frac{4}{9}$*

Proof. Differentiating the closed form of p_{MT} , I obtain $\frac{\partial p_{MT}}{\partial p_G} = \frac{1}{2} \left[1 - \frac{18\delta_P p_G - 4}{2\sqrt{D}} \right]$ where $D = 9\delta_P^2 p_G - 4\delta_P p_G + 4$. Therefore $p'_{MT}(p_G) > 0$ if and only if $\sqrt{D} > 9\delta_P p_G - 2$ which is equivalent to $\delta_P p_G < \frac{4}{9}$ \square

Claim 5. *Threshold $\bar{\delta}_A(p_G, \delta_P)$ is decreasing in p_G .*

Proof. Recall $\bar{\delta}_A(p_G, \delta_P) = 2 \frac{p_{AL}}{p_{AL} + p_G}$ and differentiate it with respect to p_G :

$$\frac{\partial \bar{\delta}_A}{\partial p_G} = 2 \frac{p_G p'_{AL} - p_{AL}}{(p_{AL} + p_G)^2}.$$

Hence, $\bar{\delta}_A$ is decreasing in p_G if and only if $p_G p'_{AL} - p_{AL} < 0$. Recall that threshold p_{AL} is the value of p_B such that $f(p_B, p_G) = \delta_P p_B^2 - (\delta_P p_G + 2)p_B + 2\delta_P p_G(1 - p_G) = 0$. Given that $f_{p_B}(p_B, p_G) = 2\delta_P p_B - (2 + \delta_P p_G)$ and $f_{p_G}(p_B, p_G) = -\delta_P p_B + 2\delta_P(1 - 2p_G)$, the implicit function theorem implies

$$\begin{aligned} p_G p'_{AL} - p_{AL} &= \frac{-p_G f_{p_G} - p_{AL} f_{p_B}}{f_{p_B}} \\ &= 2 \frac{\delta_P p_G p_{AL} + p_{AL} - \delta_P p_{AL}^2 - \delta_P p_G(1 - 2p_G)}{f_{p_B}} \\ &= -2 \frac{p_{AL} - \delta_P p_G}{f_{p_B}} \end{aligned}$$

where the third equality is due to $f(p_B, p_G) = 0$ implying $\delta_P p_{AL}^2 = (\delta_P p_G + 2)p_{AL} - 2\delta_P p_G(1 - p_G)$ and, after plugging the value of $\delta_P p_{AL}^2$ into the numerator, $\delta_P p_G p_{AL} + p_{AL} - \delta_P p_{AL}^2 - \delta_P p_G(1 - 2p_G) = \delta_P p_G p_{AL} + p_{AL} - [(\delta_P p_G + 2)p_{AL} - 2\delta_P p_G(1 - p_G)] - \delta_P p_G(1 - 2p_G) = \delta_P p_G - p_{AL}$. Since $f_{p_B} < 0$, it's possible to conclude that $p_G p'_{AL} - p_{AL} > 0$ if and only if $p_{AL} - \delta_P p_G > 0$. Because $f(\delta_P p_G, p_G) < 0$, it can be said that $p_{AL} < \delta_P p_G$ and it's possible to conclude that $\bar{\delta}_A$ is strictly decreasing in p_G . □

Claim 6. *Threshold $\bar{\delta}_A(p_G, \delta_P)$ is increasing in δ_P .*

Proof. Recall $\bar{\delta}_A(p_G, \delta_P) = 2 \frac{p_{AL}}{p_{AL} + p_G}$ and differentiate it with respect to δ_P :

$$\frac{\partial \bar{\delta}_A}{\partial \delta_P} = \frac{2p_G}{(p_G + p_{AL})^2} p'_{AL}(\delta_P) > 0,$$

given that $p'_{AL}(\delta_P) > 0$ by Claim 1. □

8.4 Proof of Proposition 5

The proof is articulated in three steps: I start by showing that, if $\frac{\gamma}{\beta} \geq 2\frac{1-\delta_A}{2-\delta_A}$, then the incentive compatibility constraint is satisfied for every $\sigma \in [\gamma, \beta]$; I continue by illustrating that, if $\frac{\gamma}{\beta} < 2\frac{1-\delta_A}{2-\delta_A}$, there exists a $\bar{\sigma} \in (\gamma, \beta]$ such that the incentive compatibility constraint is satisfied for every $\sigma \in [\bar{\sigma}, \beta]$; I conclude by showing that principal's utility is continuous and decreasing over the set of incentive compatible passing thresholds. The proposition follows from proving that the space of incentive compatible passing thresholds is compact (Step 1 and Step 2) and principal's utility is strictly decreasing over it (Step 3).

Step 1. *If $\frac{\gamma}{\beta} \geq 2\frac{1-\delta_A}{2-\delta_A}$, the attempt limit restriction is incentive compatible for every $\sigma \in [\gamma, \beta]$.*

Proof. Consider inequality (2) and plug $p_G(\gamma) = 1$ and $p_B(\gamma) = 1 - \frac{\gamma}{\beta}$ into it. It's easy to see that the inequality is satisfied at $\sigma = \gamma$ if and only if $\frac{\gamma}{\beta} \geq 2\frac{1-\delta_A}{2-\delta_A}$ is true. To show that the incentive compatibility constraint holds for an $\sigma \in [\gamma, \beta]$, take the difference between the left and the right side of inequality (2):

$$T(\sigma) = \delta_A \frac{p_G(\sigma) + p_B(\sigma)}{2} - p_B(\sigma)$$

The difference is a linear function of σ and its derivative can be either positive or negative. If $T'(\sigma) > 0$, then the incentive compatibility holds for every $\sigma \geq \gamma$, that is, $\sigma \in [\gamma, \beta]$. If $T'(\sigma) < 0$, then the inequality in (2) might not be true for some value $\sigma \in (\gamma, \beta]$; however, the incentive compatibility constraint always holds at $\sigma = \beta$, because $p_G(\beta) > p_B(\beta) = 0$. Thus, $T'(\sigma) < 0$ implies that the incentive compatibility constraint holds for every $\sigma \leq \beta$, that is $\sigma \in [\gamma, \beta]$. \square

Step 2. *If $\frac{\gamma}{\beta} < 2\frac{1-\delta_A}{2-\delta_A}$, the attempt limit is incentive compatible for every $\sigma \in [\bar{\sigma}, \beta]$, with $\bar{\sigma} > \gamma$.*

Proof. For any $\sigma \in (\gamma, \beta)$, consider the derivative of $T(\sigma)$:

$$T'(\sigma) = \frac{1}{\beta} - \frac{\delta_A}{2} \left(\frac{1}{1-\gamma} + \frac{1}{\beta} \right). \quad (5)$$

It's strictly positive if and only if $\gamma < 1 - \frac{\beta\delta_A}{2-\delta_A}$. It's easy to see that

$$\begin{aligned}
1 - \frac{\beta\delta_A}{2-\delta_A} - 2\beta\frac{1-\delta_A}{2-\delta_A} &= \frac{2-\delta_A-\beta\delta_A-2\beta(1-\delta_A)}{2-\delta_A} \\
&= \frac{2-\delta_A-2\beta+\beta\delta_A}{2-\delta_A} \\
&= \frac{(2-\delta_A)(1-\beta)}{2-\delta_A} \\
&= 1-\beta > 0
\end{aligned}$$

Thus, $2\beta\frac{1-\delta_A}{2-\delta_A} < 1 - \frac{\beta\delta_A}{2-\delta_A}$ and it's possible to conclude that the derivative in (5) is strictly positive because assumption $\frac{\gamma}{\beta} < 2\frac{1-\delta_A}{2-\delta_A}$ implies $\gamma < 2\beta\frac{1-\delta_A}{2-\delta_A} < 1 - \frac{\beta\delta_A}{2-\delta_A}$.

Since $p_G(\beta) \geq p_B(\beta) = 0$, the incentive compatibility constraint holds at β and it holds with strict inequality if $\beta - \gamma < 1$. Because the attempt limit is not incentive compatible at $\sigma = \gamma$, $T(\gamma) < 0$. Given that $T(\gamma) < 0 \leq T(\beta)$ and function $T(\sigma)$ is a continuous strictly increasing function of σ , it's possible to invoke the intermediate value theorem. If $\beta - \gamma = 1$, then the attempt limit is incentive compatible only at $\sigma = 1$; if $\beta - \gamma < 1$, there exists a $\bar{\sigma} \in (\gamma, \beta)$ such that the attempt limit is incentive compatible for every $\sigma \in [\bar{\sigma}, \beta]$. The value of $\bar{\sigma}$ is such that the incentive compatibility constraint is binding, $\bar{\sigma} = \beta\frac{(1-\gamma)(2-\delta_A)-\delta_A}{(1-\gamma)(2-\delta_A)-\beta\delta_A}$. \square

Step 3. *The utility function of the principal adopting attempt limit is decreasing in σ over the space of incentive compatible passing thresholds.*

Proof. If the principal adopts the attempt limit, her utility as a function of an incentive compatible passing threshold is

$$W_{AL}(\sigma; \gamma, \beta) = \frac{1}{2} \frac{1}{1-\delta_P} \frac{1-\sigma}{1-\gamma} + \frac{\delta_P}{1-\delta_P} \frac{1}{4} \left(\frac{1-\sigma}{1-\gamma} - \left(1 - \frac{\sigma}{\beta} \right) \right).$$

The derivative of $W_{AL}(\sigma; \gamma, \beta)$ with respect to σ is

$$\frac{-2\beta + \delta_P(1 - \gamma - \beta)}{4\beta(1 - \gamma)(1 - \delta_P)}$$

and it's negative if and only if $-2\beta + \delta_P(1 - \gamma - \beta) < 0$. Because $0 \leq \gamma \leq 1/2 \leq \beta \leq 1$ and $\delta_P \in (0, 1)$, it can be said that $\delta_P(1 - \gamma - \beta) < 1 - \gamma - \beta \leq 1 - \beta \leq \frac{1}{2}$. Therefore, $-2\beta + \delta_P(1 - \gamma - \beta) < -1 + \frac{1}{2} < -\frac{1}{2}$ \square

8.5 Proof of Proposition 6

I split the proof into three parts, one for each test signal: bad-certifying test, good-certifying test and two-sided certifying test. Given a test signal π with parameters γ and β , principal's utility as function of a passing threshold σ is

$$W_{UT}(\sigma; \gamma, \beta) = \frac{1}{2} \frac{1}{1 - \delta_P} \frac{1 - \sigma}{1 - \gamma} \left(1 + \delta_P \frac{\sigma - \gamma}{1 - \gamma} \right) - \frac{1}{2} \frac{1}{1 - \delta_P} \left(1 - \frac{\sigma}{\beta} \right) + \frac{1}{4} \frac{\delta_P}{1 - \delta_P} \frac{\sigma}{\beta} \left(\frac{1 - \sigma}{1 - \gamma} - \left(1 - \frac{\sigma}{\beta} \right) \right)$$

Given the assumptions on γ and β , it's easy to see that principal's utility is a concave function of σ : by taking the second derivative of $W_{UT}(\sigma)$ with respect to σ ,

$$W_{UT}''(\sigma; \gamma, \beta) = - \frac{\delta_P \left(2\beta^2 - (1 - \gamma)(1 - \beta - \gamma) \right)}{2\beta^2 (1 - \delta_P) (1 - \gamma)^2},$$

it's possible to state that $W_{UT}(\sigma; \gamma, \beta)$ is concave if and only if $2\beta^2 - (1 - \gamma)(1 - \beta - \gamma) \geq 0$ and, given $0 \leq \gamma \leq 1/2 \leq \beta \leq 1$, it can be said that $2\beta^2 - (1 - \gamma)(1 - \beta - \gamma) \geq \frac{1}{2} - \frac{1}{2} \geq 0$. Thus, the utility maximizing passing threshold can be found by looking at the first order conditions. To ease notation, when I fix the values of γ and β , I write the principal's utility function as $W_\xi(\sigma)$ with ξ denoting test restriction used by the principal.

Bad-certifying test. *The optimal passing threshold is $\hat{\sigma}_{UT} = \gamma$ for $\delta_P \leq 1/2$ and $\hat{\sigma}_{UT} \in (\gamma, \beta)$ otherwise.*

Proof. It's easy to rule out $\hat{\sigma}_{UT} = \beta = 1$ because it would mean failing everyone and principal's utility would be zero: she would be better off by lowering the passing threshold by an arbitrary $\varepsilon \in (0, 1)$.

Differentiating $W_{UT}(\sigma)$ with respect σ , $W'_{UT}(\sigma) = \frac{13\delta_P - 18\delta_P\sigma - 2}{4(1-\delta_P)}$, it can be seen that $W'_{UT}(1/2) \leq 0$ if and only if $\delta_P \leq 1/2$. \square

Good-certifying test. *The optimal passing threshold is $\hat{\sigma}_{UT} = \beta$.*

Proof. By fixing $\gamma = 0$ and $\beta = 1$, principal utility function is $W_{UT}(\sigma) = \frac{1+\delta_P}{1-\delta_P} \frac{\sigma}{2}$ which is strictly increasing in σ . Therefore, the principal sets σ up to the highest possible value, that is, $\sigma = \beta$. \square

Two-sided-certifying test. *The optimal passing threshold is $\hat{\sigma}_{UT} \in (\gamma, \beta)$.*

Proof. Fixing $\gamma = 0$ and $\beta = 1$ and simplifying the derivative with respect to σ gives $W'_{UT}(\sigma) = \frac{\delta_P}{9} \frac{11-16\sigma}{1-\delta_P}$. It's easy to verify that $W'_{UT}(1/4) > 0$ and $W'_{UT}(3/4) < 0$. Because $W_{UT}(\sigma)$ is a concave function of σ , it's possible to conclude the utility maximizing σ is an interior point, that is, $\hat{\sigma}_{UT} \in (\gamma, \beta)$, and, in particular, $\hat{\sigma}_{UT} = \frac{11}{16}$. \square

8.6 Proof of Proposition 7

Given a test signal π with parameters γ and β , principal's utility as function of a passing threshold σ is

$$W_{MT}(\sigma; \gamma, \beta) = \frac{\delta_P}{1-\delta_P} \left[\frac{3}{4} \left(\frac{1-\sigma}{1-\gamma} \right) - \frac{1}{4} \left(1 - \frac{\sigma}{\beta} \right) \right],$$

which is a linear function of σ . By taking the derivative of $W_{MT}(\sigma; \gamma, \beta)$ with respect to σ , it can be seen that $W'_{MT}(\sigma; \gamma, \beta) = \frac{\delta_P/4}{1-\delta_P} \left[\frac{1}{\beta} - \frac{3}{1-\gamma} \right] < 0$ if and only if $\beta > \frac{1-\gamma}{3}$. Because $0 \leq \gamma \leq 1/2 \leq \beta \leq 1$, it's easy to observe that $\frac{1-\gamma}{3} \leq \frac{1}{3} < \frac{1}{2} \leq \beta$. Thus, $W'_{MT}(\sigma; \gamma, \beta) < 0$ for every σ and the principal sets $\hat{\sigma}_{MT} = \gamma$.

8.7 Proof of Proposition 8

I split the proof into three steps: the first two are devoted to show that the optimal threshold is γ , regardless of the test restriction adopted; the last step is used to compare test settings.

Step 1. *The optimal mandatory training setting is better than the optimal unrestricted testing setting for $\delta_P \in [\frac{4}{9}, \frac{1}{2}]$.*

Proof. I assume $\delta_P \leq 1/2$. From Proposition 6 and 7, the optimal passing thresholds are $\hat{\sigma}_{UT} = \hat{\sigma}_{MT} = \gamma$ and, by plugging γ into the principal's payoff function, I conclude that $(MT, \gamma) \succeq (UT, \gamma)$ if

$$W_{MT}(\gamma) = \frac{\delta_P}{1 - \delta_P} \frac{5}{8} \geq \frac{1}{1 - \delta_P} \frac{4 + \delta_P}{16} = W_{UT}(\gamma)$$

which is equivalent to $\delta_P \geq \frac{4}{9}$. Therefore, MT is preferred to UT for $\delta_P \in [\frac{4}{9}, \frac{1}{2}]$. \square

Step 2. *The optimal mandatory training setting is better than the optimal unrestricted testing setting for $\delta_P > \frac{1}{2}$.*

Proof. Consider $W_{UT}(\sigma)$ and $W_{MT}(\sigma)$ and evaluate them at their optimal thresholds $\hat{\sigma}_{UT}$ and $\hat{\sigma}_{MT}$. By the envelope theorem, their derivatives are such that

$$\frac{\partial W_{MT}(\hat{\sigma}_{MT})}{\partial \delta_P} = \frac{5}{8} \left(\frac{1}{1 - \delta_P} \right)^2 > \frac{45\delta_P^2 + 8\delta_P - 4}{[12\delta_P(1 - \delta_P)]^2} = \frac{\partial W_{UT}(\hat{\sigma}_{UT})}{\partial \delta_P}$$

for every $\delta_P \geq \frac{1}{2}$. Since $W_{MT}(\gamma) > W_{UT}(\gamma)$ at $\delta_P = \frac{1}{2}$ and $W_{MT}(\gamma)$ increases in δ_P faster than $W_{UT}(\hat{\sigma}_{UT})$, I conclude that $MT \succeq_P UT$ for $\delta_P \geq \frac{4}{9}$. \square

Step 3. *The optimal test threshold is independent of the test restriction and the optimal test restrictions are chosen according to Proposition 8.*

Proof. From Proposition 5, I know that the optimal attempt limit threshold is $\hat{\sigma}_{AL} = \gamma$ for $\delta_A \geq \frac{2}{3}$ and $\hat{\sigma}_{AL} = 1$ for $\delta_A < \frac{2}{3}$. If $\hat{\sigma}_{AL} = 1$, then $(MT, \gamma) \succ (AL, 1)$ and test setting $(MT, 1)$ is never implemented. Steps 1 and 2 imply that $\hat{\sigma}_{UT} > \gamma$ is never used. Then the optimal passing threshold is γ , regardless of the restriction implemented, and the passing probabilities are constant in equilibrium. By Proposition 3, if $\delta_A \geq \frac{2}{3}$, mandatory training and unrestricted testing are never chosen in equilibrium. If $\delta_A < \frac{2}{3}$, Steps 1 and 2 imply that UT is chosen if $\delta_P < \frac{4}{9}$, whereas MT is selected if $\delta_P \geq \frac{4}{9}$. \square

8.8 Proof of Proposition 9

First, I show that MT is always suboptimal. Then Proposition 9 follows from comparing UT with AL . From Propositions 6 and 7, principal chooses $\hat{\sigma}_{UT} = \gamma$ and $\hat{\sigma}_{MT} = \beta$. By plugging the optimal thresholds into the principal's payoff function,

$$W_{UT}(\gamma) = \frac{1}{1 - \delta_P} \frac{1 + \delta_P}{4} > \frac{1}{1 - \delta_P} \frac{\delta_P}{2} = W_{MT}(\beta),$$

it's possible to conclude that MT is suboptimal for any value of $\delta_P \in (0, 1)$. Thus, the principal ends up choosing between two policies: she compares (UT, β) with the optimal setting making use of the attempt limit. From Corollary 1, the optimal attempt limit threshold is $\hat{\sigma}_{AL} = 2\beta \frac{1 - \delta_A}{2 - (1 + \beta)\delta_A}$ and, after plugging $\hat{\sigma}_{AL}$ into $W_{AL}(\sigma)$, it can be seen that the principal chooses the AL setting over the UT one if

$$W_{AL}(\hat{\sigma}_{AL}) = \frac{1}{1 - \delta_P} \frac{2 - \delta_A + \delta_P(1 - \delta_A)}{2(4 - 3\delta_A)} \geq \frac{1}{1 - \delta_P} \frac{1 + \delta_P}{4} = W_{UT}(\beta),$$

which is equivalent to $\delta_P \leq \frac{\delta_A}{2 - \delta_A}$.

8.9 Proof of Proposition 10

I split the proof into four steps: in the first step I show that (AL, γ) is the optimal test setting for every $\delta_A \geq \frac{4}{5}$; in the second step I illustrate that there is a value $\bar{\delta}_P(\delta_A)$ such that unrestricted testing is suboptimal for $\delta_P \leq \bar{\delta}_P$; the third step shows the optimality conditions for mandatory training to be the optimal test restriction; finally, I prove the rest of the proposition.

Step 1. *If $\delta_A \geq \frac{4}{5}$, the optimal policy is (AL, γ) .*

Proof. Assume that $\delta_A \geq \frac{4}{5}$. Proposition 5 and 7 imply that the optimal passing threshold are $\hat{\sigma}_{AL} = \gamma$ and $\hat{\sigma}_{MT} = \gamma$; because both test restrictions have the same identical threshold - and equal passing probabilities - attempt limit is preferred to mandatory training by Proposition 3. By Proposition 6, the optimal unrestricted passing threshold is an interior solution $\hat{\sigma}_{UT} \in (\gamma, \beta)$. It can be seen that attempt

limit is always preferred to unrestricted testing:

$$\begin{aligned}
W_{AL}(\gamma) - W_{UT}(\hat{\sigma}_{UT}) &= \frac{1}{2} \frac{1}{1 - \delta_P} + \frac{\delta_P}{1 - \delta_P} \frac{\Delta p_\theta(\gamma)}{4} \\
&\quad - \left[\frac{1}{2} \frac{\Delta p_\theta(\hat{\sigma}_{UT})}{1 - \delta_P} + \frac{\delta_P}{1 - \delta_P} \frac{p_G(\hat{\sigma}_{UT})(1 - p_G(\hat{\sigma}_{UT}))}{2} + \frac{\delta_P}{1 - \delta_P} \frac{(1 - p_B(\hat{\sigma}_{UT}))\Delta p_\theta(\hat{\sigma}_{UT})}{4} \right] \\
&= \frac{1}{2} \frac{1 - p_G(\hat{\sigma}_{UT})}{1 - \delta_P} + \frac{1}{2} \frac{p_B(\hat{\sigma}_{UT})}{1 - \delta_P} + \frac{\delta_P}{1 - \delta_P} \frac{\Delta p_\theta(\gamma)}{4} \\
&\quad - \left[\frac{\delta_P}{1 - \delta_P} \frac{p_G(\hat{\sigma}_{UT})(1 - p_G(\hat{\sigma}_{UT}))}{2} + \frac{\delta_P}{1 - \delta_P} \frac{(1 - p_B(\hat{\sigma}_{UT}))\Delta p_\theta(\gamma)}{4} \right] \\
&= \frac{1}{2} \frac{(1 - p_G(\hat{\sigma}_{UT}))(1 - \delta_P p_G(\hat{\sigma}_{UT}))}{1 - \delta_P} + \frac{1}{2} \frac{p_B(\hat{\sigma}_{UT})}{1 - \delta_P} \left(1 + \frac{\delta_P \Delta p_\theta(\gamma)}{2} \right) > 0,
\end{aligned}$$

where the second equality is due to $\beta = 1 - \gamma$ implying $\Delta p_\theta(\sigma)$ is constant over $[\gamma, \beta]$. \square

Step 2. *Unrestricted testing is suboptimal for $\delta_P \leq \max\{\frac{8}{11} \frac{\delta_A}{1 - \delta_A}, 1\}$.*

Proof. I've just shown that unrestricted testing is suboptimal for $\delta_A \geq \frac{4}{5}$ and, in fact, $\frac{8}{11} \frac{\delta_A}{1 - \delta_A} > 1$ for every $\delta_A \geq \frac{4}{5}$. Assume that $\delta_A < \frac{4}{5}$. Proposition 5 implies that the optimal passing threshold for the attempt limit restriction is $\hat{\sigma}_{AL} = \beta \frac{(1 - \gamma)(2 - \delta_A) - \delta_A}{(1 - \gamma)(2 - \delta_A) - \beta \delta_A}$ for $\delta_A < \frac{4}{5}$. It can be shown that the interior solution for the optimal passing threshold while using unrestricted testing is $\hat{\sigma}_{UT} = \frac{11}{16}$. The attempt limit is preferred to unrestricted testing if and only if

$$W_{AL}(\hat{\sigma}_{AL}) = \frac{\delta_P(1 - \delta_A) + 2 - \delta_A}{12(1 - \delta_P)(1 - \delta_A)} \geq \frac{16 + 19\delta_P}{96(1 - \delta_P)} = W_{UT}(\hat{\sigma}_{UT}),$$

which is equivalent to $\delta_P \leq \delta_{AL}(\delta_A) = \frac{8}{11} \frac{\delta_A}{1 - \delta_A}$. \square

Step 3. *For $\delta_A < \frac{4}{5}$, mandatory training is the optimal restriction if and only if $\delta_P > \max\{\frac{2 - \delta_A}{1 - \delta_A} \frac{1}{6}, \frac{16}{37}\}$.*

Proof. For the mandatory training to be optimal, it must be that $\frac{\delta_P}{1 - \delta_P} \frac{7}{12} > \max\{W_{AL}(\hat{\sigma}_{AL}), W_{UT}(\hat{\sigma}_{UT})\}$, which is equivalent to $\delta_P > \max\{\frac{2 - \delta_A}{1 - \delta_A} \frac{1}{6}, \frac{16}{37}\}$. Notice that $\delta_P > \frac{16}{37}$ is the condition for $W_{MT}(\hat{\sigma}_{MT}) > W_{UT}(\hat{\sigma}_{UT})$, while $\delta_P > \frac{2 - \delta_A}{1 - \delta_A} \frac{1}{6}$ guarantees $W_{MT}(\hat{\sigma}_{MT}) > W_{AL}(\hat{\sigma}_{AL})$. \square

Step 4. *There is a $\underline{\delta}_A < \frac{4}{5}$ such that unrestricted testing is the optimal test restriction if and only if $\delta_A < \underline{\delta}_A$ and $\delta_P \in (\frac{8}{11} \frac{\delta_A}{1-\delta_A}, \frac{16}{37}]$. The attempt limit is the optimal test restriction if and only if $\delta_P < \min\{\frac{8}{11} \frac{\delta_A}{1-\delta_A}, \frac{2-\delta_A}{1-\delta_A} \frac{1}{6}\}$*

Proof. The proof is split into five steps. The first three steps are used to prove that unrestricted testing is the best policy for $\delta_A \leq \underline{\delta}_A$ and $\delta_P \in (\frac{8}{11} \frac{\delta_A}{1-\delta_A}, \frac{16}{37}]$. Step 4.4 is used to show that unrestricted testing is suboptimal for $\delta_A > \underline{\delta}_A$. Finally, Steps 4.3 and 4.5 show that attempt limit is the optimal restriction for $\delta_P < \min\{\frac{8}{11} \frac{\delta_A}{1-\delta_A}, \frac{2-\delta_A}{1-\delta_A} \frac{1}{6}\}$.

Step 4.1. *There exists a $\underline{\delta}_A < \bar{\delta}_A$, such that $\delta_{AL}(\delta_A) < \frac{16}{37}$ if and only if $\delta_A < \underline{\delta}_A$*

Proof. Since $\delta_{AL}(\delta_A)$ is a continuous and strictly increasing function of δ_A , with $\delta_{AL}(0) = 0$ and $\delta_{AL}(\frac{11}{19}) = 1$, the intermediate value theorem implies the existence of a $\underline{\delta}_A$ such that $\delta_{AL}(\delta_A) < \frac{16}{37}$ for $\delta_A < \underline{\delta}_A$ and $\delta_{AL}(\delta_A) > \frac{16}{37}$ otherwise. Because $\underline{\delta}_A < \frac{11}{19} < \frac{4}{5} = \bar{\delta}_A$, then $\underline{\delta}_A < \bar{\delta}_A$ \square

Step 4.2. *If $\delta_A < \underline{\delta}_A$, then $\frac{2-\delta_A}{1-\delta_A} \frac{1}{6} < \frac{16}{37}$.*

Proof. Suppose that this is not true. Then there exists a $\delta_A < \underline{\delta}_A$, such that $W_{AL}(\hat{\sigma}_{AL}) > W_{MT}(\gamma) > W_{UT}(\hat{\sigma}_{UT})$ for $\delta_P \in (\frac{16}{37}, \frac{2-\delta_A}{1-\delta_A} \frac{1}{6})$: indeed $\delta_P < \frac{2-\delta_A}{1-\delta_A} \frac{1}{6}$ implies $W_{AL}(\hat{\sigma}_{AL}) > W_{MT}(\gamma)$ and $\delta_P > \frac{16}{37}$ means $W_{MT}(\gamma) > W_{UT}(\hat{\sigma}_{UT})$. Since I am assuming that $\delta_A < \underline{\delta}_A$, Step 4.1 implies that $\delta_{AL}(\delta_A) < \frac{16}{37}$. By Step 2, $W_{UT}(\hat{\sigma}_{UT}) > W_{AL}(\hat{\sigma}_{AL})$ for every $\delta_P > \delta_{AL}(\delta_A)$: thus, I can conclude that $W_{UT}(\hat{\sigma}_{UT}) > W_{AL}(\hat{\sigma}_{AL})$ for $\delta_P > \frac{16}{37}$, but this contradicts the initial claim that $W_{AL}(\hat{\sigma}_{AL}) > W_{UT}(\hat{\sigma}_{UT})$ for $\delta_P \in (\frac{16}{37}, \frac{2-\delta_A}{1-\delta_A} \frac{1}{6})$. \square

Step 4.3. *If $\delta_A < \underline{\delta}_A$, then $\frac{8}{11} \frac{\delta_A}{1-\delta_A} < \frac{2-\delta_A}{1-\delta_A} \frac{1}{6}$.*

Proof. Suppose that this is not true. Then there exists a $\delta_A < \underline{\delta}_A$, such that $W_{AL}(\hat{\sigma}_{AL}) < W_{MT}(\gamma) < W_{UT}(\hat{\sigma}_{UT})$ for $\delta_P \in (\frac{2-\delta_A}{1-\delta_A} \frac{1}{6}, \frac{8}{11} \frac{\delta_A}{1-\delta_A})$: on the one hand $\delta_P > \frac{2-\delta_A}{1-\delta_A} \frac{1}{6}$ implies $W_{AL}(\hat{\sigma}_{AL}) < W_{MT}(\gamma)$; on the other hand, $\delta_P < \frac{8}{11} \frac{\delta_A}{1-\delta_A}$ implies $W_{MT}(\gamma) < W_{UT}(\hat{\sigma}_{UT})$, because $\frac{8}{11} \frac{\delta_A}{1-\delta_A} < \frac{16}{37}$ by Step 4.1 and $W_{MT}(\gamma) < W_{UT}(\hat{\sigma}_{UT})$ for $\delta_P < \frac{16}{37}$. However, $\delta_P < \frac{8}{11} \frac{\delta_A}{1-\delta_A}$ implies $W_{AL}(\hat{\sigma}_{AL}) > W_{UT}(\hat{\sigma}_{UT})$ by Step 2. Thus, a contradiction. \square

Step 4.4. *If $\delta_A \geq \underline{\delta}_A$, then unrestricted testing is never optimal.*

Proof. Suppose that this is not true. If unrestricted testing is optimal, then $\delta_P > \delta_{AL}(\delta_A)$ by Step 2. However, if $\delta_P > \delta_{AL}(\delta_A)$, then $\delta_P > \frac{11}{37}$ by Step 4.1. Because $\delta_P > \frac{11}{37}$ implies $W_{MT}(\gamma) > W_{UT}(\hat{\sigma}_{UT})$, a contradiction is reached. Thus, unrestricted testing cannot be optimal. \square

Step 4.5. *If $\delta_A \geq \underline{\delta}_A$, then $\delta_{AL}(\delta_A) > \frac{2-\delta_A}{1-\delta_A} \frac{1}{6}$.*

Proof. Suppose that this is not true. Thus, for some $\delta_A > \underline{\delta}_A$, $W_{MT}(\gamma) < W_{AL}(\hat{\sigma}_{AL}) < W_{UT}(\hat{\sigma}_{UT})$ for $\delta_P \in (\delta_{AL}(\delta_A), \frac{2-\delta_A}{1-\delta_A} \frac{1}{6})$: indeed $W_{MT}(\gamma) < W_{AL}(\hat{\sigma}_{AL})$ for $\delta_A < \frac{2-\delta_A}{1-\delta_A} \frac{1}{6}$ and $W_{AL}(\hat{\sigma}_{AL}) < W_{UT}(\hat{\sigma}_{UT})$ for $\delta_P > \delta_{AL}(\delta_A)$. However, this contradicts Step 4.4 claiming that unrestricted testing is suboptimal for $\delta_A > \underline{\delta}_A$. \square

\square

8.10 Proof of Proposition 11

I split the proof into three steps: in the first step I show that the principal chooses (AL, γ, π_{BCT}) ; in the second step I prove that (MT, γ, π_{BCT}) for $\delta_A < \frac{2}{3} \leq \delta_P$; finally, I conclude by illustrating that if both δ_A and δ_P are below $\frac{2}{3}$, the optimal test signal is π_{GCT} . In this section I denote by $W_\xi(\sigma; \pi)$ the utility function of a principal choosing test restriction $\xi \in \{AL, UT, MT\}$, passing threshold $\sigma \in [\gamma, \beta]$ and test signal $\pi \in \{\pi_{BCT}, \pi_{GCT}, \pi_{TSCT}\}$.

Step 1. *If $\delta_A \geq \frac{2}{3}$, the optimal test setting is (AL, γ, π_{BCT})*

Proof. First, I show that any setting involving π_{GCT} is suboptimal. Then I do the same for π_{TSCT} .

Step 1.1. *If $\delta_A \geq \frac{2}{3}$, any test setting involving π_{GCT} is suboptimal.*

Proof. Proposition 8 implies that the optimal test settings for a bad-certifying test is (AL, γ) . Proposition 9 states that, if $\delta_P \leq \frac{\delta_A}{2-\delta_A}$, the optimal setting is $(AL, \hat{\sigma}_{AL}^{GCT})$ - where $\hat{\sigma}_{AL}^{GCT} = \beta \frac{(1-\gamma)(2-\delta_A)-\delta_A}{(1-\gamma)(2-\delta_A)-\beta\delta_A}$ -

and, if $\delta_P > \frac{\delta_A}{2-\delta_A}$, the optimal setting is (AL, β) :

$$W_{AL}(\gamma, \pi_{BCT}) = \frac{1}{2} \frac{1}{1-\delta_P} \cdot 1 + \frac{\delta_P}{1-\delta_P} \cdot \frac{1}{4} \left(1 - \frac{1}{2}\right) = \frac{4+\delta_P}{8(1-\delta_P)}$$

$$W_{UT}(\beta, \pi_{GCT}) = \frac{1+\delta_P}{4(1-\delta_P)}$$

$$W_{AL}(\hat{\sigma}_{AL}^{GCT}, \pi_{GCT}) = \frac{1}{1-\delta_P} \left[\frac{1}{2} (1 - \hat{\sigma}_{AL}^{GCT}) + \frac{\delta_P}{4} \hat{\sigma}_{AL}^{GCT} \right]$$

It can be seen that for every $\delta_P \in (0, 1)$, test setting (AL, γ, π_{BCT}) dominates any setting using π_{GCT}

$$W_{AL}(\gamma, \pi_{BCT}) - W_{UT}(\beta, \pi_{GCT}) = \frac{4+\delta_P}{8(1-\delta_P)} - \frac{1+\delta_P}{4(1-\delta_P)} = \frac{2-\delta_P}{8(1-\delta_P)} > 0$$

$$\begin{aligned} W_{AL}(\gamma, \pi_{BCT}) - W_{AL}(\hat{\sigma}_{AL}^{GCT}, \pi_{GCT}) &= \frac{4+\delta_P}{8(1-\delta_P)} - \frac{1}{1-\delta_P} \left[\frac{1}{2} (1 - \hat{\sigma}_{AL}^{GCT}) + \frac{\delta_P}{4} \hat{\sigma}_{AL}^{GCT} \right] \\ &= \frac{1}{1-\delta_P} \frac{\hat{\sigma}_{AL}^{GCT}}{2} + \frac{\delta_P}{1-\delta_P} \frac{1}{16} \frac{\delta_A}{2-\frac{3}{2}\delta_A} > 0 \end{aligned}$$

□

Step 1.2. If $\delta_A \geq \frac{2}{3}$, any test setting involving π_{TSCT} is suboptimal.

Proof. I start with assuming that $\delta_A \geq \frac{4}{5}$. Setting (AL, γ, π_{BCT}) is better than (AL, γ, π_{TSCT}) because both are able to postpone testing the bad agent by one period and always pass strong candidates, but the former is better at failing weak candidates. If $\delta_A \in [\frac{2}{3}, \frac{4}{5})$, Proposition 10 states that the optimal two-sided certifying test policy is either (MT, γ) or $(AL, \hat{\sigma}_{AL}^{TSCT})$. If it's the former, then $(AL, \gamma, \pi_{BCT}) \succ (MT, \gamma, \pi_{TSCT})$: given that both policies provide a utility level that is constant for $\delta_A \geq \frac{2}{3}$, the difference between the two remains the same over this interval; since (AL, γ, π_{BCT}) is the optimal policy for $\delta_A \geq \frac{4}{5}$, then it must be better than (MT, γ, π_{TSCT}) for $\delta_A \geq \frac{2}{3}$. If it's the latter, then $(AL, \gamma, \pi_{BCT}) \succ (AL, \hat{\sigma}_{AL}^{TSCT}, \pi_{TSCT})$: indeed the level of utility provided by $(AL, \hat{\sigma}_{AL}^{TSCT}, \pi_{TSCT})$ is

$$W_{AL}(\hat{\sigma}_{AL}^{TSCT}, \pi_{TSCT}) = \frac{1}{2} \frac{1}{1-\delta_P} \frac{1-\hat{\sigma}_{AL}^{TSCT}}{1-\gamma} + \frac{\delta_P}{1-\delta_P} \frac{\Delta p_\theta(\hat{\sigma}_{AL}^{TSCT})}{4},$$

since $\Delta p_\theta(\sigma)$ is constant over $[\gamma, \beta]$ in a two-sided certifying test and the first term is increasing in δ_A , then $W_{AL}(\hat{\sigma}_{AL}^{TSCT}, \pi_{TSCT})$ is increasing in δ_A over the interval $[\frac{2}{3}, \frac{4}{5}]$; because $W_{AL}(\gamma, \pi_{BCT}) > W_{AL}(\hat{\sigma}_{AL}^{TSCT}, \pi_{TSCT})$ at $\delta_A = \frac{4}{5}$, then $W_{AL}(\gamma, \pi_{BCT}) > W_{AL}(\hat{\sigma}_{AL}^{TSCT}, \pi_{TSCT})$ for $\delta_A \geq \frac{2}{3}$. \square

\square

Step 2. *If $\delta_A < \frac{2}{3} \leq \frac{2}{3}$, the optimal test setting is (MT, γ, π_{BCT})*

Proof. If $\delta_A < \frac{2}{3} \leq \frac{2}{3}$, Propositions 8, 9 and 10 imply that there is only one optimal setting for each test signal: (MT, γ, π_{BCT}) , (UT, β, π_{GCT}) and (MT, γ, π_{TSCT}) . By a reasoning similar to the one in Step 1.2, the principal prefers $(MT, \gamma, \pi_{BCT}) \succ (MT, \gamma, \pi_{TSCT})$: both postpone testing by one period, both are able to successfully pass strong candidates, but the bad-certifying test is better at failing weak candidates. The principal prefers (MT, γ, π_{BCT}) over (UT, β, π_{GCT}) if and only if

$$W_{MT}(\gamma, \pi_{BCT}) = \frac{\delta_P}{1 - \delta_P} \left[\frac{3}{4} - \frac{1}{8} \right] \geq \frac{1 + \delta_P}{4(1 - \delta_P)} = W_{UT}(\beta, \pi_{GCT})$$

which is equivalent to $\delta_P \geq \frac{2}{3}$. \square

\square

Step 3. *If $\delta_A < \frac{2}{3}$ and $\delta_P < \frac{2}{3}$, the optimal test setting is involves a good-certifying test.*

Proof. I split Step 3 into three small steps. First, I illustrate that any test involving a bad-certifying test is suboptimal. Step 3.2 shows that a two-sided certifying test setting using either mandatory training or unrestricted testing is never optimal. Finally, I prove that any optimal setting involving attempt limit must use a good-certifying test signal. \square

\square

Step 3.1. *If $\delta_A < \frac{2}{3}$ and $\delta_P < \frac{2}{3}$, the optimal test setting never involves a bad-certifying test.*

Proof. Proposition 8 states that any optimal bad-certifying test uses either (MT, γ) or (UT, γ) if and only if $\delta_A < \frac{2}{3}$. Step 2 has just shown that (MT, γ, π_{BCT}) is optimal if and only if $\delta_P \geq \frac{2}{3}$. Thus, it's

not optimal. Finally, test setting (UT, β, π_{BCT}) always dominates (UT, γ, π_{BCT}) :

$$W_{UT}(\gamma, \pi_{BCT}) = \frac{1}{2} \frac{1}{1 - \delta_P} - \frac{1}{4} \frac{1}{1 - \delta_P} + \frac{\delta_P}{1 - \delta_P} \frac{1}{16} = \frac{4 + \delta_P}{16(1 - \delta_P)} < \frac{1 + \delta_P}{4(1 - \delta_P)} = W_{UT}(\beta, \pi_{GCT})$$

Thus, a bad-certifying test is never used if $\delta_A < \frac{2}{3}$ and $\delta_P < \frac{2}{3}$. \square

Step 3.2. *If $\delta_A < \frac{2}{3}$ and $\delta_P < \frac{2}{3}$, neither (MT, γ, π_{TSCT}) nor $(UT, \hat{\sigma}_{UT}^{TSCT}, \pi_{TSCT})$ is an optimal test setting.*

Proof. Test setting (MT, γ, π_{TSCT}) is always worse than (MT, γ, π_{BCT}) , which is dominated in its turn by (UT, β, π_{GCT}) for $\delta_P < \frac{2}{3}$. Test setting $(UT, \hat{\sigma}_{UT}^{TSCT}, \pi_{TSCT})$ is always dominated by (UT, β, π_{GCT}) :

$$W_{UT}(\hat{\sigma}_{UT}^{TSCT}, \pi_{TSCT}) = \frac{16 + 19\delta_P}{96(1 - \delta_P)} < \frac{1 + \delta_P}{4(1 - \delta_P)} = W_{UT}(\beta, \pi_{GCT})$$

Thus, any optimal setting never combines π_{TSCT} with either MT or UT . \square

Step 3.3. *If $\delta_A < \frac{2}{3}$ and $\delta_P < \frac{2}{3}$, the optimal test setting is never $(AL, \hat{\sigma}_{AL}^{TSCT}, \pi_{TSCT})$.*

Proof. To begin, notice that Step 1.2 has already shown that $W_{UT}(\hat{\sigma}_{UT}^{TSCT}, \pi_{TSCT})$ is increasing in δ_A . Secondly, note that $\frac{\delta_A}{2 - \delta_A} < \min\{\frac{8}{11} \frac{1}{1 - \delta_A}, \frac{2 - \delta_A}{1 - \delta_A} \frac{1}{6}\}$. Therefore, I start considering the set of values of δ_P and δ_A such that $\delta_P \geq \frac{\delta_A}{2 - \delta_A}$. Since $W_{AL}(\hat{\sigma}_{AL}^{TSCT}, \pi_{TSCT})$ is increasing in δ_A , it has the highest value at $\delta_A = \frac{2\delta_P}{1 + \delta_P}$ for every $\delta_P \leq \frac{1}{2}$ and $\delta_A = \frac{2}{3}$ for every $\delta_P \in [\frac{1}{2}, \frac{2}{3}]$. For $\delta_P \in [\frac{1}{2}, \frac{2}{3}]$ and $\delta_A = \frac{2}{3}$

$$\begin{aligned} W_{UT}(\beta, \pi_{GCT}) - W_{AL}(\hat{\sigma}_{AL}^{TSCT}, \pi_{TSCT}) &= \frac{1 + \delta_P}{4(1 - \delta_P)} - \frac{1}{1 - \delta_P} \left[\frac{2}{3}(1 - \hat{\sigma}_{AL}^{TSCT}) + \frac{\delta_P}{12} \right] \\ &= \frac{1 + \delta_P}{4(1 - \delta_P)} - \frac{1}{1 - \delta_P} \left[\frac{2}{3} \left(1 - \frac{1}{2} \right) + \frac{\delta_P}{12} \right] \\ &= \frac{1 + \delta_P}{4(1 - \delta_P)} - \frac{4 + \delta_P}{12(1 - \delta_P)} > 0 \end{aligned}$$

For $\delta_P < \frac{1}{2}$ and $\delta_A = \frac{2\delta_P}{1+\delta_P}$,

$$\begin{aligned}
W_{UT}(\beta, \pi_{GCT}) - W_{AL}(\hat{\sigma}_{AL}^{TSCT}, \pi_{TSCT}) &= \frac{1 + \delta_P}{4(1 - \delta_P)} - \frac{1}{1 - \delta_P} \left[\frac{2}{3}(1 - \hat{\sigma}_{AL}^{TSCT}) + \frac{\delta_P}{12} \right] \\
&= \frac{1 + \delta_P}{4(1 - \delta_P)} - \frac{1}{12(1 - \delta_P)} \left(\frac{2 - \delta_A}{1 - \delta_A} + \delta_P \right) \\
&= \frac{1 + \delta_P}{4(1 - \delta_P)} - \frac{1}{12(1 - \delta_P)} \left(\frac{2}{1 - \delta_P} + \delta_P \right) \\
&= \frac{1 + \delta_P}{4(1 - \delta_P)} - \frac{2 + \delta_P - \delta_P^2}{12(1 - \delta_P)} \\
&= \frac{1 - \delta_P - 2\delta_P^2}{12(1 - \delta_P)} > 0
\end{aligned}$$

Now I consider the set of values δ_P and δ_A , for which $\delta_P < \frac{\delta_A}{2 - \delta_A}$. For this set of values, the principal finds it optimal to use the attempt limit for both types of test signal. For every $\delta_P < \frac{1}{2}$ and $\delta_A = \frac{2\delta_P}{1 + \delta_P}$, $W_{AL}(\hat{\sigma}_{AL}^{GCT}, \pi_{TSCT}) = W_{UT}(\beta, \pi_{GCT}) > W_{AL}(\hat{\sigma}_{AL}^{TSCT}, \pi_{TSCT})$. However, the difference between the two decreases in δ_A :

$$\begin{aligned}
\frac{\partial \left(W_{AL}(\hat{\sigma}_{AL}^{GCT}, \pi_{TSCT}) - W_{AL}(\hat{\sigma}_{AL}^{TSCT}, \pi_{TSCT}) \right)}{\partial \delta_A} &= \frac{\partial W_{AL}(\hat{\sigma}_{AL}^{GCT}, \pi_{TSCT})}{\partial \delta_A} - \frac{\partial W_{AL}(\hat{\sigma}_{AL}^{TSCT}, \pi_{TSCT})}{\partial \delta_A} \\
&= \frac{2 - \delta_P}{2(1 - \delta_P)(4 - 3\delta_A)^2} - \frac{1}{12(1 - \delta_P)(1 - \delta_A)^2} < 0
\end{aligned}$$

Thus, if the difference $W_{AL}(\hat{\sigma}_{AL}^{GCT}, \pi_{TSCT}) - W_{AL}(\hat{\sigma}_{AL}^{TSCT}, \pi_{TSCT})$ is still positive at $\delta_A = \frac{2}{3}$ for any $\delta_P < \frac{1}{2}$, it means that the principal always prefers π_{GCT} to π_{TSCT} . For $\delta_P < \frac{1}{2}$ and $\delta_A = \frac{2}{3}$,

$$W_{AL}(\hat{\sigma}_{AL}^{GCT}, \pi_{TSCT}) - W_{AL}(\hat{\sigma}_{AL}^{TSCT}, \pi_{TSCT}) = \frac{4 + \delta_P}{12(1 - \delta_P)} - \frac{4 + \delta_P}{12(1 - \delta_P)} = 0,$$

which means that the two-sided-certifying test is never optimal. Because all the test signals but π_{GCT} are suboptimal for $\delta_A < \frac{2}{3}$ and $\delta_P < \frac{2}{3}$, I can conclude that the good-certifying test is the optimal test signal for this set of values. \square