Optimal Education Design*

Gea M. Lee† Seung Han Yoo‡

July 2026

Abstract

This paper studies the optimal design of education. We model education as an endogenous institution through which a social planner designs incentive-compatible matching paths. These paths jointly determine workers' educational allocations and the information transmitted to firms, while inducing human capital investment. The resulting problem exhibits a dual non-monotonicity arising from the feasibility constraints and the social welfare function. Their interaction generates a single-peaked reduced-form problem and a virtual cost of human capital that incorporates both marginal and inframarginal education costs. We show that the optimal mechanism is partially informative and strictly dominates both full separation and pooling, providing a new mechanism-design rationale for overeducation.

Keywords and Phrases: job market, grand mechanism, information design, mechanism design

JEL Classification Numbers: C72, D44, D82

---

*First draft: September 2025. We thank audiences at various seminars and conferences for their helpful comments. The usual disclaimer applies. Yoo acknowledges that this research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. RS-2023-00243709).

†School of Economics, Singapore Management University, (e-mail: `gmlee@smu.edu.sg`).

‡Department of Economics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, Republic of Korea, 02841 (e-mail: `shyoo@korea.ac.kr`).

1 Introduction

Education systems vary considerably across countries, leading to markedly different educational outcomes. For instance, in 2023, the attainment of tertiary education among 25–34-year-olds varied substantially across countries: 69.7% in South Korea, 65.5% in Japan, 51.8% in the United States, and 44.2% in the European Union.[1] These education systems and choices, in turn, have profound implications for overall welfare in each country – these differences often result in disproportionate monetary burdens as well as opportunity costs at both the social and individual levels.[2]

For decades, education has been modeled through two canonical lenses: the signaling view of Spence (1973) and the human capital approach of Schultz (1961) and Becker (1962). Whether viewed as a costly signal or as an investment technology, education has largely been treated as a fixed institution rather than an object of design. Then, the question arises as to why different societies exhibit such persistent differences in educational patterns.[3] Some countries might become trapped in a "bad equilibrium" among multiple equilibria under a given system. A more fundamental question, however, is how the education system itself should be designed.

This paper proposes a novel framework that treats the education system as an *endogenous object*. The labor market has two sides – informed and uninformed – and a planner. On one side, individuals privately know their investment costs and decide whether to acquire human capital; on the other, firms price workers competitively. Individuals reach the market only by going through the education system, which specifies the education levels and the probabilities with which individuals are matched to them. These tendencies in how individuals are matched across education levels are institutionalized in the system. Upon observing each individual's education level, firms update their beliefs about the individual's human capital accordingly. The education system thus serves as the *institution* through which the two sides interact.

In the model, the planner designs this institution as a menu of contracts, one for each cost type. Each contract specifies a *matching path* – a mapping that determines the probabilities with which an individual is assigned to different education levels, together with the

[1]The OECD defines tertiary education as post-secondary education corresponding to levels 5 to 8 of the International Standard Classification of Education (ISCED), ranging from short-cycle tertiary programs to doctoral or equivalent qualifications.

[2]For example, extensive evidence shows that the social costs of private tutoring and shadow education are prevalent in both developed and developing countries. See, e.g., Stevenson and Baker (1992) for early evidence from Japan; Bray (1999, 2024) for seminal and recent policy analyses; and Zhang and Bray (2020) for a comparative review. See also Dang and Rogers (2008), Kim and Lee (2010), Kim et al. (2025), among others.

[3]Whether these differences originate in historical circumstances, deliberate policy, or unintended policy failures, they systematically shape individuals' educational choices and welfare.

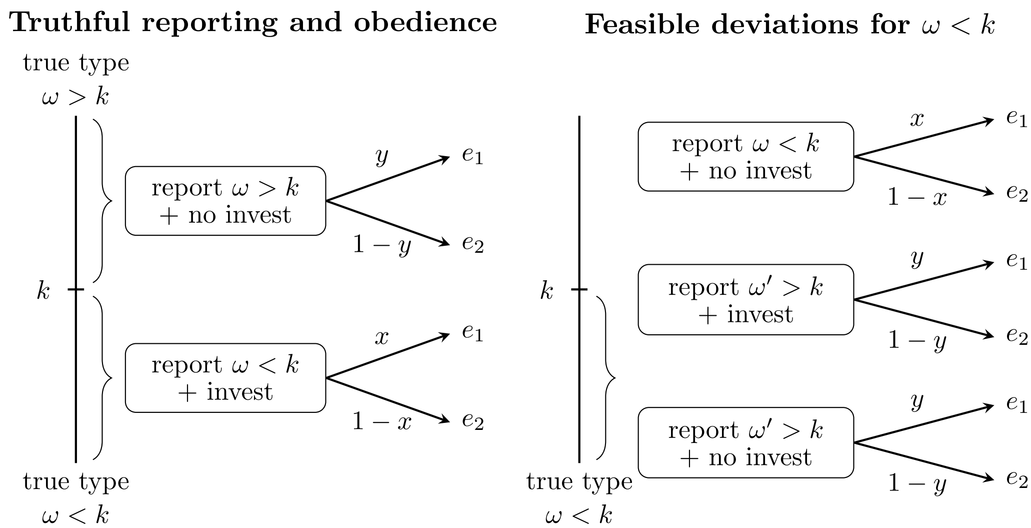

Figure 1: Equilibrium paths and feasible deviations.

education levels themselves – and a designated investment decision.[4] In this institutional design problem, an individual's strategic choice is thus not a probability distribution over messages. Rather, the planner determines the set of feasible matching paths available in the institution, and each individual simply selects one of these available paths while also choosing whether to invest. Because neither the cost type nor the investment is observable, the planner must design the menu so that individuals sort themselves: Each type willingly takes the contract designated for him and makes the designated investment. These two requirements – self-selection and obedience – are the incentive compatibility conditions of our design problem.

Beyond screening, this institutional design also governs the flow of information to firms. The beliefs firms form about workers' human capital depend on the matching paths chosen by the planner, and these beliefs in turn determine workers' expected returns to investment and hence their incentives to acquire human capital. The planner therefore designs the *entire education system*: the incentives to acquire human capital, educational assignment, and the transmission of information to firms. The distribution of human capital – and hence firms' prior – is itself endogenous, and the planner trades off the benefits of human capital formation and information transmission against the costs of investment and education in order to maximize social welfare.

Under this joint design of incentives and information, it is not immediate that the presence of only two quality levels is sufficient to restrict attention to two education levels – higher and basic education. We show, however, that this restriction is indeed without loss of

[4] A matching path comprises a diverse set of educational systems – such as public versus private institutions, national school systems, and specialized enrichment programs – and an individual's choice of matching path is typically not observable to firms.

generality.[5] Moreover, the mechanism induces a cutoff rule for human capital investment: Individuals below a cost threshold k invest, while those above it do not. The planner then chooses the probabilities with which high- and low-quality individuals are assigned to higher education, together with the corresponding levels of higher and basic education, subject to incentive compatibility and individual rationality. Figure 1 illustrates the strategic choices of individuals described above in the direct-mechanism representation; reporting $\omega < k$ corresponds to choosing the contract with matching path $(x, 1-x)$, while reporting $\omega > k$ corresponds to choosing the contract with matching path $(y, 1-y)$. The left panel depicts truthful reporting and obedience, whereas the right panel illustrates the feasible report–investment deviations for an agent with $\omega < k$.[6]

One implication of designing the grand mechanism is that the informativeness of education is itself endogenous. For a fixed y, changing x does not merely redistribute a given population of high-quality workers across education levels; by changing investment incentives, it also changes the cutoff $k(x)$ and hence the size of that population.[7] Under the uniform distribution, for example, the masses of high-quality workers assigned to higher and basic education are $xk(x)$ and $(1-x)k(x)$, respectively. More importantly, partial informativeness is not simply the result of "manipulating" firms' beliefs; rather, it reflects the planner's joint optimization of human capital acquisition, educational assignment, and information transmission, with incentive compatibility linking these dimensions and the planner balancing the benefits of human capital and information against the cost of investment and education. This result gives rise to a *virtual cost of human capital*, related to but distinct from the virtual cost in standard mechanism design.[8]

The first main result provides a reduction of the planner's problem. Under a separable cost structure following Spence (1973), our single-crossing property implies that the incentive compatibility constraint for low-quality individuals binds at the optimum. As a result, low-quality individuals always receive basic education at no cost, while higher education is fully informative of high quality: $y^* = 0$ and $e_2^* = 0$. The planner's problem therefore reduces to choosing x and e_1, with the informativeness of the education system governed entirely by x.[9]

[5] We allow a signal structure of an education mechanism to adopt any number of education levels, including a continuous case.

[6] The case $\omega > k$ is symmetric. For an agent with $\omega < k$, there are three possible deviations: he may report truthfully but choose not to invest; he may misreport $\omega' > k$ while investing; or he may combine both deviations by misreporting $\omega' > k$ and choosing not to invest.

[7] Although the principal directly chooses x and e_1, the constraints induce a one-to-one relationship between x and k, as shown formally in Online Appendix C.

[8] Our virtual cost of human capital investment bears an intriguing relationship to the well-known virtual cost in standard mechanism design.

[9] Throughout the paper, we use "reduction" to refer to the step that reduces the planner's choice variables from (x, y, e_1, e_2) to (x, e_1), and "reformulation" to refer to the subsequent step that rewrites the reduced problem as a one-dimensional problem in the investment threshold k.

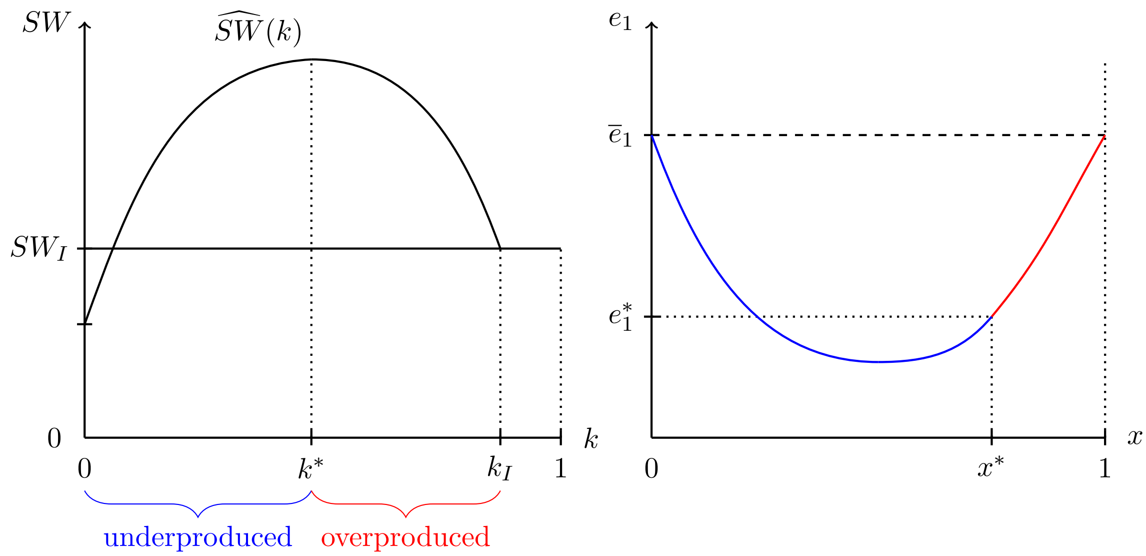

Figure 2: The social welfare and the relationship between x and e_1

Building on this reduction, the next result establishes existence and uniqueness. Despite this simplification, the interaction between incentive compatibility and human capital accumulation remains nontrivial. Nevertheless, we show that the problem can be reformulated as if the planner chose only the threshold k, even though he directly chooses (x, e_1).[10] This reformulation yields existence and uniqueness conditions characterized by the virtual human capital cost.

The final main result characterizes the optimal education design. A well-known "secret" of the signaling literature is that, from an ex ante social welfare perspective, the pooling equilibrium dominates the separating equilibrium despite the informational benefits of signaling. Our framework overturns this conclusion by identifying a welfare-maximizing *partially informative* education system. The fully uninformative mechanism assigns heterogeneous individuals to a uniform education level, thereby ignoring type-specific cost differences and eliminating the wage differential that supports human capital investment.

The dominance over the fully informative mechanism is driven by a *dual non-monotonicity* – one in the feasible set and one in the social welfare function – a feature absent from the existing literature. On the feasibility side, at both endpoints, basic education is fully informative of low quality, but for opposite reasons. When $x = 1$, all high-quality individuals are assigned to higher education – like a separating equilibrium – leaving basic education held exclusively by low-quality individuals. Interestingly, at the other extreme, we have the

[10] As shown in Section 4, under the human capital condition, x and e_1 are substitutive, since both contribute to human capital accumulation. Under the binding incentive condition, however, the relationship becomes complementary, since an increase in x requires a higher e_1 to deter low-quality individuals from mimicking. Together, these forces generate a one-to-one correspondence between the educational design (x, e_1) and the induced investment threshold k.

same informativeness. A key relationship we establish is that, given $y^* = 0$, the endogenous investment cutoff k is a strictly increasing function of x. Hence, as x approaches its lowest feasible value, so does the investment cutoff k. Consequently, almost no individuals invest, so the population is almost entirely low quality. Basic education is then again highly informative of low quality, but this time because high-quality individuals are exceedingly scarce. Hence, in both extreme cases, the expected wage differential approaches its maximum, and the binding incentive compatibility condition for low-quality individuals pushes e_1 toward its maximum level $\overline{e}_1$ to deter deviation, generating the non-monotone relationship between x and e_1, as depicted in the right panel of Figure 2.[11]

On the welfare side, the effect of increasing x is also non-monotonic: Social welfare first increases and then decreases with x. Although the expected wage differential is maximized at both extremes of x – through the exclusion and the scarcity – social welfare is not: It depends not only on the expected wage differential but also on the *fraction* of high-quality individuals who receive it. When x is small, the expected wage differential is high, but only a small fraction of high-quality individuals is assigned to higher education, so expanding higher education generates a large aggregate human-capital gain. As x increases, the investment cutoff k rises – because of the endogenous relationship between k and x – and more individuals undertake human capital investment. At the same time, however, an increasing fraction of high-quality individuals is already assigned to higher education, so the gain from further expansion gradually diminishes, while aggregate information costs continue to rise.

Putting the feasibility and welfare sides together yields the main characterization. For relatively low values of x, an increase in x is accompanied by a decrease in e_1 along the feasible path, so both changes raise welfare and k is under-produced (the blue region). For relatively high values of x, however, an increase in x requires a higher e_1, creating a conflict between feasibility and welfare, so k is over-produced (the red region). This yields the reformulated welfare function that is single-peaked in the investment threshold k, implying an interior optimum $k^* \in (0, k_I)$, as described in the left panel of Figure 2, where k_I is the fully informative mechanism's threshold. This interpretation also provides a *rationale for overeducation* that is unique to our framework. In the over-produced region of the reformulation, increasing k near k_I requires increases in both x and e_1. While a higher x assigns additional high-quality individuals to higher education, a higher e_1 is not applied only to these marginal entrants; it also raises the education level of all inframarginal individuals already assigned to higher education. Hence, further increases in k operate on both the *marginal and inframarginal* populations, leading to socially excessive educational investment. This force is encapsulated by our virtual cost function, giving rise to what we term the

[11] For intermediate values of x, the expected wage differential falls, so e_1 falls below $\overline{e}_1$. Thus, as x rises, the feasible path starts at $\overline{e}_1$, declines, and eventually returns to $\overline{e}_1$. The right panel of Figure 2 is constructed from the human capital condition and the binding incentive condition under a uniform distribution of cost types on $[0, 1]$; the corresponding equations and the more general non-monotonicity result are provided in Section 4. In general, the constraint locus need not be convex-shaped, but it is characterized by an initial decline followed by an eventual increase.

information-cost trap. At the fully informative threshold k_I, the marginal human capital benefit is exhausted, while this inframarginal information cost remains strictly positive.

We also derive comparative statics with respect to the wage differential, the return to human capital investment, and the distribution of cost types. Finally, we relax the separability assumption. While separability provides a convenient sufficient condition for reducing the planner's problem, it is not essential.[12]

Our results carry several policy implications for the design of education systems. First, we offer a new, non-mismatch-based perspective on overeducation, attributing it to the systematic over-expansion of higher-education access rather than to labor market frictions. Second, building on the empirical literature on the determinants of individuals' intrinsic investment costs, we derive implications for education policy. Finally, although the model's key objects are not directly observable, we derive *tractable testable implications* of the optimal education system.

This paper studies an institutional design problem in which the principal jointly designs incentives and information transmission. The interaction between two distinct incentive compatibility requirements – truthful reporting and obedient behavior – has been widely studied since Myerson (1982). A key departure of our framework is that the principal controls the information transmitted to the uninformed side; information design is *integrated into* mechanism design itself.

This feature connects the paper to the information design literature following Kamenica and Gentzkow (2011) and Bergemann and Morris (2016), but our framework differs in its basic design object. The planner does not choose an experiment over a fixed state; instead, he designs an incentive-compatible education mechanism whose matching paths jointly determine investment incentives, educational assignments, and the information observed by firms. The paper is also related to incentive-compatible information design, where incentive compatibility restricts the feasible set of signal structures available to the principal. This issue was first formally analyzed by Aoyagi and Yoo (2022) in the context of a matching platform.

Our approach is thus also related to the literature on grand mechanisms that combine allocation and information disclosure. In contrast to dynamic screening models such as Eső and Szentes (2007) and Li and Shi (2017), where disclosure is privately tailored to individual agents to change their valuations, the information design component of this grand mechanism features *public* signals: The public disclosure aims to change beliefs of *the other side* valuing workers. Under population uncertainty, Kim and Yoo (2022) investigate a grand mechanism problem with public signals.

The environment follows the education-signaling framework of Spence (1973), but education is treated as an object of institutional design. This perspective connects the paper to two strands that study how educational rules affect either the information conveyed to

[12] The separable specification follows the canonical signaling framework of Spence (1973) and is broadly consistent with the dynamic complementarity in skill formation emphasized by Cunha and Heckman (2007).

the market or individuals' incentives to exert effort. One strand studies education as an information-transmission mechanism. In Ostrovsky and Schwarz (2010), given the distributions of workers and jobs, each school chooses an information policy that shapes the distribution of its students' posterior means. In equilibrium, coarse grading can improve a school's expected placement by changing firms' posterior beliefs and thereby improving the assignment of its students to jobs. A second strand emphasizes how educational informativeness affects effort choices. Dubey and Geanakoplos (2010) show that fully revealing grades can weaken effort incentives when students care only about relative rankings. More recently, Krishna et al. (2026) show that pooling test-performance intervals can raise all students' interim payoffs by weakening costly competition for rank within the pool. Our framework combines these two dimensions. Firms' posterior beliefs determine the return to human capital investment, so the planner jointly designs investment incentives and information transmission through an incentive-compatible education mechanism. Moreover, because individuals' investment decisions determine the distribution of human capital, the population about which information is transmitted is itself endogenous. Partial informativeness therefore emerges as a feature of the education system as an institution, rather than either dimension in isolation.

This paper is organized as follows. Section 2 presents the model, and Section 3 provides a preliminary analysis. The main results are developed in Section 4, with comparative statics presented in Section 5. Section 6 discusses policy implications, and Section 7 concludes. Proofs of the results in Sections 3–5 are collected in the Appendix. The Online Appendix contains the full mechanism formulation, the proofs for Section 2, and an extension beyond separable cost functions with supporting results and proofs.

2 Model

2.1 Baseline model

The economy has two sides – individuals and firms – and a social planner. On the informed side, each individual privately observes his investment cost $\omega \in \Omega \equiv [\underline{\omega}, \overline{\omega}]$, with $\overline{\omega} > \underline{\omega} \geq 0$, and decides whether to acquire human capital. On the uninformed side, firms observe only education and hire workers in a competitive market. Individuals enter the labor market only through the education system, which specifies both the education levels and the probabilities with which individuals are assigned to them. We refer to each such probability distribution over education levels as a *matching path*. Thus, the education system is the institution through which the two sides interact in the model.

The education system operates in two stages. First, each individual chooses a human capital level $\theta \in \Theta \equiv \{\theta_H, \theta_L\}$, where θ_H and θ_L denote high and low human capital, respectively. Second, each individual chooses a matching path from the set of feasible matching paths specified by the education system, after which the individual is probabilistically assigned to an education level according to the chosen path. Upon observing the realized education level e, firms update their beliefs about the individual's human capital via Bayes' rule and determine the competitive wage accordingly.

The human capital investment outcome is observable only to the individual, and thus is not contractible. Yet, by observing e, each firm updates beliefs μ about the probability that the agent is $\theta = \theta_H$ via Bayes rule. We consider a market environment in which the wage is determined competitively with the expected wage $\mathbb{E}^{\mu}[W \mid e] = \mu w_H + (1-\mu) w_L$, where w_H and w_L are the competitive market wages for high-human-capital and low-human-capital individuals, respectively, satisfying $w_H > w_L \geq 0$.

If an individual with type ω chooses a human capital level $\theta \in \Theta \equiv \{\theta_H, \theta_L\}$, his payoff from the expected wage with a matched education level e is given by

$$u(\mu, e, \theta, \omega) = \mathbb{E}^{\mu}[W \mid e] - c(e, \theta) - \omega \cdot \mathbf{1}_{\{\theta=\theta_H\}}, \tag{1}$$

where $c(e, \theta)$ denotes the cost of education e for θ. To emphasize the role of the beliefs of firms – we refer to firms, which share a common belief, as the receiver – we express u as a function of μ.[13] We assume that the type ω is drawn from a cumulative distribution G with density $g(\omega) > 0$ for all $\omega \in (\underline{\omega}, \overline{\omega})$, where the distribution is common knowledge, and that the cost function $c(e, \theta)$ satisfies $c(0, \theta) = 0$ for all θ and $c_e(e, \theta) > 0$ for all $e > 0$. The social planner aims to maximize social welfare, the aggregate payoff of all individuals.

2.2 Education design

By the revelation principle for environments with hidden actions (Myerson, 1982), attention can be restricted without loss to mechanisms in which the agent reports his type and then receives a recommendation on the investment, with incentive compatibility imposed jointly on report and obedience. The principal *recommends* a human capital investment to each agent, after which an agent of type ω chooses $\theta \in \Theta$; following the investment, the principal probabilistically matches the agent with an observable education level $e \geq 0$. Unlike signaling, it is not the agent who chooses e but the principal who matches a type-ω agent with e. As such, the education design requires neither equilibrium refinements nor off-the-path beliefs for education levels e, in contrast to signaling equilibria.

Formally, a direct mechanism Γ consists of functions ρ and π with

$$\rho : \Omega \to \Delta(\Theta) \quad \text{and} \quad \pi : \Omega \to \Delta(\mathbb{R}_+), \tag{2}$$

where $\Delta(X)$ denotes the set of probability distributions over X. The recommendation rule ρ maps each report to the human capital recommendation, and the assignment rule π maps each report to a matching path. Then, for each agent's report ω, the principal commits to

[13]Alternatively, an individual's ex post payoff can be written as $u(\mathrm{w}, e, \theta, \omega) = \mathrm{w} - c(e, \theta) - \omega \cdot \mathbf{1}_{\{\theta=\theta_H\}}$ with an ex post wage w, but we instead incorporate the expected wage to (1) to highlight μ in light of the reduced-form competitive market wage.

recommending a human capital level θ with probability $\rho(\theta|\omega)$ and assigning the matching path $\pi(e|\omega)$. The commitment in this model means *institution building*: Once a society's education system is established, the social planner cannot "readily" modify it; any modification requires a different committed mechanism.

Upon hearing a recommendation on the investment, each agent can follow it or not. We denote each agent's strategy given a recommendation and its probabilistic distribution, respectively, by

$$q : \Theta \to \Delta(\Theta) \text{ and } \mathbf{q}(\theta) \equiv \big(q(\theta_H|\theta), q(\theta_L|\theta)\big). \tag{3}$$

Then, if an agent reports ω and chooses a strategy q, the probability that the agent is of high quality given a recommendation θ is $\mathbb{E}^{\rho}\big[\mathbf{q}(\theta)\cdot(1,0) \mid \omega\big] = \rho(\theta_H|\omega)[\mathbf{q}(\theta_H)\cdot(1,0)] + \rho(\theta_L|\omega)[\mathbf{q}(\theta_L)\cdot(1,0)]$, resulting in $\mathbb{E}^{\rho}\big[\mathbf{q}(\theta)\cdot(1,0) \mid \omega\big] = \rho(\theta_H|\omega)q(\theta_H|\theta_H) + \rho(\theta_L|\omega)q(\theta_H|\theta_L)$, where $\mathbf{q}(\theta)\cdot(1,0)$ is an inner product between the two vectors. Similarly, the probability that the agent is low quality is $\mathbb{E}^{\rho}\big[\mathbf{q}(\theta)\cdot(0,1) \mid \omega\big]$.

Then, by observing e, the receiver's beliefs μ_e are updated via Bayes rule such that

$$\mu_e(\mathbf{q}) = \frac{\int_\Omega \pi(e|\omega)\mathbb{E}^{\rho}\big[\mathbf{q}(\theta)\cdot(1,0) \mid \omega\big]dG(\omega)}{\int_\Omega \pi(e|\omega)\mathbb{E}^{\rho}\big[\mathbf{q}(\theta)\cdot(1,0) \mid \omega\big]dG(\omega) + \int_\Omega \pi(e|\omega)\mathbb{E}^{\rho}\big[\mathbf{q}(\theta)\cdot(0,1) \mid \omega\big]dG(\omega)}. \tag{4}$$

Note that in this model, each agent both reports a type ω and chooses an investment strategy q, so the incentive compatibility condition consists of two components, *truthful reporting* and *obedience*. The principal designs the mechanism so that agents have no incentive to misreport their type or disobey the recommended investment. That is, if truthful reporting holds, an agent of type ω has no incentive to deviate from the contract $(\rho(\cdot|\omega), \pi(\cdot|\omega))$ assigned to his type to any alternative contract $(\rho(\cdot|\omega'), \pi(\cdot|\omega'))$ for $\omega' \neq \omega$. If obedience holds, each agent follows the principal's recommendation, which is denoted by $\mathbf{q}^*(\theta_H) = (1,0)$ and $\mathbf{q}^*(\theta_L) = (0,1)$. Hence, complications can arise with respect to *combinatorial deviations*: Each agent can disobey by choosing any human capital level after misreporting his type. The detailed discussions are to be provided subsequently. Myerson (1982) first formally explored the two distinct notions of incentive compatibility. The present paper departs from Myerson (1982) in that the principal not only designs a mechanism for agents but also controls the flow of information from agents to the receiver. Unlike standard information design frameworks following Kamenica and Gentzkow (2011) and Bergemann and Morris (2016), the information structure in the present model is not exogenously given. Instead, the mechanism itself determines the *underlying information structure* through recommendations on investment decisions and incentive-compatible reporting.

By (1) and (4), if an agent reports ω and chooses a strategy q, the agent's ex-ante expected wage is given by

$$\mathbb{E}^{\pi}\big[\mathbb{E}^{\mu_e(\mathbf{q})}[W \mid E] \mid \omega\big] = \int_0^\infty \mathbb{E}^{\mu_e(\mathbf{q})}[W \mid e]\pi(e|\omega)de,$$

where W denotes the realized wage and E the realized education level. Given the continuous distribution G, we can still use the same $\mu_e(\mathbf{q})$ in (4), even when an agent *deviates* from truthful reporting. Our notion of incentive compatibility, as in standard mechanism design, requires that no agent has a profitable deviation given truthful reporting by all other agents.

In addition, for a mechanism Γ, each agent can incur costs from two different sources: education cost and human capital investment cost. First, the agent's ex-ante education cost is

$$\mathbb{E}^{\pi}\big[\mathbb{E}^{\rho}\big[\mathbf{q}(\theta)\cdot\mathbf{c}(E)\mid\omega\big]\mid\omega\big]=\int_{0}^{\infty}\big[\rho(\theta_H|\omega)\big(\mathbf{q}(\theta_H)\cdot\mathbf{c}(e)\big)+\rho(\theta_L|\omega)\big(\mathbf{q}(\theta_L)\cdot\mathbf{c}(e)\big)\big]\,\pi(e|\omega)de,$$

where we denote $\mathbf{c}(e)\equiv\big(c(e,\theta_H),c(e,\theta_L)\big)$, and second, the agent's investment cost is

$$\mathbb{E}^{\rho}\big[\mathbf{q}(\theta)\cdot(1,0)\mid\omega\big]\omega=\big[\rho(\theta_H|\omega)\big(\mathbf{q}(\theta_H)\cdot(1,0)\big)+\rho(\theta_L|\omega)\big(\mathbf{q}(\theta_L)\cdot(1,0)\big)\big]\,\omega.$$

Then, the incentive compatibility condition requires that each agent with ω report ω truthfully and obey by choosing $\mathbf{q}^*(\theta_H)=(1,0)$ and $\mathbf{q}^*(\theta_L)=(0,1)$. The formal incentive compatibility and individual rationality conditions are provided in Online Appendix A. We now derive a simpler characterization.

2.3 Incentive compatibility and individual rationality

The following proposition simplifies this otherwise complex problem.

Proposition 1 *Any incentive compatible mechanism Γ satisfies the following properties.*

(i) There exists $k\in[\underline{\omega},\overline{\omega}]$ such that

$$\rho(\theta_H|\omega)=\begin{cases}1 & \text{if } \omega<k,\\ 0 & \text{if } \omega>k.\end{cases}$$

(ii) For any pair $\omega\neq\omega'\in[\underline{\omega},k)$ or any pair $\omega\neq\omega'\in(k,\overline{\omega}]$, $\pi(e|\omega)=\pi(e|\omega')$, and further, for any signal structure $\pi:\Omega\to\Delta(\mathbb{R}_+)$, there exists a payoff-equivalent simple signal structure with two education levels.

The first result shows that the recommendation must be given as a cut-off function with a threshold k. The second establishes the within-region invariance of $\pi(e|\omega)$ within $\omega<k$ and $\omega>k$, and further shows that the maximum number of "necessary" education levels is two, denoted by e_1 and e_2. Note that the proof of the cardinality result in Proposition 1 must incorporate the incentive compatibility *constraints*.

Then, for $i=1,2$, the belief function in (4) becomes

$$\mu_i\equiv\Pr(\omega<k\mid e_i)=\frac{\int_{\underline{\omega}}^{k}\pi(e_i|\omega)dG(\omega)}{\int_{\underline{\omega}}^{k}\pi(e_i|\omega)dG(\omega)+\int_{k}^{\overline{\omega}}\pi(e_i|\omega)dG(\omega)},\tag{5}$$

which yields $\mathbb{E}^{\mu}[W \mid e_i] = \mu_i w_H + (1-\mu_i) w_L$.[14] Note that k is a threshold for human capital investment, which is *endogenously* determined through the incentive compatibility condition derived below.

If an agent reports type ω truthfully and obeys the recommendation θ, by Proposition 1, the payoff in (1) yields his expected payoff

$$U\big(\theta, \omega; \pi, (e_i)_{i=1,2}\big) \equiv \pi(e_1|\omega)\{\mathbb{E}^{\mu}[W|e_1] - c(e_1,\theta)\} + \pi(e_2|\omega)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2,\theta)\} - \mathbf{1}_{\{\theta=\theta_H\}}\omega, \tag{6}$$

where the obedient quality decision is given by $\theta = \mathbf{1}_{\{\omega<k\}}\theta_H + (1-\mathbf{1}_{\{\omega<k\}})\theta_L$. Then, a mechanism Γ is said to be incentive compatible (IC) if for each $\omega, \omega' \in \Omega$,

$$\begin{aligned} &U\big(\theta, \omega; \pi, (e_i)_{i=1,2}\big) \\ &\geq \pi(e_1|\omega')\{\mathbb{E}^{\mu}[W|e_1] - c(e_1,\theta')\} + \pi(e_2|\omega')\{\mathbb{E}^{\mu}[W|e_2] - c(e_2,\theta')\} - \mathbf{1}_{\{\theta'=\theta_H\}}\omega \text{ for all } \theta' \in \Theta. \end{aligned} \tag{7}$$

In the model, it is the principal, not an agent, who chooses π and $(e_i)_{i=1,2}$; an agent reports ω' and chooses θ'. As explained earlier, our incentive compatibility condition (IC) must account for deviations along two dimensions, one with respect to reporting a type ω' given his true type ω, and another with respect to choosing a human capital θ' upon a recommendation θ, which is given as θ_H for $\omega' < k$ and θ_L for $\omega' > k$. Thus, there are three possible ways an agent can deviate: only misreporting his type; only disobeying the recommendation; or both (a combinatorial deviation).

We replace IC by the following two constraints to simplify (7). We call the first one the truthful (incentive compatibility) condition for each $\theta \in \Theta$ in the sense that the agent still chooses his "designated quality" choice, while he may misreport a type ω'. The second one is called the human capital (investment) condition. Formally, a mechanism Γ is said to satisfy the truthful condition (TC) if for each $\omega, \omega' \in \Omega$ and every $\theta \in \Theta$,

$$\begin{aligned} &\pi(e_1|\omega)\{\mathbb{E}^{\mu}[W|e_1] - c(e_1,\theta)\} + \pi(e_2|\omega)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2,\theta)\} \\ &\geq \pi(e_1|\omega')\{\mathbb{E}^{\mu}[W|e_1] - c(e_1,\theta)\} + \pi(e_2|\omega')\{\mathbb{E}^{\mu}[W|e_2] - c(e_2,\theta)\}. \end{aligned} \tag{8}$$

One can find that the right-hand side is in fact the payoff when an agent deviates combinatorially.[15] Further, a mechanism Γ is said to satisfy the human capital condition (HC) if for each $\omega < k$ and every $\omega' > k$,

$$\begin{aligned} &\pi(e_1|\omega)\{\mathbb{E}^{\mu}[W|e_1] - c(e_1,\theta_H)\} + \pi(e_2|\omega)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2,\theta_H)\} - k \\ &= \pi(e_1|\omega')\{\mathbb{E}^{\mu}[W|e_1] - c(e_1,\theta_L)\} + \pi(e_2|\omega')\{\mathbb{E}^{\mu}[W|e_2] - c(e_2,\theta_L)\}, \end{aligned} \tag{9}$$

---

[14] By Proposition 1, $\pi(e_i|\omega)$ is constant within each of the regions $\omega < k$ and $\omega > k$, so the corresponding terms can be taken outside the integrals in (5). We retain the integral representation here because it is convenient for stating the incentive compatibility and individual rationality conditions below. It is simplified further in (12).

[15] To understand it, suppose that $\omega < k$. Then, the expected payoff in (6) becomes $\pi(e_1|\omega)\{\mathbb{E}^{\mu}[W|e_1] - c(e_1,\theta)\} + \pi(e_2|\omega)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2,\theta)\} - \omega$, and the right-hand side is from the payoff $\pi(e_1|\omega')\{\mathbb{E}^{\mu}[W|e_1] - c(e_1,\theta)\} + \pi(e_2|\omega')\{\mathbb{E}^{\mu}[W|e_2] - c(e_2,\theta)\} - \omega$, when the agent disobeys by choosing $\theta' = \theta_H$ against the recommendation θ_L, following a misreport $\omega' > k$.

where k is the HC threshold type.[16]

We show that the incentive compatibility condition in (7) can be replaced by the two conditions above TC in (8) and HC in (9) for $k \in [\underline{\omega}, \overline{\omega}]$.

Lemma 1 Γ *is incentive compatible (IC) if and only if TC and HC are satisfied.*

In addition, a mechanism Γ is said to be individually rational (IR) if for each $\omega \in \Omega$ and $\theta = \mathbf{1}_{\{\omega<k\}}\theta_H + (1 - \mathbf{1}_{\{\omega<k\}})\theta_L$,

$$U\big(\theta, \omega; \pi, (e_i)_{i=1,2}\big) \geq 0. \tag{10}$$

Next, we define the principal's objective function.

2.4 Social welfare and education cost

In light of Proposition 1, to simplify notations further, we denote

$$x \equiv \pi(e_1|\omega) \text{ for } \omega < k \text{ and } y \equiv \pi(e_1|\omega) \text{ for } \omega > k, \tag{11}$$

with which for each e_1 and e_2, the updated beliefs in (5) become

$$\mu_1 = \frac{x\lambda}{x\lambda + y(1-\lambda)} \text{ and } \mu_2 = \frac{(1-x)\lambda}{(1-x)\lambda + (1-y)(1-\lambda)} \text{ where } \lambda = G(k). \tag{12}$$

Now, we derive the social welfare as

$$\begin{aligned} SW = & \int_{\underline{\omega}}^{k} \left\{ x\mathbb{E}^{\mu}[W|e_1] + (1-x)\mathbb{E}^{\mu}[W|e_2] - \big[xc(e_1,\theta_H) + (1-x)c(e_2,\theta_H)\big] - \omega \right\} dG \\ & + \int_{k}^{\overline{\omega}} \left\{ y\mathbb{E}^{\mu}[W|e_1] + (1-y)\mathbb{E}^{\mu}[W|e_2] - \big[yc(e_1,\theta_L) + (1-y)c(e_2,\theta_L)\big] \right\} dG. \end{aligned} \tag{13}$$

Then, a mechanism Γ is optimal if it maximizes SW in (13) subject to TC in (8), HC in (9) and IR in (10). For the constrained optimization problem, we have four variables x, y, e_1, e_2 with the three families of constraints.[17]

In particular, given HC in (9), the belief μ in $\mathbb{E}^{\mu}[W|e_i]$ depends on λ through the updated beliefs $(\mu_i)_{i=1,2}$ in (12). Hence, k, or equivalently $\lambda = G(k)$, is an implicit function of (x, y, e_1, e_2). As a result, a change in any one of the variables (x, y, e_1, e_2) affects not only social welfare SW through its effect on k, but also the remaining constraints through the

[16]The right-hand side is in fact the payoff when an agent deviates only with respect to a type report. Suppose that an agent's type satisfies $\omega < k$. Then, the right-hand side is the agent's payoff when he follows the recommendation θ_L after misreporting $\omega' > k$.

[17]Although k and λ appear in the objective function and the constraints, they are determined by (x, y, e_1, e_2) through HC in (9). Hence, the planner's problem can be written as a choice of the four variables (x, y, e_1, e_2). Online Appendix C provides a more formal treatment by explicitly retaining λ.

induced change in beliefs, which renders the original formulation difficult to analyze directly. The complexity is to be illustrated in the subsequent section. Importantly, in this framework, without additional structure on the cost function, the incentive compatibility condition for the low human capital agent is not necessarily binding. In the signaling model by Spence (1973), the intuitive criterion of Cho and Kreps (1987) guarantees the binding incentive compatibility condition for the low type, a property that is standard in the signaling approach.[18] Hence, in addition to the standard properties of $c(e,\theta)$ in the baseline model, we impose the separability assumption for the main analysis as our way of ensuring the binding condition.

Assumption 1 *The cost function is separable such that $c(e,\theta) = \beta(\theta)\gamma(e)$ for all $e \geq 0$ with $\beta(\theta_L) > \beta(\theta_H) > 0$.*

The separability is in fact equivalent to a simple property, the proportional relationship between $c(e,\theta_L)$ and $c(e,\theta_H)$ such that for all $e \geq 0$,

$$c(e,\theta) = \beta(\theta)\gamma(e) \Leftrightarrow c(e,\theta_L) = \alpha c(e,\theta_H) \text{ given some } \alpha > 1.^{19} \tag{14}$$

When investment generates a cost reduction, the associated rate of *return* is given by $r(e) \equiv \frac{c(e,\theta_H)-c(e,\theta_L)}{c(e,\theta_L)}$, and under separability, $r(e)$ becomes constant with $r = \frac{1-\alpha}{\alpha}$, yielding a one-to-one relationship $\alpha = \frac{1}{1-r}$. Hence, we can call $\alpha = \frac{\beta(\theta_L)}{\beta(\theta_H)}$ *the cost-reduction multiplier* of the investment return in this model, and for example, the cost-reduction multiplier of $c(e,\theta) = \frac{e^2}{\theta}$ is $\alpha = \frac{\theta_H}{\theta_L}$.

Although separability is only a sufficient condition for our analysis, it is a natural benchmark, having been the standard specification in the signaling literature since the original model of Spence (1973). Even without separability, the low-quality agent's incentive condition may still bind for a wide class of cost functions. In Online Appendix C, we show that whenever the binding condition is satisfied, our main qualitative results continue to hold. To characterize this broader class of environments, we define $Q(e)$ for any $e \geq 0$ with $c(e,\theta_L) > c(e,\theta_H)$ such that

$$Q(e) \equiv \frac{c(e,\theta_L)}{c(e,\theta_L) - c(e,\theta_H)}. \tag{15}$$

Since $r(e) < 0$, we have $Q(e) = 1/|r(e)|$. Thus, $Q(e)$ is the inverse of the absolute rate of cost reduction: It measures the low-quality agent's investment cost relative to the cost saving associated with high quality. Under separability, $Q(e)$ becomes constant, equal to $\frac{\alpha}{\alpha-1}$, just as $r(e)$ is constant.

In addition, to allow for an interior cutoff under the fully informative signal, we assume the following range for Δw; otherwise, agents in the population are either all θ_H or all θ_L.

[18] In contrast, off-the-path beliefs play no role in our education design problem; instead, separability delivers a qualitatively similar binding result. A comparable binding condition also arises under the screening approach of Rothschild and Stiglitz (1976), in which firms make wage offers first.

[19] Formally, for each $\theta \neq \theta'$ and $\alpha \neq 0$, $c(e,\theta') = \alpha c(e,\theta)$ for all $e \geq 0$ if and only if $c(e,\theta)$ is separable in the sense that for each θ, there exist corresponding functions $\beta(\theta)$ and $\gamma(e)$ such that $c(e,\theta) = \beta(\theta)\gamma(e)$ for all $e \geq 0$.

Assumption 2 $\Delta w \in \left(\frac{\alpha}{\alpha-1}\underline{\omega}, \frac{\alpha}{\alpha-1}\overline{\omega}\right)$.

The fully informative signal matches an agent reporting $\omega < k$ (or $\omega > k$) with some e deterministically so that the signal perfectly reveals quality θ. Suppose that the principal assigns e_1 for $\omega < k$, and $e_2 = 0$ for $\omega > k$. If it is optimal for the *best* type $\underline{\omega}$ to make the human capital investment, and similarly, it is optimal for the *worst* type $\overline{\omega}$ not to make the investment, the former (resp. the latter) requires $w_H - c(e_1, \theta_H) - \underline{\omega} > w_L$ (resp. $w_H - c(e_1, \theta_H) - \overline{\omega} < w_L$). The stated range then follows because separability implies, as shown in the next section, that TC for low-quality agents binds as $w_L = w_H - c(e_1, \theta_L)$, where TC is given by (8).[20] We can alternatively interpret the condition as requiring the absolute value of the return to human capital, $|r| = \frac{\alpha-1}{\alpha}$, multiplied by the wage differential Δw, to be greater than the cost of the best type $\underline{\omega}$ but lower than that of the worst type $\overline{\omega}$.

We next simplify the social welfare in (13), which also enables us to relate our model to the standard information design.

Lemma 2 *The social welfare is derived as*

$$\begin{aligned} SW = {} & \lambda\{w_H - [xc(e_1, \theta_H) + (1-x)c(e_2, \theta_H)]\} \\ & + (1-\lambda)\{w_L - [yc(e_1, \theta_L) + (1-y)c(e_2, \theta_L)]\} - \int_{\underline{\omega}}^{k} \omega dG. \end{aligned} \tag{16}$$

The social welfare above can be rewritten as

$$SW = \tau^{\lambda}(e_1)V(e_1, \mu_1) + \tau^{\lambda}(e_2)V(e_2, \mu_2) - \int_{\underline{\omega}}^{k} \omega dG, \tag{17}$$

where $\tau^{\lambda}(e_i) \equiv \pi(e_i|\theta_H)\lambda + \pi(e_i|\theta_L)(1-\lambda)$ denotes the ex ante probability that signal e_i is generated, and the corresponding value function can be defined as

$$V(e, \mu) \equiv \mu[w_H - c(e, \theta_H)] + (1-\mu)[w_L - c(e, \theta_L)]. \tag{18}$$

Now, we decompose the role of e into two components: information transmission and cost. That is, if, "hypothetically," the costly aspect can be disentangled from the informational aspect, one can easily identify that for a fixed e, $V(e, \mu)$ is linear in information transmission via μ such that $V(e, \lambda) = \tau^{\lambda}(e_1)V(e, \mu_1) + \tau^{\lambda}(e_2)V(e, \mu_2)$, implying that if education serves purely as a signal – with no education costs – the principal is indifferent across all signal structures.

The welfare representation in (17) of (16) clarifies the relationship between our approach and standard information-design frameworks and highlights two important departures from them. First, the principal must *endogenously* collect information from agents through incentive-compatible reporting, where the value function $V(e, \mu)$ depends not only

[20] Outside this range, no fully separating outcome with an interior cutoff is feasible, since either all agents become high quality or all remain low quality.

on beliefs μ but also on the endogenous information cost e. Second, the underlying prior $\lambda = G(k)$ in (12) is itself *endogenous* rather than exogenously given, since it is determined through the mechanism's investment recommendations. Hence, our framework incorporates both endogenous information costs and an endogenous prior.

3 Preliminary analysis

In this section, we consider two extreme and yet important signal structures: the fully informative signal $x = 1, y = 0$ and the fully uninformative signal $x = y$. We first study these two polar cases and then present a simple example that strikingly illustrates how a partially informative mechanism can strictly dominate the fully uninformative signal. To do so, we rewrite TC, HC, and IR in (8)–(10); this new formulation will also be used in the full analysis in the next section.

Before analyzing them, we provide a benchmark case, in which there is no human capital investment stage; that is, we fix the prior λ. With no human capital investment, the setting only with TC and IR for a fixed λ makes it equivalent to the standard signaling problem but with the market design angle in Section 2. That is, the principal still controls education e, while each agent can choose a matching path.

In the market design model with a fixed λ, given (17), the principal solves

$$\max_{x,y,e_1,e_2} \tau^\lambda(e_1)V(e_1, \mu_1) + \tau^\lambda(e_2)V(e_2, \mu_2) \text{ subject to } TC, IR.$$

As noted above, $V(e, \mu)$ is linear in μ for a fixed e, so the dominance of the fully uninformative signal over the fully informative signal is immediate:

$$V(0, \lambda) = \tau^\lambda(e_1)V(0, \mu_1) + \tau^\lambda(0)V(0, \mu_2) > \tau^\lambda(e_1)V(e_1, \mu_1) + \tau^\lambda(0)V(0, \mu_2).$$

Without education costs, the linearity implies the "irrelevance" of the signal. Under the fully informative allocation, however, incentive compatibility requires the high-quality agent to choose a strictly positive education level, $e_1 > 0$. It is precisely this education cost that generates the strict inequality above, as summarized below.

Proposition 2 *Suppose there is no human capital investment stage. Then, the fully uninformative signal dominates the fully informative signal.*

With no human-capital investment, the dominance of pooling highlights a stark contrast between real-world signaling practices and the market-design benchmark. What Proposition 2 shows is that banning a signal is optimal from the social planner's point of view, in theory.

We therefore introduce the direct mechanism Γ in (2). Through the joint choice of ρ and π, the planner determines both the allocation of human-capital investment and the information conveyed by education. In particular, the induced share λ of high-quality individuals becomes a function of (x, y, e_1, e_2), as do the posterior beliefs μ_1 and μ_2. The resulting interaction between allocation and information is the main technical difficulty of the model. For the analysis, the difference in the expected wage between e_1 and e_2, $\mathbb{E}^\mu[W|e_1] - \mathbb{E}^\mu[W|e_2]$, plays a key role, which can be rewritten as $\mathbb{E}^\mu[W|e_1] - \mathbb{E}^\mu[W|e_2] = (\mu_1 - \mu_2)\Delta w$, where $\Delta w \equiv w_H - w_L$. We denote the difference in the updated beliefs $\mu_1 - \mu_2$ by

$$D(\mu) \equiv \mu_1 - \mu_2, \tag{19}$$

where μ_1 and μ_2 are defined in (12).

Before simplifying TC, HC and IR, it can be shown that, without loss of generality, we focus on $x > y$ together with $e_1 > e_2$ in what follows.

Lemma 3 *For any incentive compatible mechanism Γ, $x > y$ if and only if $e_1 > e_2$ or $x < y$ if and only if $e_1 < e_2$.*

Then, with $x > y$ by the lemma above and $D(\mu)$ in (19), we rewrite TC in (8) as

$$\begin{aligned} D(\mu)\Delta w - c(e_1, \theta_H) + c(e_2, \theta_H) &\geq 0 \text{ (TC}_H), \\ D(\mu)\Delta w - c(e_1, \theta_L) + c(e_2, \theta_L) &\leq 0 \text{ (TC}_L). \end{aligned} \tag{20}$$

When $x > y$, reporting $\omega < k$ increases the probability of receiving e_1 by exactly $x - y$. Since this probability enters both sides of TC proportionally, the common factor $x - y$ cancels out, yielding (20). Similarly, the human capital condition (9) can be written as

$$k = (x-y)D(\mu)\Delta w - [xc(e_1, \theta_H) - yc(e_1, \theta_L)] - [(1-x)c(e_2, \theta_H) - (1-y)c(e_2, \theta_L)] \text{ (HC)}. \tag{21}$$

As discussed earlier, μ depends on k or λ in HC, so $D(\mu)$ itself is a function of x, y, e_1, e_2. In addition, IR in (10) is given as

$$\begin{aligned} x\left[D(\mu)\Delta w - c(e_1, \theta_H) + c(e_2, \theta_H)\right] + \mathbb{E}^\mu[W|e_2] - c(e_2, \theta_H) &\geq 0, \text{ (IR}_H), \\ y\left[D(\mu)\Delta w - c(e_1, \theta_L) + c(e_2, \theta_L)\right] + \mathbb{E}^\mu[W|e_2] - c(e_2, \theta_L) &\geq 0. \text{ (IR}_L). \end{aligned} \tag{22}$$

An education mechanism is optimal if the principal chooses the four variables x, y, e_1, e_2 to maximize the objective function SW in (13) subject to the constraints in (20)–(22).

To examine the role of endogeneity, consider the first-order condition of SW in (16) with respect to e_1:

$$\begin{aligned} \frac{\partial SW}{\partial e_1} = &\underbrace{\frac{\partial \lambda}{\partial e_1}\left[\Delta w - xc(e_1, \theta_H) + yc(e_1, \theta_L) - (1-x)c(e_2, \theta_H) + (1-y)c(e_2, \theta_L)\right]}_{\text{human capital effect}} \\ &- \underbrace{\left[\lambda x c_e(e_1, \theta_H) + (1-\lambda) y c_e(e_1, \theta_L)\right]}_{\text{information cost effect}} - \underbrace{\frac{dk}{de_1} k g(k)}_{\text{human capital cost effect}} \\ = &\underbrace{\frac{\partial \lambda}{\partial e_1}\Delta w[1 - (x-y)D(\mu)]}_{\text{net human capital effect}} - [\lambda x c_e(e_1, \theta_H) + (1-\lambda) y c_e(e_1, \theta_L)], \end{aligned} \tag{23}$$

where the second equality follows from HC in (21). Given the endogeneity, a change in e_1 affects not only the direct costs and benefits associated with education, but also the measure of individuals who undertake the investment. This is captured by $\partial\lambda/\partial e_1$, which, by the implicit function theorem applied to (21), is given by

$$\frac{\partial \lambda}{\partial e_1} = -\frac{-[xc_e(e_1,\theta_H) - yc_e(e_1,\theta_L)]}{(x-y)D_\lambda(\mu)\Delta w - 1}.$$

Furthermore, the net human capital effect in the second equality combines the human capital effect and the human capital cost effect in the first equality through the endogenous k (or equivalently λ).[21] The formula therefore highlights the key trade-off of the model between the net human capital effect and the information cost effect.

While (23) highlights the complexity created by the endogeneity of k, we show in Section 4 that separability yields a single-crossing relationship among the objective function and the constraints. This leads to a binding incentive constraint for low-quality agents – the reduced problem. To illustrate the case that separability is violated, consider the following running example.

Example 1 (Non-separable costs)

$$c(e,\theta_H) = e + \frac{e^2}{2} \text{ and } c(e,\theta_L) = te + e^2 \text{ for } t > 1. \tag{24}$$

One can verify that separability holds only when $t = 2$. For $t \neq 2$, in contrast to the constant cost-reduction multiplier α implied by separability in (14), define

$$\alpha(e) \equiv \frac{c_e(e,\theta_L)}{c_e(e,\theta_H)}, \tag{25}$$

which now depends on e. For the example, $\alpha(e) = \frac{t+2e}{1+e}$ is strictly increasing in e for $t \in (1,2)$ and strictly decreasing for $t > 2$. Then, substituting $\frac{\partial\lambda}{\partial e_1}$ and factoring out $c_e(e_1,\theta_H)$, we can rewrite (23) as

$$\frac{\partial SW}{\partial e_1} = c_e(e_1,\theta_H)\left\{\left(\frac{x - y\alpha(e_1)}{(x-y)D_\lambda(\mu)\Delta w - 1}\right)\Delta w[1-(x-y)D(\mu)] - \lambda x - (1-\lambda)y\alpha(e_1)\right\}. \tag{26}$$

Then, unlike under separability, the first term in (26) can no longer be expressed in terms of a constant marginal cost ratio α, which complicates the characterization of its sign, let alone the indeterminacy of the denominator $(x-y)D_\lambda(\mu)\Delta w - 1$. This provides a glimpse of the difficulties that arise once separability is violated. Note, however, that even when

[21] The second equality follows from two substitutions. First, since $\lambda = G(k)$, $\frac{\partial\lambda}{\partial e_1} = g(k)\frac{dk}{de_1}$. Second, substituting HC, $k = x[c(e_1,\theta_L) - c(e_1,\theta_H)] + (1-x)[c(e_2,\theta_L) - c(e_2,\theta_H)]$, into the bracketed term yields $\Delta w - (x-y)\big[c(e_1,\theta_L) - c(e_2,\theta_L)\big]$. Using the binding TC_L, $D(\mu)\Delta w = c(e_1,\theta_L) - c(e_2,\theta_L)$ then gives $\Delta w[1-(x-y)D(\mu)]$.

separability is violated, we can still have the reduced problem, so long as the relevant single-crossing condition holds. For instance, in (24), when $t \neq 2$, separability is violated, yet the binding incentive constraint may still hold. We revisit this example in Section 4.

Now, we analyze the two polar cases: the fully informative signal and the fully uninformative signal. First, the fully informative case is a special case of the partially informative case with $1 \geq x > y \geq 0$ in Section 4. Hence, the results for the general analysis also apply to this benchmark, where Proposition 3 and Lemma 4 show that TC_L is binding, resulting in $e_2^* = 0$ and $y^* = 0$ for any solution. Next, given the binding TC_L, one can find that $x = 1, y = 0$ result in $\mathbb{E}^{\mu}[W|e_1] = w_H$ and $\mathbb{E}^{\mu}[W|e_2] = w_L$, together with $D(\mu) = 1$. We denote by k_I the fully informative signal's threshold and $\overline{e}_1$ its higher education level. Then, by incorporating $x = 1, y = 0$ and $e_2 = 0$ to the binding TC_L in (20) and HC in (21), we derive $\overline{e}_1$ and k_I, respectively, such that

$$\Delta w = c(\overline{e}_1, \theta_L) \text{ and } k_I = \Delta w - c(\overline{e}_1, \theta_H). \tag{27}$$

With IR in (22) being always satisfied, the social welfare in (16) becomes

$$SW_I = \lambda_I \left[w_H - c(\overline{e}_1, \theta_H)\right] + (1-\lambda_I) w_L - \int_{\underline{\omega}}^{k_I} \omega dG = w_L + \int_{\underline{\omega}}^{k_I} G(\omega) d\omega, \tag{28}$$

where the second equality follows from the integration by parts and $k_I = \Delta w - c(\overline{e}_1, \theta_H)$. By combining the two outcomes in (14) and (27), the fully informative HC threshold becomes

$$k_I = c(\overline{e}_1, \theta_L) - c(\overline{e}_1, \theta_H) = \frac{\alpha - 1}{\alpha} \Delta w \text{ and } \overline{e}_1 = \sqrt{\theta_L \Delta w}. \tag{29}$$

The range $\Delta w \in \left(\frac{\alpha}{\alpha-1}\underline{\omega}, \frac{\alpha}{\alpha-1}\overline{\omega}\right)$ in Assumption 2 guarantees the interior case. While the closed-form expression in (29) relies on separability, the reformulation itself does not. Online Appendix C shows that the same procedure extends beyond separable costs once the binding TC_L is obtained.[22]

The other polar case is the fully uninformative signal, characterized by $x = y$ with $e_1 = e_2 = e$, which results in $\mathbb{E}^{\mu}[W|e] = \mathbb{E}^{\lambda}[W] \equiv \lambda_U w_H + (1-\lambda_U) w_L$.[23] We denote by k_U the fully uninformative signal's threshold. In this case, the only relevant constraints are HC and IR_L in (21) & (22), and by incorporating $x = y$ to them, we derive the fully uninformative signal's condition for e and k_U such that

$$\mathbb{E}^{\lambda_U}[W] - c(e, \theta_L) \geq 0 \text{ and } k_U = c(e, \theta_L) - c(e, \theta_H). \tag{30}$$

[22] The identity $k_I = c(e_1, \theta_L) - c(e_1, \theta_H)$ holds generally without separability. Under separability, however, $k_I = \frac{c(e_1,\theta_L) - c(e_1,\theta_H)}{c(e_1,\theta_L)} c(e_1, \theta_L) = \frac{\alpha-1}{\alpha} \Delta w$, where the second equality uses $c(e_1, \theta_L) = \alpha c(e_1, \theta_H)$ and the binding TC_L, which implies $c(e_1, \theta_L) = \Delta w$.

[23] Because all education levels induce the same posterior, any such signal is payoff-equivalent to a deterministic signal with $e_1 = e_2 = e$.

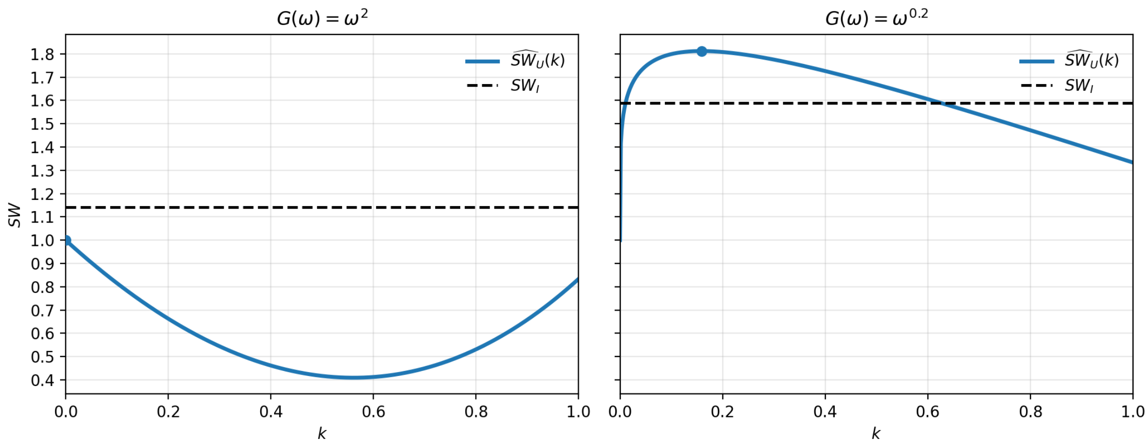

Figure 3: Comparison between fully informative and uninformative case

The social welfare in (16) becomes $SW_U = \mathbb{E}^{\lambda_U}[W] - \lambda_U[c(e,\theta_H) - c(e,\theta_L)] - c(e,\theta_L) - \int_{\underline{\omega}}^{k_U} \omega dG$. Then, the principal solves

$$\max_{e \geq 0} SW_U(e) \text{ subject to } \mathbb{E}^{\lambda_U}[W] - c(e,\theta_L) \geq 0 \text{ and } k_U = \frac{\alpha-1}{\alpha} c(e,\theta_L), \tag{31}$$

where $k_U = c(e,\theta_L) - c(e,\theta_H) = \frac{\alpha-1}{\alpha} c(e,\theta_L)$ by the separability.

Since k_U is strictly increasing in e via $k_U = c(e,\theta_L) - c(e,\theta_H)$, maximizing the social welfare with respect to e can be equivalently reformulated as maximizing the same objective now with respect to k_U. The transition results in the social welfare

$$\widehat{SW}_U(k_U) \equiv \mathbb{E}^{\lambda_U}[W] + \left[\lambda_U - \frac{\alpha}{\alpha - 1}\right] k_U - \int_{\underline{\omega}}^{k_U} \omega dG, \tag{32}$$

where we apply $k_U = c(e,\theta_L) - c(e,\theta_H)$ together with the property in (14).[24] This simple case reduces the number of variables to one, k_U, making the main trade-off in (23) particularly transparent. The derivative of the social welfare with respect to k_U yields

$$\frac{dSW_U(k_U)}{dk_U} = \underbrace{\frac{\partial \lambda}{\partial k_U} \frac{\partial \mathbb{E}^{\lambda_U}[W]}{\partial \lambda}}_{\text{human capital effect}} + \underbrace{\left[\lambda_U - \frac{\alpha}{\alpha-1}\right] + k_U g(k_U)}_{\text{information cost effect}} - \underbrace{k_U g(k_U)}_{\text{human capital cost effect}}. \tag{33}$$

The reformulation transforms the maximization over e into a maximization over k_U:

$$\max_{k_U \in [\underline{\omega}, \overline{\omega}]} \widehat{SW}_U(k_U) \text{ subject to } \mathbb{E}^{\lambda_U}[W] - c(e,\theta_L) \geq 0 \text{ where}$$

$$k_U = \begin{cases} \underline{\omega} & \text{if } c(e,\theta_L) < \frac{\alpha}{\alpha-1}\underline{\omega}, \\ \frac{\alpha-1}{\alpha} c(e,\theta_L) & \text{if } c(e,\theta_L) \in \left[\frac{\alpha}{\alpha-1}\underline{\omega}, \frac{\alpha}{\alpha-1}\overline{\omega}\right], \\ \overline{\omega} & \text{if } c(e,\theta_L) > \frac{\alpha}{\alpha-1}\overline{\omega}. \end{cases} \tag{34}$$

[24] To see this, first write $SW_U = \mathbb{E}^{\lambda_U}[W] - \lambda_U[c(e,\theta_H) - c(e,\theta_L)] - c(e,\theta_L) - \int_{\underline{\omega}}^{k_U} \omega dG$ and $c(e,\theta_L) = k_U \frac{c(e,\theta_L)}{k_U} = k_U \frac{c(e,\theta_L)}{(\alpha-1)c(e,\theta_H)} = k_U \frac{\alpha}{\alpha-1}$.

One can also easily verify that the two problems are equivalent through $\frac{\partial SW_U}{\partial e} = \frac{\partial SW_U}{\partial k_U}\frac{\partial k_U}{\partial e}$.

The characterization of the fully uninformative case shows that the shape of the objective function SW_U does not necessarily admit an interior solution k_U. Differentiating SW_U twice with respect to k_U yields $g'(k_U)\Delta w + g(k_U)$, which may be positive or negative for some k_U, Δw and $G(\omega)$. To illustrate it, let us consider another example.

Example 2 (Simple separable costs)

$$c(e,\theta) = \frac{e^2}{\theta}. \tag{35}$$

With Example 2 and $\theta_H = 2, \theta_L = 1, \Delta w = 1.5$, consider two different distributions $G(\omega) = \omega^2$ and $G(\omega) = \omega^{0.2}$, both with support $\omega \in [0,1]$. The latter yields a strictly concave function $\widehat{SW}_U(k_U)$, whereas the former yields a strictly convex one, as illustrated in Figure 3. Furthermore, the dominance of two polar cases cannot be determined in general; the fully uninformative signal dominates the fully informative signal for $G(\omega) = \omega^{0.2}$, whereas the reverse holds for $G(\omega) = \omega^2$.

Finally, to illustrate the role of partial informativeness, consider Example 2 and $\theta_H = 2, \theta_L = 1, \Delta w = 1.5$, now with the uniform distribution $G(\omega) = \omega$ on $[0,1]$. Then, by incorporating the fully informative solution in (29) into (28), we have $SW_I = w_L + \frac{1}{2}\left(\frac{\alpha-1}{\alpha}\Delta w\right)^2$, where $\alpha = \frac{\theta_H}{\theta_L} = 2$. On the other hand, the fully uninformative signal in (34) yields

$$\widehat{SW}_U(k_U) = \mathbb{E}^{\lambda_U}[W] + \lambda_U k_U - \frac{\alpha}{\alpha-1}k_U - \int_{\underline{\omega}}^{k_U} \omega dG \tag{36}$$

which is strictly convex. Hence, its optimum is attained at one of the boundaries, $\widehat{SW}_U(0) = w_L$ and $\widehat{SW}_U(1) = w_H - \frac{\alpha}{\alpha-1} + \frac{1}{2}$. In contrast to the fully uninformative welfare function, which may be either concave or convex depending on G as depicted above, the reduced welfare under partial informativeness is single-peaked under a regularity condition.[25] For the same uniform specification,

$$\widehat{SW}(k) = \mathbb{E}^{\lambda}[W] + \lambda k - \frac{\alpha}{\alpha-1}\lambda k - \int_{\underline{\omega}}^{k} \omega dG, \tag{37}$$

which becomes $w_L + \frac{3}{2}k - \frac{3}{2}k^2$ and maximized at $k^* = \frac{1}{2}$. The formal derivation is provided in the next section. The key distinction is that partial informativeness implements a given human-capital cutoff k by assigning higher education only to the fraction $G(k) = k$ of the population, but under the fully uninformative signal, by contrast, everyone receives the same education level. Thus, the only difference between (36) and (37) is the information-cost term:

$$-\underbrace{\frac{\alpha}{\alpha-1}\lambda k}_{\text{partially informative}} \quad \text{versus} \quad -\underbrace{\frac{\alpha}{\alpha-1}k}_{\text{fully uninformative}}.$$

[25]Note that an increasing reverse hazard rate is sufficient for the reformulated welfare under partial informativeness to be single-peaked. By contrast, it does not guarantee single-peakedness under the fully uninformative mechanism. For example, both $G(\omega) = \omega$ and $G(\omega) = \omega^2$ satisfy an increasing reverse hazard rate, yet the corresponding fully uninformative welfare functions are strictly convex.

Hence, $\widehat{SW}(k) - \widehat{SW}_U(k) = \frac{\alpha}{\alpha-1}(1-\lambda)k$, so whenever the same cutoff k is feasible under both mechanisms, $\widehat{SW}(k^*) > \widehat{SW}(k) \geq \widehat{SW}_U(k)$. The welfare gain is exactly the mass of individuals who do not invest, $1-\lambda$, multiplied by the education cost that can be avoided by assigning them costless basic education. The next section characterizes the optimal partially informative mechanism in general.

4 Optimal education design

In this section, we present the main results of this paper: the existence and the characterization of the optimal education design. We begin with a benchmark in which we suppose that the principal can observe each agent's type ω. As discussed in the model, $V(e,\mu)$ is linear in information transmission via μ, so the principal is indifferent across all signal structures – *i.e.*, the fully uninformative signal achieves the *same* ex-ante payoff, provided that the information cost is fixed. Hence, with no incentive compatibility condition required, in this complete information benchmark, it is optimal for the principal to choose no education.

By choosing $e_1 = e_2 = 0$, the only allocation problem of the principal is whether he requires each agent to acquire human capital or not. Then, the social welfare in (16) becomes $\lambda w_H + (1-\lambda)w_L - \int_{\underline{\omega}}^{k} \omega dG$, and the principal solves

$$\begin{gathered}\max_{k\in[\underline{\omega},\overline{\omega}]} \lambda w_H + (1-\lambda)w_L - \int_{\underline{\omega}}^{k} \omega dG \\ \text{subject to } \mathrm{IR}_H, \mathrm{IR}_L \text{ where } \lambda = G(k).\end{gathered} \tag{38}$$

Since both IR conditions in (22) are always satisfied, the first-order condition for an interior optimum k is given by

$$g(k)\big[\Delta w - k\big] = 0, \tag{39}$$

which yields the first-best $\widehat{k}$ for the human capital accumulation as

$$\widehat{k} = \begin{cases} \Delta w & \text{if } \Delta w \leq \overline{\omega}, \\ \overline{\omega} & \text{if } \Delta w > \overline{\omega}, \end{cases} \tag{40}$$

where Assumption 2 ensures $\Delta w > \underline{\omega}$, but not the other corner case. With no information cost in (23) or (33), for the first-best case, the optimal k is determined by the balance between the human capital effect and the human capital cost effect. That is, an increase in k increases the overall human capital cost in the society, whereas it increases each agent's expected wage through $\lambda = G(k)$. The trade-off yields the optimal $\widehat{k}$ in (40).

Now, we consider the original education design problem that maximizes the social welfare in (16) subject to the IC & IR constraints in (20), (22) together with HC in (21). The principal chooses $x, y \in [0,1]$ and $e_1, e_2 \in \mathbb{R}_+$ to solve

$$\begin{gathered}\max_{x,y,e_1,e_2} SW \\ \text{subject to } \mathrm{TC}_H, \mathrm{TC}_L, \mathrm{IR}_H, \mathrm{IR}_L, \mathrm{HC}, x, y \in [0,1], e_1, e_2 \in \mathbb{R}_+.\end{gathered} \tag{41}$$

The main existence and characterization in Theorems 1 and 2 do not require a unique solution, but comparative statics analysis does. For the latter, we define the *virtual cost of human capital investment* as

$$\psi(\omega) \equiv \frac{\alpha}{\alpha-1}\omega + \frac{1}{\alpha-1}\frac{G(\omega)}{g(\omega)}. \tag{42}$$

If $\psi(\omega)$ is strictly increasing in $\omega \in [\underline{\omega}, \overline{\omega}]$, then, as shown in Theorem 1, there is a unique solution to (41).

Assumption 3 *$\psi(\omega)$ is strictly increasing in $\omega \in [\underline{\omega}, \overline{\omega}]$.*

The definition in (42) also allows us to relate the optimal solution to the complete information benchmark in (40) as well as to the standard mechanism design problem. The strict monotonicity of $\psi(k)$ holds if the well-known monotone (reverse) hazard rate condition is satisfied: $\frac{G(\omega)}{g(\omega)}$ is nondecreasing. Unlike the standard virtual cost function, however, our $\psi(\omega)$ depends on the term $\alpha/(\alpha-1)$, which reflects the cost-reduction multiplier α in (14).[26]

4.1 The reduced problem

We simplify the social welfare maximization problem in (41). The main step is to prove that TC_L is binding at any optimal solution. Despite the endogenous human capital accumulation, the separability property in Assumption 1 enables us to obtain the binding incentive constraint for low-quality agents.

Proposition 3 *Suppose Assumptions 1 and 2 hold. If an incentive compatible and individually rational mechanism Γ is optimal, then TC_L is binding.*

To motivate the relationship between the separability property and the binding TC_L, revisit Example 1 in (24), where separability fails whenever $t \neq 2$. One can show that for $t \in (1,2)$, the relevant single-crossing property may fail, whereas for $t > 2$, it continues to hold and TC_L remains binding. Thus, separability is sufficient for single crossing but not necessary. We revisit this distinction in Online Appendix C. The general geometric intuition is illustrated in Figure 4.

In Figure 4, the red line depicts the iso-welfare curve (its dimension differs from that of the iso-welfare curve in Figure 2), and the blue line depicts the iso-incentive compatibility curve for high-quality agents with θ_H, so the shaded area depicts the set of (e_1, e_2) pairs such that SW increases, while TC_H is satisfied. In the right panel, Figure 4(b), if e_1 increases along HC, the social welfare can increase, but TC_H is violated. This means that if TC_H is binding at the optimum, such a conflict between TC_H and SW prevents us from obtaining the binding TC_L. In contrast, in the left panel, Figure 4(a), if the separability property is satisfied, a decrease in e_1 always increases SW without having a conflict with TC_H. Hence,

[26] One can verify that $Q(e)$ in (15) of the generalized model reduces to $\alpha/(\alpha-1)$ under Assumption 1.

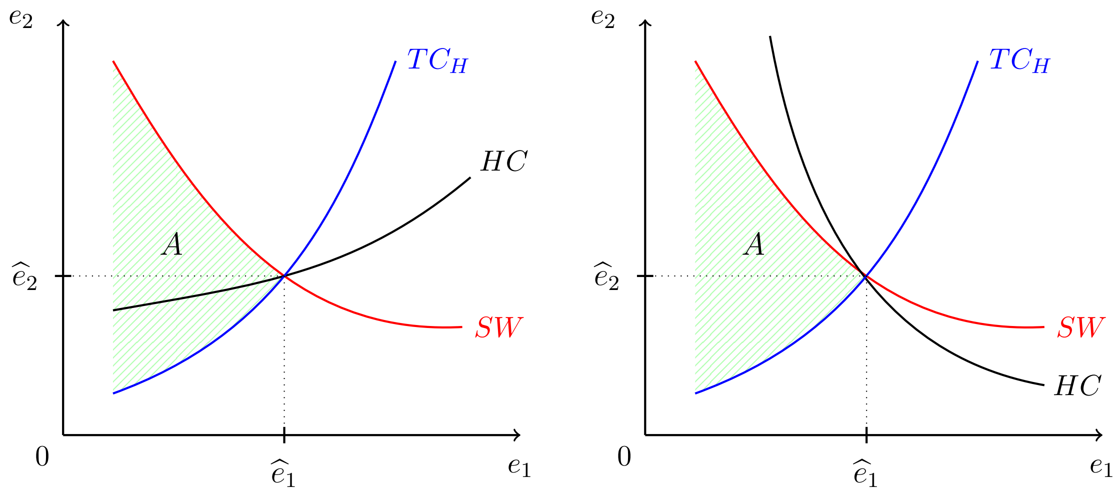

Figure 4: (a) Separability holds and (b) Separability is violated

we can have a solution only when TC_L is binding. The separability essentially guarantees the *single-crossing* property of the iso-welfare curve relative to HC and TC_H, as described in Figure 4(a). The proof also shows that IR_L satisfies the single-crossing. Once this key step is established, the following result for y^* and e_2^* can be subsequently obtained.

Lemma 4 *Suppose Assumptions 1 and 2 hold. If an incentive compatible and individually rational mechanism Γ is optimal, then $y^* = 0$, and $e_2^* = 0$.*

Similar to $\frac{\partial SW}{\partial e_1}$ in (23), $\frac{\partial SW}{\partial y}$ also consists of three effects: the human capital effect, the information cost effect, and the human capital cost effect. In general, the sign of the change in human capital accumulation λ with respect to y is ambiguous because changing y affects both posterior beliefs and the investment threshold. However, by Proposition 3, $D(\mu)\Delta w - c(e_1, \theta_L) + c(e_2, \theta_L) = 0$. Hence,

$$\frac{\partial \lambda}{\partial y} = -\frac{-D(\mu)\Delta w + (x-y)D_y(\mu)\Delta w + c(e_1, \theta_L) - c(e_2, \theta_L)}{(x-y)D_\lambda(\mu)\Delta w - 1} \tag{43}$$

simplifies to $-\frac{(x-y)D_y(\mu)\Delta w}{(x-y)D_\lambda(\mu)\Delta w - 1} < 0$, so $\frac{\partial \lambda}{\partial y} < 0$ in (43) and $\frac{\partial SW}{\partial y} < 0$ as well. We can also utilize the left panel of Figure 4 for the intuition for $e_2^* = 0$; as long as $e_2 > 0$, one can decrease e_1 to increase SW, without violating any constraint.

The corresponding updated beliefs are:

$$\mu_1 = 1 \text{ and } \mu_2 = \frac{(1-x^*)\lambda^*}{(1-x^*)\lambda^* + (1-\lambda^*)} \text{ where } \lambda^* = G(k^*). \tag{44}$$

Since $y^* = 0$, only high-quality individuals are assigned to higher education, so higher education perfectly reveals high quality, implying $\mu_1 = 1$. By contrast, basic education contains all low-quality individuals together with any high-quality individuals not assigned to higher education, depending on the optimal value of x^*.

4.2 Reformulation: Existence and uniqueness

Lemma 4 identifies two of the four optimal choice variables, $y^* = 0$ and $e_2^* = 0$, leaving only x and e_1 to be determined. The original problem therefore reduces substantially, and the social welfare in (16) becomes

$$SW(x, e_1) = \lambda\{w_H - xc(e_1, \theta_H)\} + (1-\lambda)w_L - \int_{\underline{\omega}}^{k} \omega dG.$$

Further, the binding TC_L implies that TC_H is slack, and both IR_H and IR_L are always satisfied as well.[27] That is, the binding TC_L from (20) and HC in (21) are the only remaining effective constraints. First, the binding TC_L is given by

$$D(\mu)\Delta w - c(e_1, \theta_L) = 0, \tag{45}$$

and HC in (21) is

$$k = xD(\mu)\Delta w - xc(e_1, \theta_H) = x[c(e_1, \theta_L) - c(e_1, \theta_H)] \text{ where } D(\mu) = \frac{1-\lambda}{1-x\lambda}. \tag{46}$$

Hence, the principal solves the following simple problem, instead of (41):

$$\begin{aligned} &\max_{x, e_1} SW(x, e_1) \\ &\text{subject to binding } \mathrm{TC}_L, \mathrm{HC}, x \in [0,1], e_1 \in \mathbb{R}_+. \end{aligned} \tag{47}$$

Under HC, both x and e_1 contribute to the human capital accumulation, so an increase in one requires a reduction in the other in order to maintain the same k; x and e_1 are *substitutive.* On the other hand, under the binding TC_L, as an increase in x raises the expected wage differential $D(\mu)$, e_1 must also rise to preserve the binding truth-telling incentives for individuals with θ_L, holding k constant; they are *complementary.*

This combination of the two constraints makes it possible to map each feasible value of k into the corresponding pair (x, e_1) and thereby reformulate the original problem in terms of k alone. As shown in the proof of Theorem 1, the social welfare can then be written as

$$\widehat{SW}(k) \equiv \mathbb{E}^{\lambda}[W] + \left[1 - \frac{\alpha}{\alpha - 1}\right]\lambda k - \int_{\underline{\omega}}^{k} \omega dG. \tag{48}$$

Then, an alternative formulation of (47) is given by

$$\max_{k} \widehat{SW}(k) \text{ subject to } k \in [\underline{\omega}, k_I]. \tag{49}$$

The first main result shows that the original problem with $y^* = 0$, and $e_2^* = 0$ in (47) is equivalent to (49), which enables us to pursue the simpler problem without loss of generality.

[27] By incorporating $y^* = 0$, and $e_2^* = 0$, IR_H and IR_L are $x[D(\mu)\Delta w - c(e_1, \theta_H)] + \mathbb{E}^{\mu}[W|e_2] \geq 0$, $y[D(\mu)\Delta w - c(e_1, \theta_L)] + \mathbb{E}^{\mu}[W|e_2] \geq 0$.

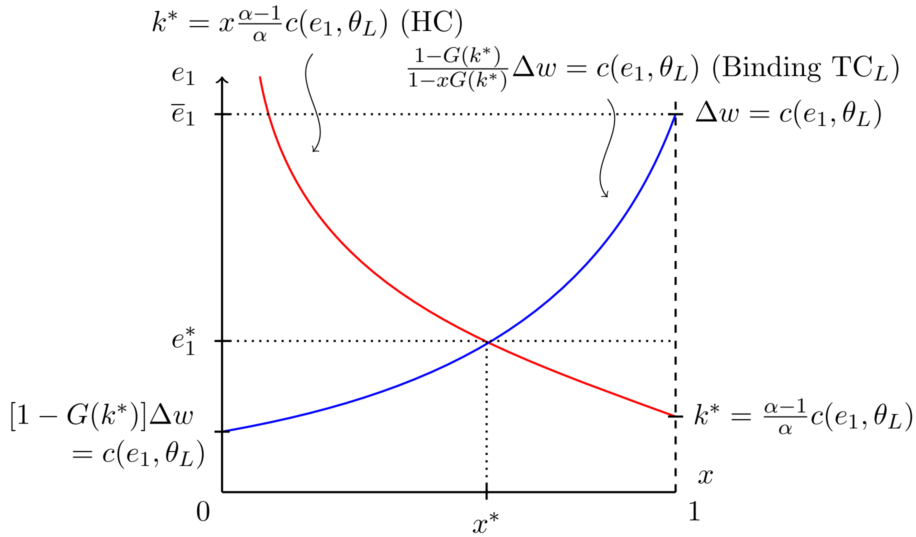

Figure 5: A system of simultaneous equations

Theorem 1 *Suppose Assumptions 1 and 2 hold. Then, the optimal education mechanism problem in (47) is equivalent to (49) such that*

$$\max_{x\in[0,1],e_1\in\mathbb{R}_+}\left\{SW(x,e_1): \textit{binding } TC_L, HC\right\} = \max_{k\in[\underline{\omega},k_I]}\widehat{SW}(k).$$

If, in addition, Assumption 3 holds, the solution to (49) is unique.

Because $\widehat{SW}(k)$ is continuous and $[\underline{\omega},k_I]$ is compact, there exists a solution k^* to the alternative problem in (49), even without the monotonicity of $\psi(k)$. Importantly, once the optimal human capital accumulation k^* is obtained, the unique corresponding solution (x^*,e_1^*) in the original problem (47) is derived by solving the following system of simultaneous equations:

$$k^* = x\frac{\alpha-1}{\alpha}c(e_1,\theta_L)\ (\mathrm{HC})\ \text{and}\ \frac{1-G(k^*)}{1-xG(k^*)}\Delta w = c(e_1,\theta_L)\ (\text{binding }\mathrm{TC}_L), \tag{50}$$

where the above two equations are from the binding TC_L in (45) and HC in (46) together with the separability, *i.e.*, $c(e_1,\theta_L) - c(e_1,\theta_H) = \frac{\alpha-1}{\alpha}c(e_1,\theta_L)$. As shown in the proof of the theorem above, the maximum feasible human capital accumulation is the fully informative signal's threshold k_I from (29) in Section 3, which can be attained only when $x=1$. Hence, the maximum attainable k given (49) is k_I, whereas the fully uninformative signal in (34) allows us to attain even the upper bound $k_U = \overline{\omega}$. Yet, one should note that a higher k does not necessarily improve social welfare in this model because of both the human capital cost effect and the information cost effect.

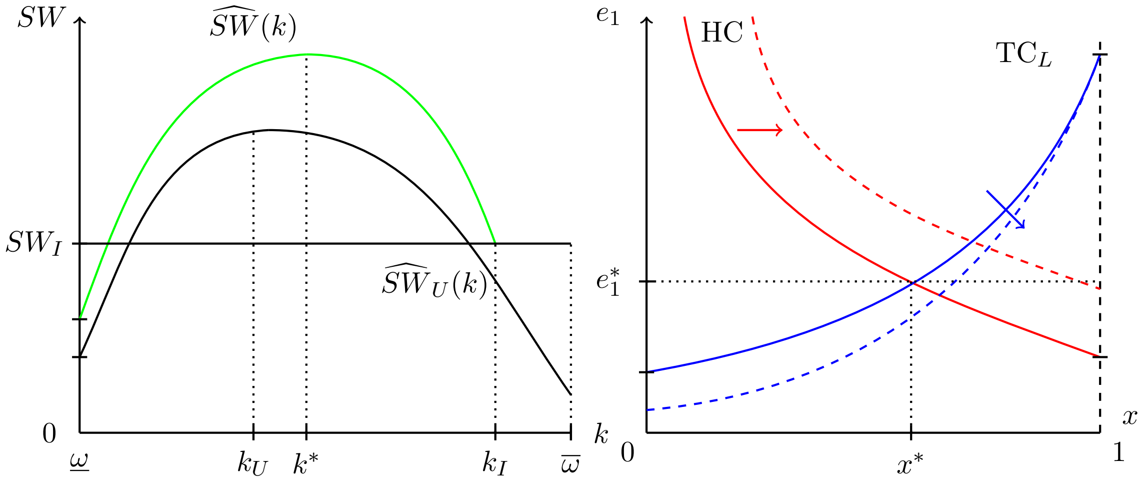

Figure 6: Social welfares and simultaneous equations

4.3 Main characterization

The reformulation allows the planner's problem to be characterized through the one-dimensional welfare function $\widehat{SW}(k)$ in (49). To establish its global structure, we differentiate this function, yielding

$$\frac{d\widehat{SW}(k)}{dk} = \underbrace{g(k)\Delta w}_{\text{human capital effect}} + \underbrace{\left[1 - \frac{\alpha}{\alpha - 1}\right][g(k)k + G(k)]}_{\text{information cost effect}} - \underbrace{g(k)k}_{\text{human capital cost effect}} \tag{51}$$
$$= g(k)\left[\Delta w - \psi(k)\right],$$

where $\psi(\omega) \equiv \frac{\alpha}{\alpha-1}\omega + \frac{1}{\alpha-1}\frac{G(\omega)}{g(\omega)}$ is the virtual cost of human capital investment in (42). If $\psi(k)$ is strictly increasing, then $\widehat{SW}(k)$ satisfies the *single-peaked property*: There exists a unique k^* such that it is strictly increasing for $k < k^*$ and strictly decreasing for $k > k^*$, as depicted (green line) in the left panel of Figure 6. We now present the main characterization result implied by (51).

Theorem 2 *Suppose Assumptions 1-3 hold. Then there exists a unique threshold $k^* \in (\underline{\omega}, k_I)$ satisfying*

$$k^* = \psi^{-1}(\Delta w), \tag{52}$$

which characterizes the optimal education mechanism Γ, where $x^* \in (0, 1)$, $y^* = 0$, and $e_1^* > e_2^* = 0$, with (x^*, e_1^*) being the unique solution to the system of simultaneous equations in (50). Moreover, the optimal education mechanism is partially informative and strictly dominates both the fully informative and fully uninformative mechanisms in terms of social welfare.

To develop the underlying intuition, we return to the original formulation in (47), where the planner directly chooses x and e_1, and show how the single-peakedness of the reformulated

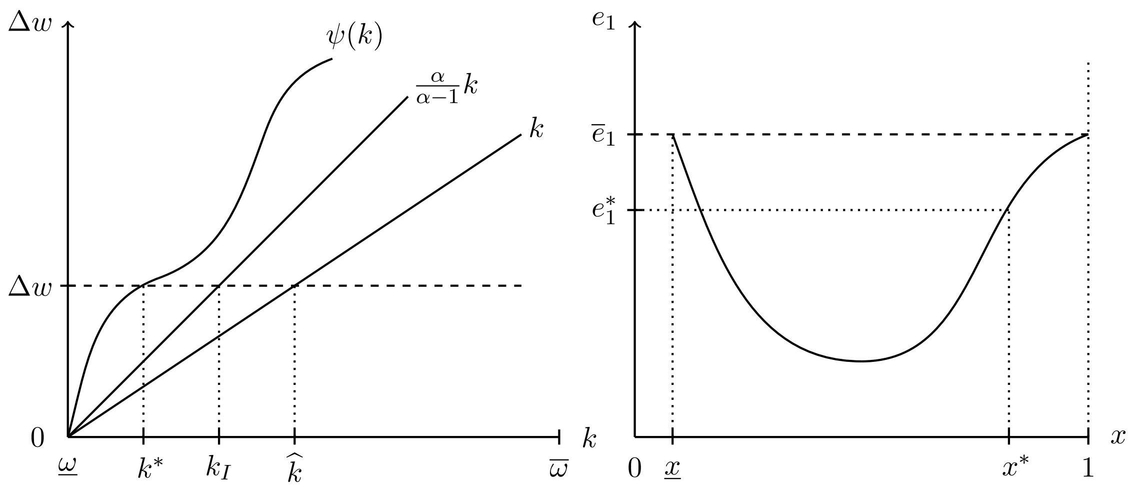

Figure 7: The optimal k^* and the relationship between x and e_1

problem arises from a dual non-monotonicity in these original choice variables. Providing intuition for this general case through its reformulation in (51) requires more careful reasoning than in the fully uninformative case of Section 3, where e in the original problem (31) maps one-to-one into k_U in its reformulation (32). Given the two sides of the problem in (47), namely, the objective function SW and its constraints, we first consider the non-monotonicity arising from the latter. As will be shown by Lemma 6 in Online Appendix C, x is an implicit strictly increasing function of k, and equivalently we may write the induced cutoff as $k(x)$. Substituting $\lambda = G(k(x))$ into the binding TC$_L$ condition in (45) yields

$$\frac{1 - G(k(x))}{1 - xG(k(x))}\Delta w = c(e_1, \theta_L). \tag{53}$$

Now, as $x \to 1$, the mechanism converges to the fully informative benchmark, so the wage differential $D(\mu)\Delta w$ converges to its maximum value, Δw. This requires e_1 to be high in order to deter low-quality agents from deviating to the higher education level. Interestingly, a similar force arises as $x \to \underline{x}$. By the monotonicity of the cutoff k with respect to x, we have $k \to \underline{\omega}$. Although agents of both quality levels choose e_2, almost all individuals choosing e_2 are low-quality agents. Thus, $\frac{1-G(k(x))}{1-xG(k(x))} \to 1$ on the left-hand side of (53) and $D(\mu)\Delta w \to \Delta w$ again, implying that e_1 must also be high. This generates the non-monotonic relationship between x and e_1, as depicted in Figure 7. In particular, this results in a positive relationship between x and e_1 for sufficiently large x, but a negative relationship for sufficiently small x. To see this graphically, if we increase k in the left panel of Figure 6, then the downward-sloping human capital locus in the (x, e_1) space shifts rightward, while the upward-sloping incentive condition locus rotates clockwise around the fixed intercept at $(1, \overline{e}_1)$, as illustrated in the right panel of Figure 6.

Having established the non-monotonicity from the constraints, we now examine the objective function in (47). First, given $y^* = e_2^* = 0$, differentiating SW with respect to e_1

Region	Feasibility	Welfare	Implication
Small x	$(x\uparrow, e_1\downarrow)$	$(x\uparrow, e_1\downarrow)$	feasibility and welfare are aligned
Large x	$(x\uparrow, e_1\uparrow)$	$(x\downarrow, e_1\downarrow)$	feasibility and welfare are in conflict

Table 1: Feasibility and welfare trajectories of (x, e_1)

yields

$$\frac{\partial SW}{\partial e_1} = \frac{\partial k}{\partial e_1} g(k)\Delta w[1 - xD(\mu)] - \lambda x c_e(e_1, \theta_H) < 0, \tag{54}$$

where $\frac{\partial k}{\partial e_1} = -\frac{xc_e(e_1,\theta_H)}{xD_\lambda(\mu)\Delta w g(k)-1} < 0$ and $D_\lambda(\mu) = -\frac{1-x}{(1-x\lambda)^2} < 0$. On the other hand, the derivative of SW with respect to x yields

$$\frac{\partial SW}{\partial x} = \underbrace{\frac{\partial k}{\partial x} g(k)\Delta w[1 - xD(\mu)]}_{\text{net human capital effect}} - \underbrace{\lambda c(e_1, \theta_H)}_{\text{information cost effect}}, \tag{55}$$

where

$$\frac{\partial k}{\partial x} = -\frac{D(\mu)\Delta w + xD_x(\mu)\Delta w - c(e_1, \theta_H)}{xD_\lambda(\mu)\Delta w g(k) - 1} > 0,$$

given $D_x(\mu) = \frac{(1-\lambda)\lambda}{(1-x\lambda)^2} > 0$ and $D(\mu)\Delta w - c(e_1, \theta_H) > D(\mu)\Delta w - c(e_1, \theta_L) = 0$. Unlike the negative sign of $\frac{\partial SW}{\partial e_1}$, $\frac{\partial SW}{\partial x}$ contains two opposing effects in (55): the net human capital effect and the information cost effect. The former captures the welfare gain from increasing the fraction of individuals who acquire higher education, whereas the latter reflects the additional education costs incurred by high-quality agents. As shown on the constraint side, the wage differential converges to its maximum value Δw both as $x \to \underline{x}$ and as $x \to 1$. By contrast, the welfare effect depends on the aggregate gain $xD(\mu)\Delta w$. As $x \to \underline{x}$, the proportion of individuals receiving higher education remains small despite $D(\mu)\Delta w \to \Delta w$, so the potential gain from expanding higher education remains large and $\frac{\partial SW}{\partial x} > 0$. In contrast, as $x \to 1$, both x and $D(\mu)$ approach one, so the human capital gain is already substantial while the information cost effect becomes dominant, implying $\frac{\partial SW}{\partial x} < 0$. Hence, if x is small, the positive net human capital effect dominates the information cost effect, but if x is large, the relationship reverses. Table 1 summarizes how the non-monotonicity arising from the constraints interacts with the non-monotonicity arising from the objective function. If x is relatively small, an increase in x is accompanied by a decrease in e_1 along the constraint locus in the right panel of Figure 7. Both changes increase SW by (54) and (55), so feasibility and welfare are aligned. However, if x is relatively large, an increase in x is accompanied by an increase in e_1 along the constraint locus. Both changes then decrease SW by (54) and (55), so feasibility and welfare are in conflict.

This dual non-monotonicity in the original problem corresponds to the single-peakedness of $\widehat{SW}(k)$ in the reformulated problem. The resulting optimal k^* in the left panel of Figure

7 is thus lower than the fully informative cutoff k_I in (29). In the first-best case with observable ω, the information cost effect is absent, and human capital accumulation is therefore highest. By contrast, achieving k_I under the fully informative signal requires socially excessive investment in both information generation and human capital, so the optimal cutoff satisfies $k^* < k_I$.

The reformulation makes the welfare dominance over the fully informative and fully uninformative signals particularly transparent. Consider first the fully informative signal. We reorganize the terms in (51) according to whether they apply to the marginal individual or to individuals already assigned to higher education:

$$\frac{d\widehat{SW}(k)}{dk} = \underbrace{\left[\Delta w - \frac{\alpha}{\alpha-1}k\right] g(k)}_{\text{net marginal effect}} - \underbrace{\frac{1}{\alpha-1}G(k)}_{\text{inframarginal information-cost effect}}. \tag{56}$$

The net marginal effect includes the human capital benefit generated by the newly assigned marginal individual, net of both the marginal human capital investment cost and the marginal information cost.[28] By contrast, the second term captures the inframarginal information cost. Expanding higher education requires not only assigning an additional marginal individual to higher education, but also increasing the education level of all individuals already assigned to higher education. The net marginal effect is positive for all $k < k_I = \frac{\alpha-1}{\alpha}\Delta w$ and equals zero at $k = k_I$, whereas the inframarginal information cost is zero at the lower boundary and increases with k. The derivative therefore crosses zero once at an interior threshold $k^* \in (\underline{\omega}, k_I)$. In particular, $\left.\frac{d\widehat{SW}(k)}{dk}\right|_{k=k_I} = -\frac{1}{\alpha-1}G(k_I) < 0$. This observation reveals what we term the *information-cost trap*. At the fully informative threshold, the human capital benefit generated by the marginal individual is exactly offset by the marginal human capital investment and information costs, while the inframarginal information cost remains strictly positive. Full informativeness is therefore socially excessive. Unlike the Myersonian virtual-cost term, the inframarginal term here captures a genuine resource cost of information acquisition.

Consider next the fully uninformative signal. Recall from Section 3 that $\widehat{SW}_U$ may be convex or concave; in the convex case, the optimal k_U arises at a corner, $k_U = \underline{\omega}$ or $k_U = \overline{\omega}$. The left panel of Figure 6 considers the case in which $\widehat{SW}_U$ is concave, as in the right panel of Figure 3 in Section 3, but regardless of its shape, Theorem 2 shows the dominance of the optimal k^* over the fully uninformative signal. The proof strategy is as follows. If the optimal k_U is no greater than k_I, we can choose $k = k_U$, since such a threshold is feasible in the reformulation. As established in the proof of Theorem 2, it then follows from $\widehat{SW}_U(k_U)$ in (32) and $\widehat{SW}(k)$ in (49) that $\widehat{SW}(k^*) > \widehat{SW}(k_U) \geq \widehat{SW}_U(k_U)$. If, instead, the optimal k_U is greater than k_I, then k_U is not feasible in the reformulated problem, whose domain is $k \in [\underline{\omega}, k_I]$. However, by showing $SW_I > \widehat{SW}_U(k_U)$, we establish the dominance of $\widehat{SW}(k^*)$

[28]Indeed, $\frac{\alpha}{\alpha-1}k = k + \frac{1}{\alpha-1}k$, where k is the marginal human capital investment cost and $\frac{1}{\alpha-1}k$ is the marginal information cost.

over $\widehat{SW}_U(k_U)$ – *i.e.*, $\widehat{SW}(k^*) > SW_I > \widehat{SW}_U(k_U)$ – nevertheless. By (28) and (32), for $k_U > k_I$, we have

$$\begin{aligned} SW_I - \widehat{SW}_U(k_U) &= -\lambda_U \Delta w + \frac{\alpha}{\alpha-1} k_U - \int_{k_I}^{k_U} G(\omega) d\omega \\ &\geq -\frac{\alpha}{\alpha-1} k_I + \frac{\alpha}{\alpha-1} k_U - \int_{k_I}^{k_U} G(\omega) d\omega > 0, \end{aligned}$$

where $\Delta w = \frac{\alpha}{\alpha-1} k_I$, $\lambda_U \leq 1$ and $\frac{\alpha}{\alpha-1} > 1$. This is illustrated in the left panel of Figure 6.

For closed-form solutions, we revisit Example 2 in (35) with support $[\underline{\omega}, \overline{\omega}]$ where $G(\omega) = \frac{\omega - \underline{\omega}}{\overline{\omega} - \underline{\omega}}$ for $\omega \in [\underline{\omega}, \overline{\omega}]$. Then, (51) yields

$$k^* = \frac{\underline{\omega} + (\alpha-1)\Delta w}{\alpha+1}.$$

Now, we consider the two conditions, TC$_L$ and HC:

$$\frac{1-G(k)}{1-xG(k)} \Delta w = \frac{e_1^2}{\theta_L} \text{ and } k = x \frac{\alpha-1}{\alpha} \frac{e_1^2}{\theta_L}. \tag{57}$$

For any given $k \in (\underline{\omega}, k_I)$, there is a unique (x, e_1) that solves the two conditions above, given by

$$x_k = \frac{\alpha k(\overline{\omega} - \underline{\omega})}{(\alpha-1)(\overline{\omega} - k)\Delta w + \alpha k(k - \underline{\omega})} \text{ and } e_k = \sqrt{\frac{[(\alpha-1)(\overline{\omega} - k)\Delta w + \alpha k(k - \underline{\omega})]\theta_L}{(\alpha-1)(\overline{\omega} - \underline{\omega})}}.$$

Previously, the relationships between k and x and between k and e_1 were characterized implicitly by (50), but now with these closed-form solutions, we find a positive relationship between k and x:

$$\frac{dx}{dk} = \frac{[\overline{\omega}(\alpha-1)\Delta w - \alpha k^2]\alpha(\overline{\omega} - \underline{\omega})}{[(\alpha-1)(\overline{\omega} - k)\Delta w + \alpha k(k - \underline{\omega})]^2} > 0 \text{ for all } k \in [\underline{\omega}, k_I],$$

and a non-monotonic relationship between k and e_1:

$$\frac{de_1}{dk} = \frac{\alpha(2k - \underline{\omega} - k_I)\theta_L}{2(\alpha-1)(\overline{\omega} - \underline{\omega}) e_k}.$$

Finally, the set of (x, e_1) that solves (57) can be mapped one-to-one onto $k \in [\underline{\omega}, k_I]$. It contains the following combinations:

$$(x, e_1) = \begin{cases} \left(\frac{\underline{\omega}}{k_I}, \overline{e}_1\right) & \text{if } k = \underline{\omega}, \\ (x_k, e_k) & \text{if } k \in (\underline{\omega}, k_I), \\ (1, \overline{e}_1) & \text{if } k = k_I. \end{cases}$$

This set shows that x rises but e_1 falls when k increases in the range $k < \frac{\underline{\omega}+k_I}{2}$, while both x and e_1 rise when k increases in the range $k > \frac{\underline{\omega}+k_I}{2}$. Thus, increasing k above the threshold $\frac{\underline{\omega}+k_I}{2}$ requires an expansion of higher education – that is, increases in both x and e_1.

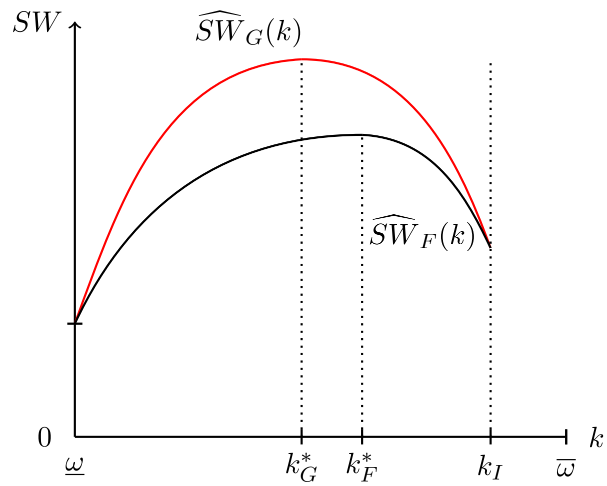

Figure 8: Social welfare and comparative statics analysis

5 Comparative statics analysis

In this section, we conduct comparative statics analyses with respect to three key parameters in the model: $\Delta w, \alpha$, and $\frac{G(\omega)}{g(\omega)}$. Regarding the last parameter, we assume that F dominates G in terms of the reverse hazard rate (RHR) such that $\phi_F(\omega) > \phi_G(\omega)$ for all $\omega \in (\underline{\omega}, \overline{\omega})$, where

$$\phi_F(\omega) \equiv \frac{f(\omega)}{F(\omega)} \text{ and } \phi_G(\omega) \equiv \frac{g(\omega)}{G(\omega)}. \tag{58}$$

As is well known, this assumption implies that F first-order stochastically dominates G, *i.e.*, $F(\omega) < G(\omega)$ for all $\omega \in (\underline{\omega}, \overline{\omega})$.

First, changes in k^* with respect to changes in one of $\Delta w, \alpha$, and $\frac{G(\omega)}{g(\omega)}$ can be immediately identified by (51) for which we denote by k_F^* (resp. k_G^*) the threshold of the distribution F (resp. G). Second, regarding changes in optimal social welfare SW, the simple alternative formulation in (49) allows us to apply the envelope theorem.

Proposition 4 *Suppose Assumptions 1-3 hold. The comparative statics analyses satisfy the following properties.*

(i) k^* *increases, as* Δw *or* α *increases.*

(ii) $k_F^* > k_G^*$ *if* F *dominates* G *in terms of RHR.*

(iii) SW^* *increases, as* Δw *or* α *increases.*

(iv) $SW_F^* < SW_G^*$ *if* F *dominates* G *in terms of RHR.*

The comparative statics analyses in (i) and (iii) with respect to the wage differential Δw or the cost-reduction multiplier α of the investment are intuitive. The increase in the relative benefits of making the human capital investment provides stronger investment incentives,

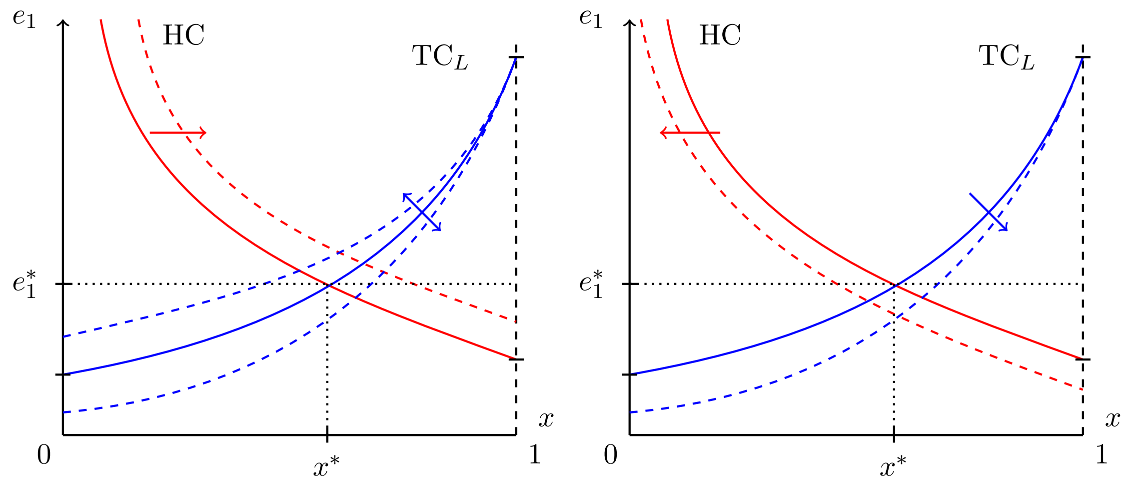

Figure 9: Comparative statics analyses: (a) Δw and (b) α

resulting in a higher threshold k^* and higher welfare SW^*. Now, consider two distributions F and G such that the former dominates the latter in terms of the reverse hazard rate. This dominance means that F is the "worse" distribution in our model, as it contains more high cost types. However, for the worse case, we have the higher threshold in (ii) above. Yet, this does not necessarily imply that a larger fraction of agents invests. Given $F(k) < G(k)$ and $k_F^* > k_G^*$, the ranking between $F(k_F^*)$ and $G(k_G^*)$ is indeterminate. As illustrated in Figure 8, for each k, the worse distribution yields the lower social welfare, resulting in $\widehat{SW}_F(k_F^*) < \widehat{SW}_G(k_G^*)$, in spite of $k_F^* > k_G^*$.

The comparative statics for (x^*, e_1^*) are analytically ambiguous. The indeterminacy resembles that of a standard demand-supply analysis, in which an exogenous parameter change shifts both curves in the same direction, leaving the effect on the equilibrium price and quantity ambiguous. Here, however, an additional source of ambiguity arises. Recall the system of simultaneous equations from (50) and Figure 5, where

$$k^* = x\frac{\alpha - 1}{\alpha}c(e_1, \theta_L) \text{ (HC) and } \frac{1 - G(k^*)}{1 - xG(k^*)}\Delta w = c(e_1, \theta_L) \text{ (binding TC}_L\text{)}.$$

For both conditions, as we change one of the parameters, k^* changes as well, following Proposition 4, which becomes an additional source for the indeterminacy. For changes in the wage differential Δw, the cost-reduction multiplier α, and the distribution, we indeed have both the demand-supply-type source and the endogenous human capital source. For example, a change in Δw affects not only Δw itself but also $\frac{1-G(k^*)}{1-xG(k^*)}$ in the binding TC$_L$. As illustrated in the left panel of Figure 9, the iso-binding TC$_L$ line can shift both to the right and to the left.

We combine (HC) and (TC$_L$) by substituting out the term $c(e_1, \theta_L)$ such that $\frac{1-G(k^*)}{1-xG(k^*)}\Delta w - k^*\frac{\alpha}{x(\alpha-1)} = 0$. To conduct the comparative statics analyses, define

$$\Phi(x, \Delta w, \alpha, G) \equiv \frac{1 - G(k^*)}{1 - xG(k^*)}\Delta w - k^*\frac{\alpha}{x(\alpha - 1)}. \tag{59}$$

In this way, we determine how x^* changes with respect to changes in a parameter among $\Delta w, \alpha, G/g$.

Proposition 5 *Suppose Assumptions 1-3 hold. The comparative statics analyses satisfy the following properties*

(i) Let Δw increase. If $\Phi(x, \Delta w, \alpha, G)$ decreases (resp. increases), then x^* increases (resp. decreases).

(ii) Let α increase. If $\Phi(x, \Delta w, \alpha, G)$ decreases (resp. increases), then x^* increases (resp. decreases).

(iii) Let F dominate G in terms of RHR. If $\Phi(x, \Delta w, \alpha, F) >$ (resp. $<$) $\Phi(x, \Delta w, \alpha, G)$, then x^* increases (resp. decreases).

For example, a sufficient condition for the decreasing $\Phi(x, \Delta w, \alpha, G)$ with respect to Δw is the decreasing $\frac{1-G(k^*)}{1-xG(k^*)}\Delta w$, since by Proposition 4, k^* is increasing in Δw. Even when x increases or decreases, due to the non-monotonic nature of the relationship between x and e_1 from (50) as in the right panel of Figure 7, the sign of e_1^* is indeterminate in general. The closed-form solution in (57) enables us to determine the signs of further comparative statics analysis, in which case, e_1^* increases with Δw, $\underline{\omega}$, and $\overline{\omega}$; and x^* decreases with $\overline{\omega}$.

6 Discussions

6.1 New perspectives on education design

Our framework offers two perspectives on the design of education systems. First, it provides a new interpretation of overeducation. The existing literature typically conceptualizes overeducation as a labor-market mismatch arising from credential inflation, structural imbalances between skill supply and occupational demand, or an excess supply of educated workers (Freeman, 1976; Leuven and Oosterbeek, 2011). By contrast, our model suggests that overeducation can arise from a systematic expansion of higher-education access rather than from a labor-market friction. The optimal mechanism avoids this by deliberately assigning a fraction $1-x^*$ of high-quality individuals to basic education. From this perspective, the partial informativeness of the education system is not a distortion but a valuable policy instrument that allows the planner to adjust the investment threshold in response to changes in the population's cost-type distribution.

The information-cost trap at $x = 1$ carries direct implications for education reform. At the fully informative boundary, the human capital benefit from inducing one additional individual to invest is exactly offset by the cost of human capital investment and higher education at $\overline{e}_1$. What remains is a pure, uncompensated information cost borne by all individuals who already invest. This result sharpens the overeducation perspective: The

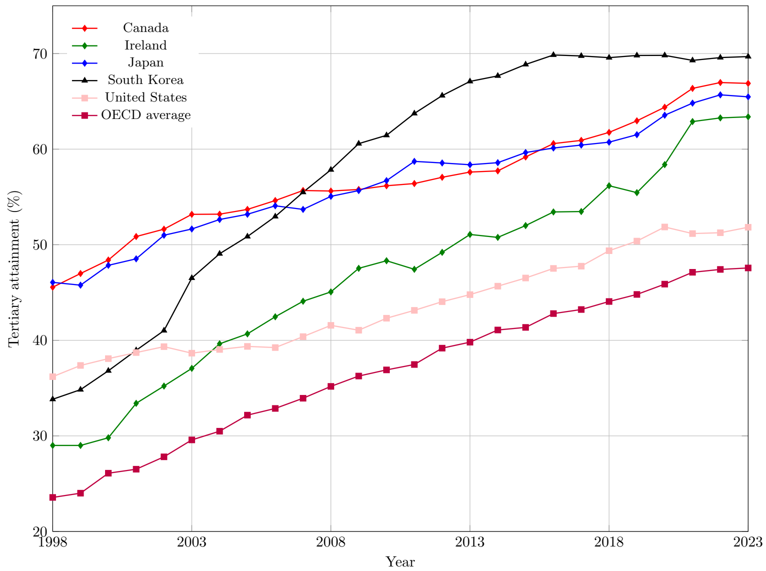

Figure 10: Tertiary Education Attainment

welfare cost of over-expansion is not merely about pushing the investment threshold beyond its optimum, but about imposing an inframarginal burden on this entire group. This trap also reveals that effective reform requires jointly adjusting both x and e_1 rather than either instrument alone. The non-monotone feasible path establishes a unique pair (x, e_1) for each investment threshold k.

A second perspective concerns the role of population composition. The model's cost type ω captures an individual's intrinsic cost of human capital investment, reflecting the cumulative effect of pre-mechanism circumstances – encompassing innate ability and early skill formation (Cunha et al., 2010), family background, school quality, financial constraints (Carneiro and Heckman, 2002), and early childhood environments (Currie, 2009). Policies that lower these baseline costs shift the cost-type distribution G toward lower-cost types, contracting the optimal investment threshold and raising welfare. This contraction directly reduces the inframarginal information cost borne by those who already invest while inducing fewer high-cost types to invest, thereby easing the aggregate investment cost burden. By contrast, expansion policies leave G unchanged and instead raise the threshold by increasing the higher-education access probability x. The welfare effects of implementing both policies simultaneously depend on the initial position of the threshold. Our model shows that population-level improvements may not justify expansion policies: In an overinvestment region, expansion counteracts these improvements by undermining the cost-reducing threshold contraction they deliver.

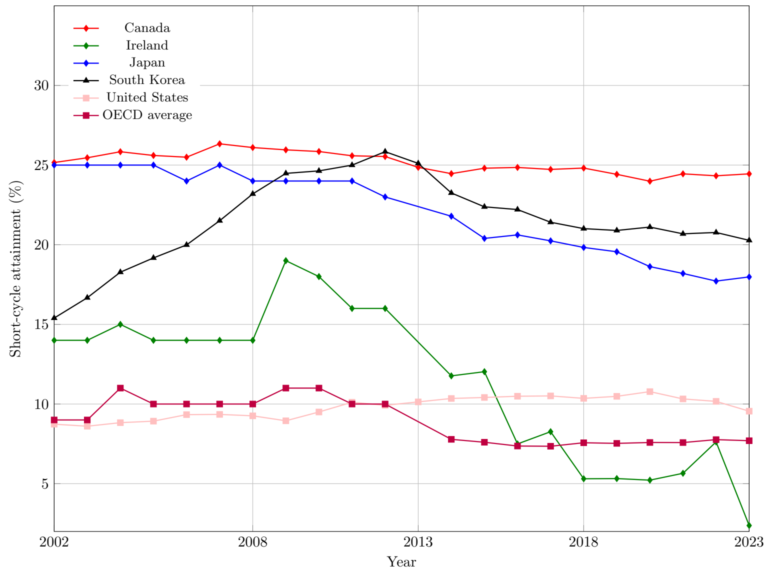

Figure 11: Short-Cycle Tertiary Attainment

6.2 Testable implications

Translating the model's theoretical predictions into empirical implications requires identifying observable counterparts for the key objects x, k, and e_1. There are two challenges. First, neither x nor k is directly observable. Second, what is observable in data is the attainment rate of higher education – the share of individuals who attain higher education – which equals the product of x and $G(k)$ at the optimum, with neither factor separately identified from $xG(k)$. Since the relationship between x and k is one-to-one and strictly monotonic (see Online Appendix C) and $xG(k)$ is strictly monotonic in x or k for a given G, the optimality of the observable attainment rate $xG(k)$ is equivalent to the joint optimality of x and k. While e_1 is the observable education signal in the model, its real-world counterpart – the graduation or curriculum standard of higher education, or program duration – serves as a natural empirical proxy. The joint observation of $xG(k)$ and e_1 generates tractable empirical implications.

When the attainment rate $xG(k)$ is low and the education standard e_1 is high – few individuals attain higher education but at a high academic standard – the economy lies near the underinvestment region, implying that expanding higher-education access raises welfare. Conversely, when both attainment and education standards are high – many individuals attain higher education at a high academic standard – the economy lies near the overinvestment region, implying that further expansion reduces welfare. Between these two extremes, the single-peaked property of the welfare function provides additional identifying information. Since $xG(k)$ is strictly monotonic in k, welfare is also single-peaked in the attainment rate. The identification challenge in the intermediate region is therefore about locating the optimal benchmark $x^*G(k^*)$, which requires structural information about Δw, α, and G.

Overall, these findings emphasize the need for targeted and context-specific approaches in the design of education systems. Policy decisions about the expansion of higher education access should be contingent on the prevailing share of high-quality individuals in the population. To associate our findings with some real-world examples, we turn to OECD data on tertiary educational attainment. Figure 10 displays the share of 25–34-year-olds with tertiary education in Canada, Ireland, Japan, and South Korea – the four countries with the highest tertiary attainment rates – compared to the United States and the OECD average.[29] All tertiary attainment rates have increased substantially over time. At the same time, the composition of tertiary education has shifted toward longer-duration programs, as presented in Figure 11.[30] The share of short-cycle tertiary education has remained relatively stable overall, though with a gradual contraction in recent years in Ireland, Japan, and South Korea alongside continued stability in Canada, the United States, and the OECD average. These trends exhibit a rising rate of tertiary attainment and a compositional shift toward longer-duration tertiary programs. Interpreting the rising attainment rate as a proxy for $xG(k)$ and the shift toward longer-duration programs as a proxy for a rising education standard e_1, our model suggests that policymakers should consider whether their education systems have entered a phase of overinvestment. Consistent with the increasing path of $xG(k)$ and e_1, the model further provides a framework to explain why the average wage premium of tertiary over non-tertiary education has increased over time—a pattern extensively documented by Katz and Murphy (1992), Goldin and Katz (2008), and Acemoglu and Autor (2011).

The comparative statics also generate two clean testable implications for changes in the attainment rate of higher education. First, an increase in the productivity gap between high- and low-quality workers, Δw, raises the optimal investment cutoff k^* and hence the optimal attainment rate $x^*G(k^*)$; economies experiencing skill-biased technical change should exhibit rising attainment rates of higher education at the optimum. Second, improvements in the effectiveness of human capital investment – raising the cost-reduction multiplier α – similarly increase the optimal attainment rate $x^*G(k^*)$; economies in which human capital investment is more effective at reducing education costs should also exhibit higher attainment rates at the optimum. These two implications provide a framework for interpreting observed trends in tertiary attainment across countries – the widely observed rises in attainment rates are consistent with optimal adjustment to increases in Δw or α, while calling for closer scrutiny when the education standard e_1 is also rising.

[29] Each marker represents the percentage of the population aged 25–34 that attained tertiary education. Data were obtained from the OECD Data Explorer (https://www.oecd.org/en/data/datasets/oecd-DE.html, accessed August 2025), except that data for Ireland from 1998 to 1999 were collected from OECD Education at a Glance (2000-2001 editions).

[30] Each marker indicates the percentage of the population aged 25–34 that attained short-cycle tertiary education. Data for Ireland, Japan, and the OECD average from 2001 to 2012 were sourced from the OECD Education at a Glance (2002-2014 editions). Data for Ireland, Japan, and the OECD average from 2014 to 2023, as well as data for Canada, South Korea, and the United States, were obtained from the OECD Data Explorer (https://www.oecd.org/en/data/datasets/oecd-DE.html, accessed August 2025). No data were available for Ireland, Japan, or the OECD average for the year 2013 from either source.

7 Concluding remarks

This paper has advanced a framework in which education is treated as an institution subject to optimal design. By modeling the education system as an endogenous institution that jointly governs information transmission and incentives for human capital investment, we have shown that its socially optimal design departs sharply from canonical benchmarks. Rather than taking the form of either full separation or pooling, the optimal design is characterized by a partially informative structure.

The central theoretical insight is that the single-peakedness of the reformulated problem reflects two interacting non-monotonicities in the original problem, one arising from the feasibility constraints and the other from the objective function. Their interaction determines how educational expansion affects both marginal entrants and individuals already assigned to higher education, thereby generating the marginal and inframarginal effects that yield an interior optimum and explain why excessive educational expansion can reduce social welfare.

This insight reframes overeducation as a systemic outcome of the design and expansion of education rather than merely as a labor market mismatch. The model also generates testable predictions: Observable patterns of educational attainment can indicate whether an education system is approaching the transition from underinvestment to overinvestment. These predictions carry clear policy relevance, since further educational expansion may reduce welfare once this transition has occurred, while shadow education may push the system further from the social optimum. Though the model abstracts from richer dynamics and institutional frictions, it offers a tractable design-based approach and advances the view that education is not merely an investment or a signal, but an institution whose design determines whether expansion raises or lowers social welfare.

Appendix: Proofs

Proof of Lemma 3. Consider TC in (8) for $\omega < k$:

$$\begin{aligned} &x\{\mathbb{E}^{\mu}[W|e_1] - c(e_1, \theta_H)\} + (1-x)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2, \theta_H)\} \\ \geq\ &y\{\mathbb{E}^{\mu}[W|e_1] - c(e_1, \theta_H)\} + (1-y)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2, \theta_H)\}, \end{aligned}$$

and consider TC in (8) for $\omega > k$:

$$\begin{aligned} &y\{\mathbb{E}^{\mu}[W|e_1] - c(e_1, \theta_L)\} + (1-y)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2, \theta_L)\} \\ \geq\ &x\{\mathbb{E}^{\mu}[W|e_1] - c(e_1, \theta_L)\} + (1-x)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2, \theta_L)\}, \end{aligned}$$

Combining them yields

$$(x-y)\{c(e_2, \theta_H) - c(e_1, \theta_H) - [c(e_2, \theta_L) - c(e_1, \theta_L)]\} \geq 0.$$

The result follows from the single-crossing property of $c(e, \theta)$. ■

Proof of Proposition 3. Suppose that for an optimal mechanism Γ, TC_L is not binding. The proof consists of two parts: $e_2 > 0$ and $e_2 = 0$.

Part 1. $e_2 > 0$. The first part consists of three steps.

Step 1. Suppose that $\Gamma = (\rho, \pi)$ is optimal with some k. Let us fix k and also x, y. Then, by the implicit function theorem given (21), we have

$$\frac{de_2}{de_1} = -\frac{xc_e(e_1, \theta_H) - yc_e(e_1, \theta_L)}{(1-x)c_e(e_2, \theta_H) - (1-y)c_e(e_2, \theta_L)}. \tag{60}$$

Given $x > y$ and $c_e(e_2, \theta_H) < c_e(e_2, \theta_L)$, it is always the case that $(1-x)c_e(e_2, \theta_H) - (1-y)c_e(e_2, \theta_L) < 0$ for $e_2 > 0$, but the sign of the numerator, $xc_e(e_1, \theta_H) - yc_e(e_1, \theta_L)$, cannot be pre-determined. This means that the sign of $\frac{de_2}{de_1}$ cannot be determined either.

Now, by taking the derivative of the social welfare with respect to e_1 and by incorporating $\frac{de_2}{de_1}$ into it, we have

$$\begin{aligned}
\frac{\partial SW}{\partial e_1} &= -\lambda xc_e(e_1, \theta_H) - (1-\lambda)yc_e(e_1, \theta_L) - [\lambda(1-x)c_e(e_2, \theta_H) + (1-\lambda)(1-y)c_e(e_2, \theta_L)]\frac{de_2}{de_1} \\
&= -\lambda xc_e(e_1, \theta_H) - (1-\lambda)yc_e(e_1, \theta_L) \\
&\quad + [\lambda(1-x)c_e(e_2, \theta_H) + (1-\lambda)(1-y)c_e(e_2, \theta_L)]\frac{xc_e(e_1, \theta_H) - yc_e(e_1, \theta_L)}{(1-x)c_e(e_2, \theta_H) - (1-y)c_e(e_2, \theta_L)} \\
&= -yc_e(e_1, \theta_L) + (1-y)c_e(e_2, \theta_L)\frac{xc_e(e_1, \theta_H) - yc_e(e_1, \theta_L)}{(1-x)c_e(e_2, \theta_H) - (1-y)c_e(e_2, \theta_L)},
\end{aligned}$$

where the last equality follows by rearranging the first term in the second line and the second term in the third line above such that

$$-\lambda xc_e(e_1, \theta_H) - (1-\lambda)yc_e(e_1, \theta_L) = -yc_e(e_1, \theta_L) - \lambda[xc_e(e_1, \theta_H) - yc_e(e_1, \theta_L)],$$
$$\lambda(1-x)c_e(e_2, \theta_H) + (1-\lambda)(1-y)c_e(e_2, \theta_L) = (1-y)c_e(e_2, \theta_L) + \lambda[(1-x)c_e(e_2, \theta_H) - (1-y)c_e(e_2, \theta_L)].$$

The last line yields

$$\begin{aligned}
&-yc_e(e_1, \theta_L) + (1-y)c_e(e_2, \theta_L)\frac{xc_e(e_1, \theta_H) - yc_e(e_1, \theta_L)}{(1-x)c_e(e_2, \theta_H) - (1-y)c_e(e_2, \theta_L)} \\
&= \frac{-(1-x)yc_e(e_1, \theta_L)c_e(e_2, \theta_H) + x(1-y)c_e(e_1, \theta_H)c_e(e_2, \theta_L)}{(1-x)c_e(e_2, \theta_H) - (1-y)c_e(e_2, \theta_L)} < 0,
\end{aligned} \tag{61}$$

where the negative sign follows from $x(1-y) > (1-x)y$ and the assumption in (14) implying

$$\frac{c_e(e_1, \theta_L)}{c_e(e_1, \theta_H)} = \frac{c_e(e_2, \theta_L)}{c_e(e_2, \theta_H)} = \alpha.$$

Step 2. Consider TC_H and taking the derivative of the term in it with respect to e_1 results in

$$\begin{aligned}
-c_e(e_1, \theta_H) + c_e(e_2, \theta_H)\frac{de_2}{de_1} &= -c_e(e_1, \theta_H) - c_e(e_2, \theta_H)\frac{xc_e(e_1, \theta_H) - yc_e(e_1, \theta_L)}{(1-x)c_e(e_2, \theta_H) - (1-y)c_e(e_2, \theta_L)} \\
&= \frac{-c_e(e_1, \theta_H)[c_e(e_2, \theta_H) - c_e(e_2, \theta_L)]}{(1-x)c_e(e_2, \theta_H) - (1-y)c_e(e_2, \theta_L)} < 0,
\end{aligned} \tag{62}$$

where $yc_e(e_1, \theta_L)c_e(e_2, \theta_H) - yc_e(e_1, \theta_H)c_e(e_2, \theta_L) = 0$ by the assumption in (14).

Consider IR_L and taking the derivative of the term in it with respect to e_1 results in

$$\begin{aligned} & -yc_e(e_1, \theta_L) - (1-y)c_e(e_2, \theta_L)\frac{de_2}{de_1} \\ = & -yc_e(e_1, \theta_L) + (1-y)c_e(e_2, \theta_L)\frac{xc_e(e_1, \theta_H) - yc_e(e_1, \theta_L)}{(1-x)c_e(e_2, \theta_H) - (1-y)c_e(e_2, \theta_L)} \\ = & \frac{-(1-x)yc_e(e_1, \theta_L)c_e(e_2, \theta_H) + x(1-y)c_e(e_1, \theta_H)c_e(e_2, \theta_L)}{(1-x)c_e(e_2, \theta_H) - (1-y)c_e(e_2, \theta_L)} < 0, \end{aligned}$$

where the sign follows as in the case of the derivative of SW with respect to e_1.

Step 3. Suppose, on the contrary, that TC_L is not binding, *i.e.*, $D(\mu)\Delta w - c(e_1, \theta_L) + c(e_2, \theta_L) < 0$. Then, if the principal decreases e_1, while holding k fixed, by Step 1, SW strictly increases. Since the term in IR_L strictly increases by Step 2 too, the decrease in e_1 still makes IR_L hold, which implies IR_H. In addition, by Step 2, the term in TC_H strictly increases, so the decrease in e_1 still makes TC_H hold, and also if the decrease is sufficiently small, TC_L holds. This means that the principal can increase the social welfare by decreasing e_1 together with changing e_2, while fixing k, x and y. We have a contradiction.

Part 2. $e_2 = 0$. At $e_2 = 0$, the approach in Part 1 may not apply because a decrease in e_1 may require a decrease in e_2, which is infeasible at the boundary. We consider two cases: $xc_e(e_1, \theta_H) - yc_e(e_1, \theta_L) < 0$ and $xc_e(e_1, \theta_H) - yc_e(e_1, \theta_L) \geq 0$.

Case 1. $xc_e(e_1, \theta_H) - yc_e(e_1, \theta_L) < 0$. By the assumption in (14), $xc_e(e_1, \theta_H) - yc_e(e_1, \theta_L) = c_e(e_1, \theta_H)(x - \alpha y)$, so $x - \alpha y < 0$. Since $c(e, \theta_L) = \alpha c(e, \theta_H)$, holding k, x, and y fixed, (21) can be written as

$$k = (x-y)D(\mu)\Delta w - [x - \alpha y]c(e_1, \theta_H) - [(1-x) - \alpha(1-y)]c(e_2, \theta_H).$$

Since $x - \alpha y < 0$ and $(1-x) - \alpha(1-y) < 0$, a sufficiently small decrease in e_1 can be accompanied by an increase in e_2 from zero so as to keep (21) satisfied. For $e_2 > 0$, the calculations in Part 1 apply. In particular, (61) can be rewritten as

$$\frac{-(1-x)yc_e(e_1, \theta_L) + x(1-y)c_e(e_1, \theta_H)\frac{c_e(e_2, \theta_L)}{c_e(e_2, \theta_H)}}{(1-x) - (1-y)\frac{c_e(e_2, \theta_L)}{c_e(e_2, \theta_H)}}.$$

Under (14), $\frac{c_e(e_2, \theta_L)}{c_e(e_2, \theta_H)} = \alpha$ for every $e_2 > 0$, so the above expression reduces to $\frac{\alpha c_e(e_1, \theta_H)(x-y)}{(1-x) - \alpha(1-y)} < 0$, which is independent of e_2. The same cancellation applies to the derivatives of the terms in TC_H and IR_L in Step 2 of Part 1, so their signs are likewise independent of e_2. Hence all three signs hold for every $e_2 > 0$. Together with the feasible local change constructed above, the argument in Step 3 of Part 1 applies, implying that the principal can strictly increase social welfare while preserving all constraints, which contradicts optimality.

Case 2. $xc_e(e_1, \theta_H) - yc_e(e_1, \theta_L) \geq 0$. In this case, by incorporating $e_2 = 0$ into (21), the social welfare is rewritten as

$$SW = \lambda\{w_H - xc(e_1, \theta_H)\} + (1-\lambda)\{w_L - yc(e_1, \theta_L)\} - \int_{\underline{\omega}}^{k} \omega dG.$$

In addition, TC and IR are given as $D(\mu)\Delta w - c(e_1, \theta_H) \geq 0$, $D(\mu)\Delta w - c(e_1, \theta_L) \leq 0$, $x\,[D(\mu)\Delta w - c(e_1, \theta_H)] + \mathbb{E}^{\mu}[W|e_2] \geq 0$, and $y\,[D(\mu)\Delta w - c(e_1, \theta_L)] + \mathbb{E}^{\mu}[W|e_2] \geq 0$. HC becomes

$$k = (x - y)D(\mu)\Delta w - [xc(e_1, \theta_H) - yc(e_1, \theta_L)].$$

Fix k and y, and let x adjust with e_1 along this locus. By the implicit function theorem,

$$\frac{\partial x}{\partial e_1} = \frac{xc_e(e_1, \theta_H) - yc_e(e_1, \theta_L)}{[D(\mu)\Delta w - c(e_1, \theta_H)] + (x - y)D_x(\mu)\Delta w}.$$

Direct differentiation gives

$$\frac{\partial \mu_1}{\partial x} = \frac{\lambda(1-\lambda)y}{[\lambda x + (1-\lambda)y]^2} \geq 0 \text{ and } \frac{\partial \mu_2}{\partial x} = -\frac{\lambda(1-\lambda)(1-y)}{[\lambda(1-x) + (1-\lambda)(1-y)]^2} < 0,$$

so $D_x(\mu) = \frac{\partial \mu_1}{\partial x} - \frac{\partial \mu_2}{\partial x} > 0$. Since TC_H implies $D(\mu)\Delta w - c(e_1, \theta_H) \geq 0$ and $x > y$, the denominator above is strictly positive. The numerator is nonnegative by the case hypothesis, and hence $\frac{\partial x}{\partial e_1} \geq 0$. Therefore,

$$\frac{\partial SW}{\partial e_1} = -\lambda c(e_1, \theta_H)\frac{\partial x}{\partial e_1} - \lambda xc_e(e_1, \theta_H) - (1-\lambda)yc_e(e_1, \theta_L) < 0.$$

Thus, decreasing e_1 together with the corresponding change in x strictly increases social welfare.

It remains to verify that the constraints continue to hold. Consider TC_H. Taking the derivative of the term in TC_H with respect to e_1 gives

$$D_x(\mu)\Delta w\frac{\partial x}{\partial e_1} - c_e(e_1, \theta_H) = \frac{D_x(\mu)\Delta w\, y[c_e(e_1, \theta_H) - c_e(e_1, \theta_L)] - c_e(e_1, \theta_H)[D(\mu)\Delta w - c(e_1, \theta_H)]}{[D(\mu)\Delta w - c(e_1, \theta_H)] + (x - y)D_x(\mu)\Delta w} \leq 0.$$

The inequality follows from $c_e(e_1, \theta_H) < c_e(e_1, \theta_L)$ and $D(\mu)\Delta w - c(e_1, \theta_H) \geq 0$. Hence, as e_1 decreases, the term in TC_H weakly increases, so TC_H continues to hold. Consider IR_L. Since $e_2 = 0$, $\mathbb{E}^{\mu}[W|e_2] = w_L + \mu_2\Delta w$. Taking the derivative of the term in IR_L with respect to e_1 gives

$$y\left[D_x(\mu)\Delta w\frac{\partial x}{\partial e_1} - c_e(e_1, \theta_L)\right] + \frac{\partial \mu_2}{\partial x}\Delta w\frac{\partial x}{\partial e_1} = \left[y\frac{\partial \mu_1}{\partial x} + (1-y)\frac{\partial \mu_2}{\partial x}\right]\Delta w\frac{\partial x}{\partial e_1} - yc_e(e_1, \theta_L).$$

Moreover,

$$y\frac{\partial \mu_1}{\partial x} + (1-y)\frac{\partial \mu_2}{\partial x} = \lambda(1-\lambda)\left[\frac{y^2}{[\lambda x + (1-\lambda)y]^2} - \frac{(1-y)^2}{[\lambda(1-x) + (1-\lambda)(1-y)]^2}\right].$$

Since $y[\lambda(1-x) + (1-\lambda)(1-y)] - (1-y)[\lambda x + (1-\lambda)y] = -\lambda(x-y) < 0$, we have

$$\frac{y}{\lambda x + (1-\lambda)y} < \frac{1-y}{\lambda(1-x) + (1-\lambda)(1-y)},$$

and $y\frac{\partial \mu_1}{\partial x} + (1-y)\frac{\partial \mu_2}{\partial x} < 0$. Since $\frac{\partial x}{\partial e_1} \geq 0$, the derivative of the term in IR_L with respect to e_1 is nonpositive. Hence, as e_1 decreases, the term in IR_L weakly increases, so IR_L continues to hold, which implies IR_H. Since TC_L is strictly slack and $x > y$ holds strictly at the original mechanism, both continue to hold for a sufficiently small decrease in e_1 by continuity. Therefore, the principal can increase social welfare by decreasing e_1 together with changing x, while fixing k and y, which contradicts optimality. ■

Proof of Theorem 1. First, we rewrite the social welfare in (16) as (48) such that

$$\begin{aligned} SW(k) &= \lambda\{w_H - xc(e_1,\theta_H)\} + (1-\lambda)w_L - \int_{\underline{\omega}}^{k} \omega dG \\ &= \mathbb{E}^\lambda[W] - \lambda xc(e_1,\theta_H) - \int_{\underline{\omega}}^{k} \omega dG \\ &= \mathbb{E}^\lambda[W] - \lambda[xc(e_1,\theta_L) - k] - \int_{\underline{\omega}}^{k} \omega dG \\ &= \mathbb{E}^\lambda[W] + \lambda k - \lambda k \frac{c(e_1,\theta_L)}{c(e_1,\theta_L) - c(e_1,\theta_H)} - \int_{\underline{\omega}}^{k} \omega dG \\ &= \mathbb{E}^\lambda[W] + \left[1 - \frac{\alpha}{\alpha-1}\right]\lambda k - \int_{\underline{\omega}}^{k} \omega dG, \end{aligned}$$

where the third and the fourth equalities follow from $k = x[c(e_1,\theta_L) - c(e_1,\theta_H)]$ in (46) and the last from the separability in (14).

Before proceeding further, show that $k > k_I$ is not feasible. Suppose, on the contrary, that $k > k_I$. By (29) and (46),

$$x[c(e_1,\theta_L) - c(e_1,\theta_H)] = k > k_I = c(\bar{e}_1,\theta_L) - c(\bar{e}_1,\theta_H),$$

which yields $e_1 > \bar{e}_1$. Then, we have a contradiction with (45):

$$\frac{1-\lambda}{1-x\lambda}\Delta w - c(e_1,\theta_L) < \frac{1-\lambda}{1-x\lambda}\Delta w - c(\bar{e}_1,\theta_L) = \frac{1-\lambda}{1-x\lambda}\Delta w - \Delta w = \frac{-(1-x)\lambda}{1-x\lambda}\Delta w \leq 0 \text{ for all } x,$$

where $D(\mu) = \frac{1-\lambda}{1-x\lambda}$ in (45) and the first equality above follows from (27).

Since it is immediate that for each $(x, e_1) \in [0,1] \times \mathbb{R}_+$, there exists a unique $k \in [\underline{\omega}, k_I]$ given (46), show that for each $k \in [\underline{\omega}, k_I]$, there exists a unique $(x, e_1) \in [0,1] \times \mathbb{R}_+$. We divide the proof into three cases: (i) $k = \underline{\omega}$ (ii) $k = k_I$ (iii) $k \in (\underline{\omega}, k_I)$.

Case (i) $k = \underline{\omega}$. This yields $D(\mu) = 1$, so by (45) and (46),

$$\Delta w = c(\bar{e}_1,\theta_L) \text{ and } \underline{\omega} = x\frac{\alpha-1}{\alpha}\Delta w,$$

where (46) is rewritten as $\underline{\omega} = x\frac{\alpha-1}{\alpha}\Delta w$ since $c(e_1,\theta_L) = \Delta w$ and $c(e_1,\theta_H) = \frac{\Delta w}{\alpha}$. Recall that $\bar{e}_1$ is defined by the equality from $\Delta w = c(\bar{e}_1,\theta_L)$ in (27). Then, a unique x is derived as

$$\underline{x} = \underline{\omega}\frac{\alpha}{\alpha-1}\frac{1}{\Delta w}.$$

Case (ii) $k = k_I$. If $x = 1$, $D(\mu) = 1$, so by (45) and (46),

$$\Delta w = c(\bar{e}_1, \theta_L) \text{ and } k_I = \frac{\alpha - 1}{\alpha}\Delta w,$$

which is identical to the fully informative case in (27).

Case (iii) $k \in (\underline{\omega}, k_I)$. By HC in (46), $k = x[c(e_1, \theta_L) - c(e_1, \theta_H)] = x(\alpha - 1)c(e_1, \theta_H)$. Denote

$$\phi(e_1) \equiv \frac{1}{(\alpha - 1)c(e_1, \theta_H)},$$

which is the separable version of ϕ introduced in Online Appendix C. Then, HC in (46) can be rewritten as $\phi(e_1) = \frac{x}{k}$. By incorporating $e_1 = \phi^{-1}\left(\frac{x}{k}\right)$ into the binding TC$_L$ (45), define

$$\Psi(x, k) \equiv \frac{1 - \lambda}{1 - x\lambda}\Delta w - c\left(\phi^{-1}\left(\frac{x}{k}\right), \theta_L\right).$$

For a sufficiently small $x > 0$, there exists a sufficiently large $e_1 = \phi^{-1}\left(\frac{x}{k}\right)$, which yields

$$\Psi(x, k) = \frac{1 - \lambda}{1 - x\lambda}\Delta w - c\left(\phi^{-1}\left(\frac{x}{k}\right), \theta_L\right) < 0,$$

where $(1 - \lambda)\Delta w$ is bounded but $c(e_1, \theta)$ is strictly increasing. Let us choose $x = 1$. By $k < k_I$, the corresponding e_1 must satisfy $e_1 = \phi^{-1}\left(\frac{1}{k}\right) < \bar{e}_1$, implying

$$\Psi(1, k) = \frac{1 - \lambda}{1 - \lambda}\Delta w - c\left(\phi^{-1}\left(\frac{1}{k}\right), \theta_L\right) = \Delta w - c(e_1, \theta_L) > \Delta w - c(\bar{e}_1, \theta_L) = 0.$$

Furthermore, the partial derivative of $\Psi(x, k)$ with respect to x yields

$$\Psi_x(x, k) = \frac{(1 - \lambda)\lambda}{(1 - x\lambda)^2}\Delta w - \frac{1}{k\phi'\left(\phi^{-1}\left(\frac{x}{k}\right)\right)} c_e\left(\phi^{-1}\left(\frac{x}{k}\right), \theta_L\right) > 0 \text{ for all } x > 0.$$

Hence, for each $k \in (\underline{\omega}, k_I)$, there exists a unique x satisfying $\Psi(x, k) = 0$, which implies the existence of a unique (x, e_1) generating k. ■

Proof of Theorem 2. The first-order condition in (51) yields

$$k^* = \begin{cases} \underline{\omega} & \text{if } \Delta w \leq \psi(\underline{\omega}), \\ \psi^{-1}(\Delta w) & \text{if } \Delta w > \psi(\underline{\omega}), \end{cases}$$

where $\Delta w \leq \psi(\underline{\omega})$ violates the range in Assumption 2. Now, we show that the partially informative signal dominates the fully informative signal (Part 1) and the fully uninformative signal (Part 2).

Part 1. Let $x = 1$. By taking the derivative of the social welfare with respect to x at $x = 1$ in (55),

$$\frac{\partial SW}{\partial x} = \frac{\partial k}{\partial x} g(k)[1 - xD(\mu)]\Delta w - \lambda c(e_1, \theta_H).$$

Then, given that $D(\mu) = 1$ in (46), we have $1 - xD(\mu) = 0$, which results in $\frac{\partial SW}{\partial x} < 0$.

Part 2. Divide the proof into two cases: $k_U \leq k_I$ and $k_U > k_I$.

Case 1. $k_U \leq k_I$. Consider $\widehat{SW}_U$ in (32) and $\widehat{SW}$ in (48). Since the principal can implement any $k \in [\underline{\omega}, k_I]$, by choosing $k = k_U$, we have $\widehat{SW}(k_U) \geq \widehat{SW}_U(k_U)$:

$$\begin{aligned}\widehat{SW}(k_U) &= \mathbb{E}^{\lambda_U}[W] - \lambda_U k_U \frac{\alpha}{\alpha - 1} + \int_{\underline{\omega}}^{k_U} G(\omega) d\omega \\ &\geq \mathbb{E}^{\lambda_U}[W] - k_U \frac{\alpha}{\alpha - 1} + \int_{\underline{\omega}}^{k_U} G(\omega) d\omega = \widehat{SW}_U(k_U)\end{aligned}$$

Hence, $\widehat{SW}(k^*) > \widehat{SW}(k_U) \geq \widehat{SW}_U(k_U)$.

Case 2. $k_U > k_I$. By (27) and (30), $e > \overline{e}_1$. Recall (28) where

$$SW_I = w_L + \int_{\underline{\omega}}^{k_I} G(\omega) d\omega$$

Hence, the difference between SW_I and $\widehat{SW}_U(k_U)$ becomes

$$\begin{aligned}SW_I - \widehat{SW}_U(k_U) =& -\lambda_U \Delta w + k_U \frac{\alpha}{\alpha - 1} - \int_{k_I}^{k_U} G(\omega) d\omega \\ >& -\lambda_U \Delta w + k_U \frac{\alpha}{\alpha - 1} - [k_U - k_I] G(k_U) \\ =& [-\Delta w + k_I - k_U] G(k_U) + k_U \frac{\alpha}{\alpha - 1} \\ =& [-c(\overline{e}_1, \theta_H) + c(e, \theta_H) - c(e, \theta_L)] G(k_U) + k_U \frac{\alpha}{\alpha - 1} \\ =& G(k_U)[c(e, \theta_H) - c(\overline{e}_1, \theta_H)] + [1 - G(k_U)] c(e, \theta_L) > 0,\end{aligned}$$

where the second-to-last equality follows from (27) where $\Delta w = c(\overline{e}_1, \theta_L)$ and $k_I = \Delta w - c(\overline{e}_1, \theta_H)$, and the last equality from (30) where $k_U = c(e, \theta_L) - c(e, \theta_H) = \frac{\alpha-1}{\alpha} c(e, \theta_L)$. Then, by Part 1, $\widehat{SW}(k^*) > SW_I > \widehat{SW}_U(k_U)$. ■

Proof of Proposition 4. The proof consists of two parts: changes in k^* and SW^*.

Part 1. Changes in k^*. Given the first-order condition in (51), $\Delta w = \psi(k^*)$, where $\psi(k) = \frac{\alpha}{\alpha-1} k + \frac{1}{\alpha-1} \frac{G(k)}{g(k)}$ and Assumption 3, it is immediate that $\frac{\partial k^*}{\partial \Delta w} > 0$ and $\frac{\partial k^*}{\partial \alpha} > 0$ for (i). Now, from $\phi_F(\omega) > \phi_G(\omega)$ in (58), $k^*_F > k^*_G$ for (ii).

Part 2. Changes in SW^* when F dominates G in terms of RHR. Consider (48) and it follows from the envelope theorem that we obtain (iii). Last, (iv) hinges on the key finding below:

$$\begin{aligned}&\widehat{SW}_F(k) - \widehat{SW}_G(k) \\ &= \mathbb{E}^{F(k)}[W] - F(k) k \frac{\alpha}{\alpha - 1} + \int_{\underline{\omega}}^{k} F(\omega) d\omega - \left\{ \mathbb{E}^{G(k)}[W] - G(k) k \frac{\alpha}{\alpha - 1} + \int_{\underline{\omega}}^{k} G(\omega) d\omega \right\} \\ &= [F(k) - G(k)][k_I - k] \frac{\alpha}{\alpha - 1} + \int_{\underline{\omega}}^{k} [F(\omega) - G(\omega)] d\omega < 0,\end{aligned}$$

where we incorporate $k_I = \frac{\alpha-1}{\alpha}\Delta w$ in (29) into the second line above. ■

Proof of Proposition 5. The derivative of Φ in (59) with respect to x is given as

$$\frac{\partial \Phi}{\partial x} = \frac{G(k^*)}{[1 - xG(k^*)]^2}\Delta w + k^* \frac{\alpha}{x^2(\alpha-1)} > 0.$$

The proof consists of two parts: changes in x^* with respect to Δw and α, and changes in x^* when F dominates G in terms of RHR.

Part 1. Changes in x^* with respect to Δw, α. We aim to find

$$\frac{\partial x^*}{\partial \Delta w} = -\frac{\partial \Phi / \partial \Delta w}{\partial \Phi / \partial x} \text{ and } \frac{\partial x^*}{\partial \alpha} = -\frac{\partial \Phi / \partial \alpha}{\partial \Phi / \partial x}.$$

The derivatives of Φ with respect to Δw and α are

$$\frac{\partial \Phi}{\partial \Delta w} = \frac{\partial \Phi}{\partial k}\frac{\partial k^*}{\partial \Delta w} + \frac{1 - G(k^*)}{1 - xG(k^*)} \text{ and } \frac{\partial \Phi}{\partial \alpha} = \frac{\partial \Phi}{\partial k}\frac{\partial k^*}{\partial \alpha} + k^* \frac{1}{x(\alpha-1)^2},$$

where for each case, the two terms have opposite signs since $\frac{\partial \Phi}{\partial k}\frac{\partial k^*}{\partial \Delta w} < 0$ and $\frac{\partial \Phi}{\partial k}\frac{\partial k^*}{\partial \alpha} < 0$ given Proposition 4 and

$$\frac{\partial \Phi}{\partial k} = \frac{-(1-x)g(k^*)}{[1 - xG(k^*)]^2} - \frac{\alpha}{x(\alpha-1)} < 0.$$

Establishing the definite sign of each of them requires the corresponding sufficient condition in (i), (ii).

Part 2. Changes in x^* when F dominates G in terms of RHR. It consists of two elements. First,

$$\frac{1 - G(k_F^*)}{1 - xG(k_F^*)}\Delta w - k_F^* \frac{\alpha}{x(\alpha-1)} < \frac{1 - G(k_G^*)}{1 - xG(k_G^*)}\Delta w - k_G^* \frac{\alpha}{x(\alpha-1)},$$

and by $F(\omega) < G(\omega)$, $\frac{1-F(k^*)}{1-xF(k^*)} > \frac{1-G(k^*)}{1-xG(k^*)}$, so

$$\frac{1 - F(k^*)}{1 - xF(k^*)}\Delta w - k^* \frac{\alpha}{x(\alpha-1)} > \frac{1 - G(k^*)}{1 - xG(k^*)}\Delta w - k^* \frac{\alpha}{x(\alpha-1)}.$$

Together, we have

$$\begin{aligned}
&\frac{1 - F(k_F^*)}{1 - xF(k_F^*)}\Delta w - k_F^* \frac{\alpha}{x(\alpha-1)} \\
&= \frac{1 - F(k_F^*)}{1 - xF(k_F^*)}\Delta w - k_F^* \frac{\alpha}{x(\alpha-1)} - \left[\frac{1 - G(k_F^*)}{1 - xG(k_F^*)}\Delta w - k_F^* \frac{\alpha}{x(\alpha-1)}\right] \\
&\quad + \frac{1 - G(k_F^*)}{1 - xG(k_F^*)}\Delta w - k_F^* \frac{\alpha}{x(\alpha-1)} - \left[\frac{1 - G(k_G^*)}{1 - xG(k_G^*)}\Delta w - k_G^* \frac{\alpha}{x(\alpha-1)}\right].
\end{aligned}$$

Φ is not differentiable in this case, but the same indeterminate-sign problem arises, which necessitates the sufficient condition in (iii). ■

References

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of Labor Economics*, volume 4, pages 1043–1171. Elsevier.

Aoyagi, M. and Yoo, S. H. (2022). Matching strategic agents on a two-sided platform. *Games and Economic Behavior*, 135:271–296.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of Political Economy*, 70(5, Part 2):9–49.

Bergemann, D. and Morris, S. (2016). Bayes Correlated Equilibrium and the Comparison of Information Structures in Games. *Theoretical Economics*, 11(2):487–522.

Bray, M. (1999). *The Shadow Education System: Private Tutoring and its Implications for Planners*. UNESCO International Institute for Educational Planning.

Bray, M. (2024). Still in Need of Confronting: Shadow Education and its Implications in the Sustainable Development Goals. *International Journal of Educational Development*, 104:102967.

Carneiro, P. and Heckman, J. J. (2002). The evidence on credit constraints in post-secondary schooling. *The Economic Journal*, 112(482):705–734.

Cho, I.-K. and Kreps, D. M. (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics*, 102(2):179–221.

Cunha, F. and Heckman, J. (2007). The technology of skill formation. *American Economic Review*, 97(2):31–47.

Cunha, F., Heckman, J. J., and Schennach, S. M. (2010). Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica*, 78(3):883–931.

Currie, J. (2009). Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood, and human capital development. *Journal of Economic Literature*, 47(1):87–122.

Dang, H.-A. and Rogers, F. H. (2008). The Growing Phenomenon of Private Tutoring: Does It Deepen Human Capital, Widen Inequalities, or Waste Resources? *The World Bank Research Observer*, 23(2):161–200.

Dubey, P. and Geanakoplos, J. (2010). Grading exams: 100, 99, 98, … or A, B, C? *Games and Economic Behavior*, 69(1):72–94.

Eső, P. and Szentes, B. (2007). Optimal information disclosure in auctions and the handicap auction. *The Review of Economic Studies*, 74(3):705–731.

Freeman, R. B. (1976). *The Overeducated American*. Academic Press.

Goldin, C. D. and Katz, L. F. (2008). *The race between education and technology*. Harvard University Press.

Kamenica, E. and Gentzkow, M. (2011). Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615.

Katz, L. F. and Murphy, K. M. (1992). Changes in relative wages, 1963–1987: supply and demand factors. *The Quarterly Journal of Economics*, 107(1):35–78.

Kim, B. and Yoo, S. H. (2022). Grand Mechanism and Population Uncertainty. Working Paper 2204, Institute of Economic Research, Korea University. Available at SSRN 4149079.

Kim, E., Goodman, J., and West, M. R. (2025). Kumon In: The Recent, Rapid Rise of Private Tutoring Centers. *Education Finance and Policy*, pages 1–21.

Kim, S. and Lee, J.-H. (2010). Private Tutoring and Demand for Education in South Korea. *Economic Development and Cultural Change*, 58(2):259–296.

Krishna, K., Lychagin, S., Olszewski, W., Siegel, R., and Tergiman, C. (2026). Pareto improvements in the contest for college admissions. *Review of Economic Studies*, 93(1):629–663.

Leuven, E. and Oosterbeek, H. (2011). Overeducation and Mismatch in the Labor Market. In *Handbook of the Economics of Education*, volume 4, pages 283–326. Elsevier.

Li, H. and Shi, X. (2017). Discriminatory information disclosure. *American Economic Review*, 107(11):3363–3385.

Myerson, R. B. (1982). Optimal Coordination Mechanisms in Generalized Principal–Agent Problems. *Journal of Mathematical Economics*, 10(1):67–81.

Ostrovsky, M. and Schwarz, M. (2010). Information disclosure and unraveling in matching markets. *American Economic Journal: Microeconomics*, 2(2):34–63.

Rothschild, M. and Stiglitz, J. E. (1976). Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information. *The Quarterly Journal of Economics*, 90(4):629–649.

Schultz, T. W. (1961). Investment in human capital. *The American Economic Review*, 51(1):1–17.

Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics*, 87(3):355–374.

Stevenson, D. L. and Baker, D. P. (1992). Shadow Education and Allocation in Formal Schooling: Transition to University in Japan. *American Journal of Sociology*, 97(6):1639–1657.

Zhang, W. and Bray, M. (2020). Comparative Research on Shadow Education: Achievements, Challenges, and the Agenda Ahead. *European Journal of Education*, 55(3):322–341.

Online Appendix

A Full mechanism

Incentive compatibility Formally, a mechanism Γ is incentive compatible if for each $\omega, \omega' \in \Omega$ and for every deviation q or equivalently $\mathbf{q}$ in (3),

$$\begin{aligned} &\mathbb{E}^{\pi}\left[\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W \mid E] \mid \omega\right] - \mathbb{E}^{\pi}\left[\mathbb{E}^{\rho}\big[\mathbf{q}^*(\theta)\cdot \mathbf{c}(E) \mid \omega\big] \mid \omega\right] - \mathbb{E}^{\rho}\big[\mathbf{q}^*(\theta)\cdot (1,0) \mid \omega\big]\omega \\ \geq\ &\mathbb{E}^{\pi}\left[\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W \mid E] \mid \omega'\right] - \mathbb{E}^{\pi}\left[\mathbb{E}^{\rho}\big[\mathbf{q}(\theta)\cdot \mathbf{c}(E) \mid \omega'\big] \mid \omega'\right] - \mathbb{E}^{\rho}\big[\mathbf{q}(\theta)\cdot (1,0) \mid \omega'\big]\omega. \end{aligned} \tag{63}$$

As explained earlier, incentive compatibility must rule out two kinds of deviations. First, an agent with true cost ω may misreport his type as ω'. Second, after receiving a recommendation θ, the agent may disobey it by choosing an alternative action rule $\mathbf{q}(\theta)$. Since G is continuous, a single agent's deviation does not affect the aggregate posterior $\mu_e(\mathbf{q}^*)$, and hence does not affect the expected wage $\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W \mid E]$. Thus, the deviation rule $\mathbf{q}$ affects only the cost terms on the right-hand side of (63).

To see this closely, if we expand (63), incentive compatibility requires that for any $\omega, \omega' \in \Omega$ and every deviation $\mathbf{q}$,

$$\begin{aligned} &\int \pi(e|\omega)\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e]de \\ &- \int \pi(e|\omega)\left[\rho(\theta_H|\omega)\big(\mathbf{q}^*(\theta_H)\cdot \mathbf{c}(e)\big) + \rho(\theta_L|\omega)\big(\mathbf{q}^*(\theta_L)\cdot \mathbf{c}(e)\big)\right] de \\ &- \left[\rho(\theta_H|\omega)\big(\mathbf{q}^*(\theta_H)\cdot (1,0)\big) + \rho(\theta_L|\omega)\big(\mathbf{q}^*(\theta_L)\cdot (1,0)\big)\right]\omega \\ \geq &\int \pi(e|\omega')\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e]de \\ &- \int \pi(e|\omega')\left[\rho(\theta_H|\omega')\big(\mathbf{q}(\theta_H)\cdot \mathbf{c}(e)\big) + \rho(\theta_L|\omega')\big(\mathbf{q}(\theta_L)\cdot \mathbf{c}(e)\big)\right] de \\ &- \left[\rho(\theta_H|\omega')\big(\mathbf{q}(\theta_H)\cdot (1,0)\big) + \rho(\theta_L|\omega')\big(\mathbf{q}(\theta_L)\cdot (1,0)\big)\right]\omega. \end{aligned}$$

With the obedience action rule $\mathbf{q}^*(\theta_H) = (1,0)$ and $\mathbf{q}^*(\theta_L) = (0,1)$, IC above can be restated as

$$\begin{aligned} &\int \pi(e|\omega)\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e]de \\ &- \int \pi(e|\omega)\left[\rho(\theta_H|\omega)c(e,\theta_H) + \rho(\theta_L|\omega)c(e,\theta_L)\right] de \\ &- \rho(\theta_H|\omega)\omega \\ \geq &\int \pi(e|\omega')\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e]de \\ &- \int \pi(e|\omega')\left[\rho(\theta_H|\omega')\big(\mathbf{q}(\theta_H)\cdot \mathbf{c}(e)\big) + \rho(\theta_L|\omega')\big(\mathbf{q}(\theta_L)\cdot \mathbf{c}(e)\big)\right] de \\ &- \left[\rho(\theta_H|\omega')\big(\mathbf{q}(\theta_H)\cdot (1,0)\big) + \rho(\theta_L|\omega')\big(\mathbf{q}(\theta_L)\cdot (1,0)\big)\right]\omega. \end{aligned} \tag{64}$$

Individual rationality In addition, a mechanism Γ is said to be individually rational if for each $\omega \in \Omega$,

$$\mathbb{E}^{\pi}\Big[\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W \mid E] \mid \omega\Big] - \mathbb{E}^{\pi}\Big[\mathbb{E}^{\rho}\Big[\mathbf{q}^*(\theta) \cdot \mathbf{c}(E) \mid \omega\Big] \mid \omega\Big] - \mathbb{E}^{\rho}\Big[\mathbf{q}^*(\theta) \cdot (1,0) \mid \omega\Big]\omega \geq 0. \quad (65)$$

Objective function The social welfare is given by

$$\int_{\Omega} \Big\{\mathbb{E}^{\pi}\Big[\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W \mid E] \mid \omega\Big] - \mathbb{E}^{\pi}\Big[\mathbb{E}^{\rho}\Big[\mathbf{q}^*(\theta) \cdot \mathbf{c}(E) \mid \omega\Big] \mid \omega\Big] - \mathbb{E}^{\rho}\Big[\mathbf{q}^*(\theta) \cdot (1,0) \mid \omega\Big]\omega\Big\} dG(\omega). \quad (66)$$

First-order conditions We have four choice variables x, y, e_1, e_2 in the objective function SW in (13) and the four constraints & HC in (20) - (22). The first-order conditions of (16) with respect to e_1, e_2, x, y are as follows:

$$\begin{aligned} \frac{\partial SW}{\partial e_1} &= \frac{\partial k}{\partial e_1} g(k) \left[\Delta w - xc(e_1, \theta_H) + yc(e_1, \theta_L) - (1-x)c(e_2, \theta_H) + (1-y)c(e_2, \theta_L)\right] \\ &\quad - \lambda x c_e(e_1, \theta_H) - (1-\lambda) y c_e(e_1, \theta_L) - \frac{\partial k}{\partial e_1} k g(k) \\ &= \frac{\partial \lambda}{\partial e_1} \Delta w[1 - (x-y)D(\mu)] - \lambda x c_e(e_1, \theta_H) - (1-\lambda) y c_e(e_1, \theta_L), \end{aligned} \quad (67)$$

where the equality follows from HC in (21). Similarly,

$$\begin{aligned} \frac{\partial SW}{\partial e_2} &= \frac{\partial k}{\partial e_2} g(k) \Delta w[1 - (x-y)D(\mu)] - \lambda(1-x)c_e(e_2, \theta_H) - (1-\lambda)(1-y)c_e(e_2, \theta_L), \\ \frac{\partial SW}{\partial x} &= \frac{\partial k}{\partial x} g(k) \Delta w[1 - (x-y)D(\mu)] - \lambda c(e_1, \theta_H) + \lambda c(e_2, \theta_H), \\ \frac{\partial SW}{\partial y} &= \frac{\partial k}{\partial y} g(k) \Delta w[1 - (x-y)D(\mu)] - (1-\lambda)c(e_1, \theta_L) + (1-\lambda)c(e_2, \theta_L). \end{aligned} \quad (68)$$

Note also that if $(x-y)D_\lambda(\mu)\Delta w g(k) - 1 \neq 0$,

$$\begin{aligned} \frac{\partial k}{\partial e_1} &= -\frac{-[x c_e(e_1, \theta_H) - y c_e(e_1, \theta_L)]}{(x-y)D_\lambda(\mu)\Delta w g(k) - 1}, \\ \frac{\partial k}{\partial e_2} &= -\frac{-[(1-x)c_e(e_2, \theta_H) - (1-y)c_e(e_2, \theta_L)]}{(x-y)D_\lambda(\mu)\Delta w g(k) - 1}, \\ \frac{\partial k}{\partial x} &= -\frac{D(\mu)\Delta w + (x-y)D_x(\mu)\Delta w - c(e_1, \theta_H) + c(e_2, \theta_H)}{(x-y)D_\lambda(\mu)\Delta w g(k) - 1}, \\ \frac{\partial k}{\partial y} &= -\frac{-D(\mu)\Delta w + (x-y)D_y(\mu)\Delta w + c(e_1, \theta_L) - c(e_2, \theta_L)}{(x-y)D_\lambda(\mu)\Delta w g(k) - 1}, \end{aligned}$$

where $(1-x)c_e(e_2, \theta_H) - (1-y)c_e(e_2, \theta_L) < 0$.

B Proofs for Section 2

Proof of Proposition 1. Show (i) of the proposition. The proof consists of two steps.

Step 1. Find the role of the obedience IC. To do so, we consider the obedience IC only with respect to q or equivalently $\mathbf{q}$ in (3) from (64), holding $\omega' = \omega$. Then IC implies

$$\begin{aligned}
&\int \pi(e|\omega)\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e]de \\
&- \int \pi(e|\omega)\left[\rho(\theta_H|\omega)c(e,\theta_H) + \rho(\theta_L|\omega)c(e,\theta_L)\right] de \\
&- \rho(\theta_H|\omega)\omega \\
\geq &\int \pi(e|\omega)\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e]de \\
&- \int \pi(e|\omega)\big[\rho(\theta_H|\omega)\big(\mathbf{q}(\theta_H)\cdot\mathbf{c}(e)\big) + \rho(\theta_L|\omega)\big(\mathbf{q}(\theta_L)\cdot\mathbf{c}(e)\big)\big] de \\
&- \big[\rho(\theta_H|\omega)\big(\mathbf{q}(\theta_H)\cdot(1,0)\big) + \rho(\theta_L|\omega)\big(\mathbf{q}(\theta_L)\cdot(1,0)\big)\big]\,\omega,
\end{aligned} \tag{69}$$

which can be rewritten as

$$\begin{aligned}
&\int \pi(e|\omega)\big[\rho(\theta_H|\omega)\big(\mathbf{q}(\theta_H)\cdot\mathbf{c}(e) - c(e,\theta_H)\big) + \rho(\theta_L|\omega)\big(\mathbf{q}(\theta_L)\cdot\mathbf{c}(e) - c(e,\theta_L)\big)\big] de \\
&+ \big[\rho(\theta_H|\omega)\big(\mathbf{q}(\theta_H)\cdot(1,0) - 1\big) + \rho(\theta_L|\omega)\big(\mathbf{q}(\theta_L)\cdot(1,0)\big)\big]\,\omega \geq 0.
\end{aligned}$$

Further, by simplifying the above, we have

$$\begin{aligned}
&\int \pi(e|\omega)\big[c(e,\theta_L) - c(e,\theta_H)\big]\left[\rho(\theta_H|\omega)q(\theta_L|\theta_H) - \rho(\theta_L|\omega)q(\theta_H|\theta_L)\right] de \\
&- \left[\rho(\theta_H|\omega)q(\theta_L|\theta_H) - \rho(\theta_L|\omega)q(\theta_H|\theta_L)\right]\omega \geq 0 \\
\Leftrightarrow\ &\left[\rho(\theta_H|\omega)q(\theta_L|\theta_H) - \rho(\theta_L|\omega)q(\theta_H|\theta_L)\right]\left\{\int \pi(e|\omega)\big[c(e,\theta_L) - c(e,\theta_H)\big]de - \omega\right\} \geq 0.
\end{aligned} \tag{70}$$

Define the two sets of types such that

$$\begin{aligned}
\Omega_+ &\equiv \left\{\omega \in \Omega : \int \pi(e|\omega)\big[c(e,\theta_L) - c(e,\theta_H)\big]de - \omega > 0\right\}, \\
\Omega_- &\equiv \left\{\omega \in \Omega : \int \pi(e|\omega)\big[c(e,\theta_L) - c(e,\theta_H)\big]de - \omega < 0\right\}.
\end{aligned} \tag{71}$$

Step 2. Show that IC implies

$$\forall\omega \in \Omega_+, \rho(\theta_H|\omega) = 1 \text{ and } \forall\omega \in \Omega_-, \rho(\theta_H|\omega) = 0. \tag{72}$$

Suppose, on the contrary, that for $\omega \in \Omega_+$ in (71), $\rho(\theta_H|\omega) < 1$ (or $\rho(\theta_L|\omega) > 0$). Then, given recommendation θ_H, θ_L, by disobeying $q(\theta_L|\theta_H) = 0, q(\theta_H|\theta_L) = 1$, we have

$$\rho(\theta_H|\omega)q(\theta_L|\theta_H) - \rho(\theta_L|\omega)q(\theta_H|\theta_L) = -\rho(\theta_L|\omega) < 0,$$

which violates IC in (70) from Step 1. Similarly, suppose that for $\omega \in \Omega_-$ in (71), $\rho(\theta_H|\omega) > 0$. Then, given recommendation θ_H, θ_L, by disobeying $q(\theta_L|\theta_H) = 1, q(\theta_H|\theta_L) = 0$, we have

$$\rho(\theta_H|\omega)q(\theta_L|\theta_H) - \rho(\theta_L|\omega)q(\theta_H|\theta_L) = \rho(\theta_H|\omega) > 0,$$

which violates IC in (70) from Step 1. Now, show that there exists $k \in [\underline{\omega}, \overline{\omega}]$ such that

$$\rho(\theta_H|\omega) = \begin{cases} 1 & \text{if } \omega < k, \\ 0 & \text{if } \omega > k. \end{cases}$$

Let $\omega \in \Omega_+$ and $\omega' \in \Omega_-$ and suppose $\omega > \omega'$. Then, to satisfy the full IC constraint in (64), substituting ω, ω' into (72) results in

$$\int \pi(e|\omega)\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e]de - \int \pi(e|\omega)c(e, \theta_H)de - \omega \geq \int \pi(e|\omega')\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e]de - \int \pi(e|\omega')c(e, \theta_L)de,$$

and

$$\int \pi(e|\omega')\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e]de - \int \pi(e|\omega')c(e, \theta_L)de \geq \int \pi(e|\omega)\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e]de - \int \pi(e|\omega)c(e, \theta_H)de - \omega'.$$

By summing them up, we have $\omega \leq \omega'$, which yields a contradiction.

Show (ii) of the proposition. The proof consists of two parts.

Part 1. Show that there exists a payoff-equivalent representation in which $\pi(e|\omega) = \pi(e|\omega')$ for $\omega, \omega' \in \Omega_+$ and similarly for $\omega, \omega' \in \Omega_-$, where Ω_+ and Ω_- are in (71). By (72), for every $\omega \in \Omega_+$, $\rho(\theta_H|\omega) = 1$. Then, from (64),

$$\begin{aligned} &\int \pi(e|\omega)\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e]de - \int \pi(e|\omega)\left[\rho(\theta_H|\omega)c(e, \theta_H) + \rho(\theta_L|\omega)c(e, \theta_L))\right]de \\ = &\int \pi(e|\omega)\left[\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e] - c(e, \theta_H)\right]de. \end{aligned}$$

For the sets Ω_+ and Ω_-, denote the corresponding maximizers by ω_+ and ω_-, respectively:

$$\begin{aligned} \omega_+ &\in \arg\max_{\omega \in \Omega_+} \left\{ \int \pi(e|\omega)\left[\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e] - c(e, \theta_H)\right]de \right\}, \\ \omega_- &\in \arg\max_{\omega \in \Omega_-} \left\{ \int \pi(e|\omega)\left[\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e] - c(e, \theta_L)\right]de \right\}, \end{aligned}$$

where note that by (72), for each $\omega \in \Omega_+$, $\rho(\theta_H|\omega) = 1$ and for each $\omega \in \Omega_-$, $\rho(\theta_H|\omega) = 0$. Next, show that for each $\omega \in \Omega_+$,

$$\int \pi(e|\omega)\left[\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e] - c(e, \theta_H)\right]de = \int \pi(e|\omega_+)\left[\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e] - c(e, \theta_H)\right]de.$$

Suppose, without loss of generality, that there exists $\omega \in \Omega_+$ such that

$$\int \pi(e|\omega_+)\left[\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e] - c(e, \theta_H)\right]de > \int \pi(e|\omega)\left[\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e] - c(e, \theta_H)\right]de,$$

which violates IC in (64), by adding the cost type $-\omega$ to both sides. The same argument applies to Ω_-: for each $\omega \in \Omega_-$,

$$\int \pi(e|\omega) \left[\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e] - c(e,\theta_L)\right] de = \int \pi(e|\omega_-) \left[\mathbb{E}^{\mu_e(\mathbf{q}^*)}[W|e] - c(e,\theta_L)\right] de.$$

Now, define

$$\pi_H(e) \equiv \pi(e|\omega_+) \qquad \text{and} \qquad \pi_L(e) \equiv \pi(e|\omega_-).$$

Then, we construct a payoff-equivalent alternative mapping $\widehat{\pi}$ such that

$$\widehat{\pi}(e|\omega) = \begin{cases} \pi_H(e) & \text{if } \omega \in \Omega_+, \\ \pi_L(e) & \text{if } \omega \in \Omega_-. \end{cases}$$

By the equalities above, replacing π by $\widehat{\pi}$ preserves all payoff-relevant terms, and hence preserves IC, IR, and the objective value. Thus, $\widehat{\pi}$ is payoff equivalent to π.

Part 2. By Part 1, (4) can be simplified as

$$\mu(e) = \frac{\pi_H(e)\lambda}{\pi_H(e)\lambda + \pi_L(e)(1-\lambda)},$$

which also yields $\mathbb{E}[W|e] = \mu(e)w_H + (1-\mu(e))w_L$. Then, type ω agent with quality θ_H obtains the expected payoff

$$\int \pi_H(e)\left[\mu(e)w_H + (1-\mu(e))w_L\right] de - \int \pi_H(e)c(e,\theta_H)de - \omega,$$

whereas each agent with quality θ_L obtains the expected payoff

$$\int \pi_L(e)\left[\mu(e)w_H + (1-\mu(e))w_L\right] de - \int \pi_L(e)c(e,\theta_L)de.$$

We now construct a payoff-equivalent signal structure using only two education levels e_1 and e_2. Let x and y denote $x \equiv \pi(e_1|\omega)$ for $\omega < k$ and $y \equiv \pi(e_1|\omega)$ for $\omega > k$ as in (11). The corresponding posteriors μ_1 and μ_2 are given by (12). The two-level structure is payoff equivalent to the original signal structure if we find x, y, e_1, e_2 satisfying the following four equalities:

$$x\mathbb{E}^{\mu_1}[W|e_1] + (1-x)\mathbb{E}^{\mu_2}[W|e_2] = \int \pi_H(e)\left[\mu(e)w_H + (1-\mu(e))w_L\right] de, \tag{73}$$

$$xc(e_1,\theta_H) + (1-x)c(e_2,\theta_H) = \int \pi_H(e)c(e,\theta_H)de, \tag{74}$$

$$y\mathbb{E}^{\mu_1}[W|e_1] + (1-y)\mathbb{E}^{\mu_2}[W|e_2] = \int \pi_L(e)\left[\mu(e)w_H + (1-\mu(e))w_L\right] de, \tag{75}$$

$$yc(e_1,\theta_L) + (1-y)c(e_2,\theta_L) = \int \pi_L(e)c(e,\theta_L)de. \tag{76}$$

If the equalities are satisfied, for each quality's expected wage and expected cost with a general signal structure on the right-hand sides, there exist the corresponding payoff-equivalent expected wage and cost with only two signals on the left-hand sides. We first characterize the range of the expected wage terms in (73) and (75). First, show

$$0 \le \int \frac{\pi_H(e)\pi_L(e)}{\pi_H(e)\lambda + \pi_L(e)(1-\lambda)} de \le 1. \tag{77}$$

The lower bound is immediate. For the upper bound, using $\mu(e) = \frac{\pi_H(e)\lambda}{\pi_H(e)\lambda+\pi_L(e)(1-\lambda)}$, observe that

$$\lambda(1-\lambda)\int \frac{\pi_H(e)\pi_L(e)}{\pi_H(e)\lambda + \pi_L(e)(1-\lambda)} de = \int \mu(e)(1-\mu(e))\left[\pi_H(e)\lambda + \pi_L(e)(1-\lambda)\right] de.$$

Since the function $f(p) = p(1-p)$ is concave on $[0,1]$, by Jensen's inequality, applied with respect to the probability density $\pi_H(e)\lambda + \pi_L(e)(1-\lambda)$, we have

$$\begin{aligned}
&\int \mu(e)(1-\mu(e))\big[\pi_H(e)\lambda + \pi_L(e)(1-\lambda)\big] de \\
&\le \left(\int \mu(e)\big[\pi_H(e)\lambda + \pi_L(e)(1-\lambda)\big] de\right)\left(1-\int \mu(e)\big[\pi_H(e)\lambda + \pi_L(e)(1-\lambda)\big] de\right) \\
&= \lambda(1-\lambda).
\end{aligned}$$

Now, denote the integral in (77) by

$$T^* \equiv \int \frac{\pi_H(e)\pi_L(e)}{\pi_H(e)\lambda + \pi_L(e)(1-\lambda)} de.$$

Then, the right-hand sides of (73) and (75), respectively, can be written as

$$\begin{aligned}
\int \pi_H(e)\left[\mu(e)w_H + (1-\mu(e))w_L\right] de &= w_H - \Delta w(1-\lambda)T^*, \\
\int \pi_L(e)\left[\mu(e)w_H + (1-\mu(e))w_L\right] de &= w_L + \Delta w\,\lambda\,T^*.
\end{aligned} \tag{78}$$

Since $T^* \in [0,1]$ by (77),

$$\begin{aligned}
\mathbb{E}^\lambda[W] \le \int \pi_H(e)\left[\mu(e)w_H + (1-\mu(e))w_L\right] de \le w_H, \\
w_L \le \int \pi_L(e)\left[\mu(e)w_H + (1-\mu(e))w_L\right] de \le \mathbb{E}^\lambda[W].
\end{aligned} \tag{79}$$

We next reduce the two expected-wage equations to one scalar equation. Let $\Phi_H(x,y)$ and $\Phi_L(x,y)$ denote the left-hand sides of (73) and (75), respectively, and define

$$T(x,y) \equiv \frac{xy}{x\lambda + y(1-\lambda)} + \frac{(1-x)(1-y)}{(1-x)\lambda + (1-y)(1-\lambda)}.$$

First, (12) yields $x(1-\mu_1)+(1-x)(1-\mu_2)=(1-\lambda)T(x,y)$ and $y\mu_1+(1-y)\mu_2=\lambda T(x,y)$. Hence

$$\Phi_H(x,y)=w_H-\Delta w(1-\lambda)T(x,y) \text{ and } \Phi_L(x,y)=w_L+\Delta w\,\lambda\,T(x,y). \tag{80}$$

Combining (80) with (78), both expected-wage equations reduce to the single condition

$$T(x,y)=T^*. \tag{81}$$

We now characterize the solution set to (81). The boundary cases are immediate. If $T^*=0$, then $\pi_H(e)\pi_L(e)=0$ a.e., so the original signal is fully separating; the target wages are (w_H,w_L), which are attained by setting $x=1$ and $y=0$. If $T^*=1$, then $\pi_H=\pi_L$ a.e., so the signal is uninformative; both target wages equal $\mathbb{E}^\lambda[W]$, which is attained by any $x=y$. Hence we focus on the interior case $T^*\in(0,1)$. To solve (81), consider the region $0<x<1, 0\le y\le x$. On this region, T is continuous. Moreover, for each $x\in(0,1)$,

$$T(x,x)=1,\ T(x,0)=\frac{1-x}{(1-x)\lambda+(1-\lambda)},\ \lim_{x\to 1}T(x,0)=0.$$

Since $T^*\in(0,1)$, there exists $x<1$ sufficiently close to 1 such that

$$T(x,0)<T^*<T(x,x)=1.$$

For such an x, $T(x,y)$ is continuous in y on $[0,x]$. Hence, by the intermediate value theorem, there exists $y\in(0,x)$ such that $T(x,y)=T^*$. Moreover, differentiating T with respect to y gives

$$\frac{\partial T}{\partial y}=\lambda\left\{\frac{x^2}{\big[x\lambda+y(1-\lambda)\big]^2}-\frac{(1-x)^2}{\big[(1-x)\lambda+(1-y)(1-\lambda)\big]^2}\right\}.$$

To determine the sign, note that

$$\frac{x}{x\lambda+y(1-\lambda)}-\frac{1-x}{(1-x)\lambda+(1-y)(1-\lambda)}=\frac{(x-y)(1-\lambda)}{\big[x\lambda+y(1-\lambda)\big]\big[(1-x)\lambda+(1-y)(1-\lambda)\big]}.$$

Hence, whenever $x>y$, $\frac{x}{x\lambda+y(1-\lambda)}>\frac{1-x}{(1-x)\lambda+(1-y)(1-\lambda)}$. Therefore, $\frac{\partial T}{\partial y}>0$ whenever $x>y$. It follows that, for each fixed x, the equation $T(x,y)=T^*$ has at most one solution $y\le x$. This establishes that the set of pairs satisfying (73) and (75) is nonempty. The general argument in Online Appendix D, which also covers the separable case, then shows that one can select a wage-matching pair $(\widehat{x},\widehat{y})$ from this set together with education levels $(\widehat{e}_1,\widehat{e}_2)$ satisfying (74) and (76). ■

Proof of Lemma 1. Let $\omega<k$. By Proposition 1, $\rho(\theta_H|\omega)=1$ for $\omega<k$. Given the recommendation θ_H, IC can be divided into three different types of deviation as follows.

(H,i) Deviation on a recommendation θ following a misreport on ω.

$$\begin{aligned}&x\{\mathbb{E}^\mu[W|e_1]-c(e_1,\theta_H)\}+(1-x)\{\mathbb{E}^\mu[W|e_2]-c(e_2,\theta_H)\}\\ \ge\ &y\{\mathbb{E}^\mu[W|e_1]-c(e_1,\theta_H)\}+(1-y)\{\mathbb{E}^\mu[W|e_2]-c(e_2,\theta_H)\}.\end{aligned}$$

(H,ii) Deviation only on a recommendation θ following a truthful report on ω.

$$\begin{aligned} & x\{\mathbb{E}^{\mu}[W|e_1] - c(e_1, \theta_H)\} + (1-x)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2, \theta_H)\} - \omega \\ \geq\ & x\{\mathbb{E}^{\mu}[W|e_1] - c(e_1, \theta_L)\} + (1-x)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2, \theta_L)\}. \end{aligned}$$

(H,iii) No deviation on a recommendation θ following a misreport on ω.

$$\begin{aligned} & x\{\mathbb{E}^{\mu}[W|e_1] - c(e_1, \theta_H)\} + (1-x)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2, \theta_H)\} - \omega \\ \geq\ & y\{\mathbb{E}^{\mu}[W|e_1] - c(e_1, \theta_L)\} + (1-y)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2, \theta_L)\}. \end{aligned}$$

Unlike (H,i), for (H,iii), following a misreport $\omega' > k$, given the principal's recommendation θ_L by $\rho(\theta_H|\omega) = 0$ for $\omega > k$, the agent follows θ_L. We can have the similar (L,i), (L,ii) and (L,iii) for $\omega > k$. Now, we show that (H,ii) is redundant. Note that for (L,i), we have

$$\begin{aligned} & y\{\mathbb{E}^{\mu}[W|e_1] - c(e_1, \theta_L)\} + (1-y)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2, \theta_L)\} \\ \geq\ & x\{\mathbb{E}^{\mu}[W|e_1] - c(e_1, \theta_L)\} + (1-x)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2, \theta_L)\}. \end{aligned}$$

Then, by using (L,i) and (H,iii), we show that (H,ii) holds:

$$\begin{aligned} & x\{\mathbb{E}^{\mu}[W|e_1] - c(e_1, \theta_H)\} + (1-x)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2, \theta_H)\} - \omega \\ \geq\ & y\{\mathbb{E}^{\mu}[W|e_1] - c(e_1, \theta_L)\} + (1-y)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2, \theta_L)\} \\ \geq\ & x\{\mathbb{E}^{\mu}[W|e_1] - c(e_1, \theta_L)\} + (1-x)\{\mathbb{E}^{\mu}[W|e_2] - c(e_2, \theta_L)\}. \end{aligned}$$

With a similar procedure, it can be shown that a combination of (H,i) and (L,iii) implies (L,ii). ■

Proof of Lemma 2. Given (12), we derive

$$\begin{aligned} & x\mathbb{E}[W|e_1] + (1-x)\mathbb{E}[W|e_2] \\ =\ & x\frac{x\lambda}{x\lambda + y(1-\lambda)}w_H + x\frac{y(1-\lambda)}{x\lambda + y(1-\lambda)}w_L \\ & + (1-x)\frac{(1-x)\lambda}{(1-x)\lambda + (1-y)(1-\lambda)}w_H + (1-x)\frac{(1-y)(1-\lambda)}{(1-x)\lambda + (1-y)(1-\lambda)}w_L, \end{aligned} \tag{82}$$

and

$$\begin{aligned} & y\mathbb{E}[W|e_1] + (1-y)\mathbb{E}[W|e_2] \\ =\ & y\frac{x\lambda}{x\lambda + y(1-\lambda)}w_H + y\frac{y(1-\lambda)}{x\lambda + y(1-\lambda)}w_L \\ & + (1-y)\frac{(1-x)\lambda}{(1-x)\lambda + (1-y)(1-\lambda)}w_H + (1-y)\frac{(1-y)(1-\lambda)}{(1-x)\lambda + (1-y)(1-\lambda)}w_L. \end{aligned} \tag{83}$$

By adding the first line of (82) and the first line of (83), we have

$$\begin{aligned} & x\lambda\frac{x\lambda}{x\lambda + y(1-\lambda)}w_H + x\lambda\frac{y(1-\lambda)}{x\lambda + y(1-\lambda)}w_L \\ & + y(1-\lambda)\frac{x\lambda}{x\lambda + y(1-\lambda)}w_H + y(1-\lambda)\frac{y(1-\lambda)}{x\lambda + y(1-\lambda)}w_L \\ =\ & x\lambda w_H + y(1-\lambda)w_L. \end{aligned}$$

Similarly, by adding the second line of (82) and the second line of (83), we have $(1-x)\lambda w_H + (1-y)(1-\lambda)w_L$. ■

C Generalized cost structures

The main role of separability in (14) is twofold. First, it guarantees the binding TC$_L$, allowing the original problem to be reduced to a one-dimensional one. Second, it yields the simple virtual cost in (42), under which single-peakedness follows from a standard monotone hazard rate condition. As long as TC$_L$ remains binding, however, the reformulation of the problem continues to be valid even without separability. Moreover, the corresponding virtual cost can still be characterized under a general cost structure.

The first step is to revisit the single-crossing arguments used in the proof of Proposition 3, where under separability, the cost-reduction multiplier $\alpha(e)$ in (25) is constant. To see why the argument may fail without separability, consider $t \in (1,2)$, in which $\alpha(e)$ is strictly increasing. With a slight abuse of notation, we denote by $TC_H(x,y,e_1,e_2)$ the function required for TC$_H$ in (20) and $HC(x,y,e_1,e_2)$ the function for HC in (21).[31] Then, for a fixed k or λ, the marginal substitution rate between e_1 and e_2 on HC is given by

$$\frac{de_2}{de_1}\bigg|_{HC(x,y,e_1,e_2)=0} = -\frac{x - y\alpha(e_1)}{(1-x)-(1-y)\alpha(e_2)}\frac{c_e(e_1,\theta_H)}{c_e(e_2,\theta_H)}.$$

In addition, consider TC$_H$ in (20) and, while fixing k or λ, the marginal substitution rate between e_1 and e_2 on TC$_H$ results in

$$\frac{de_2}{de_1}\bigg|_{TC_H(x,y,e_1,e_2)=0} = \frac{c_e(e_1,\theta_H)}{c_e(e_2,\theta_H)}.$$

Consider a point $(\widehat{e}_1,\widehat{e}_2)$. Since $\alpha(e)$ is strictly increasing with $\alpha(e)>1$, one can easily find

$$\frac{de_2}{de_1}\bigg|_{HC(\widehat{e}_1,\widehat{e}_2,x,y)=0} - \frac{de_2}{de_1}\bigg|_{TC_H(\widehat{e}_1,\widehat{e}_2,x,y)=0} < 0.$$

The comparison involving the social welfare function is different. For the fixed k or λ, the marginal substitution rate between e_1 and e_2 on SW in (16) is

$$\frac{de_2}{de_1}\bigg|_{SW(x,y,e_1,e_2)=\overline{SW}} = -\frac{\lambda x + (1-\lambda)y\alpha(e_1)}{\lambda(1-x)+(1-\lambda)(1-y)\alpha(e_2)}\frac{c_e(e_1,\theta_H)}{c_e(e_2,\theta_H)}.$$

Now, the comparison yields

$$\frac{de_2}{de_1}\bigg|_{HC(\widehat{e}_1,\widehat{e}_2,x,y)=0} - \frac{de_2}{de_1}\bigg|_{SW(\widehat{e}_1,\widehat{e}_2,x,y)=\overline{SW}} < 0,$$

[31] That is, define $HC(x,y,e_1,e_2) \equiv k-(x-y)D(\mu)\Delta w - [xc(e_1,\theta_H) - yc(e_1,\theta_L)] - [(1-x)c(e_2,\theta_H) - (1-y)c(e_2,\theta_L)]$ and $TC_H(x,y,e_1,e_2) \equiv D(\mu)\Delta w - c(e_1,\theta_H) + c(e_2,\theta_H)$.

where the negative sign holds for a sufficiently large $e_1 - e_2$. Although $(1-x)y < (1-y)x$ for $x > y$, the gap $\alpha(e_1) > \alpha(e_2)$ makes the numerator positive when $e_1 - e_2$ is large, and the denominator is negative since $(1-x) - (1-y)\alpha(e_2) < 0$.[32]

By contrast, if the separability property is satisfied for which α is a constant, we have $\frac{de_2}{de_1}\Big|_{HC(\widehat{e}_1,\widehat{e}_2,x,y)=0} - \frac{de_2}{de_1}\Big|_{SW(\widehat{e}_1,\widehat{e}_2,x,y)=\overline{SW}} > 0$. That is, the single-crossing property between HC and SW, which holds under separability, may fail once $\alpha(e)$ is increasing. The same logic applies to IR_L: Under separability the proof shows that IR_L satisfies the single-crossing property, but it too may fail when $\alpha(e)$ is increasing.[33]

The following lemma identifies sufficient conditions under which each of the single-crossing comparisons used in the proof of Proposition 3 continues to hold once separability is relaxed.

Lemma 5 *The monotonicity of the cost-reduction multiplier $\alpha(e)$ determines the relevant single-crossing properties as follows.*

(i) If $\alpha(e)$ is decreasing, then the single-crossing properties involving HC–SW and HC–IR_L continue to hold.

(ii) If $\alpha(e)$ is increasing, then the single-crossing property involving HC–TC_H continues to hold.

The lemma clarifies the role of separability. By making $\alpha(e)$ constant, separability guarantees all three single-crossing comparisons simultaneously. When separability is violated, however, the comparisons need no longer move together. Nevertheless, the failure of separability does not necessarily imply the failure of the binding condition; separability is only a sufficient condition for the property. For instance, in Example 1 in (24), when $t > 2$, $\alpha(e) = \frac{t+2e}{1+e}$ is strictly decreasing. Hence, by Lemma 5 (i), the single-crossing properties involving HC–SW and HC–IR_L continue to hold. Moreover, although the sufficiency in Lemma 5 (ii) is violated, the HC–TC_H comparison also holds for this specification.[34] Consequently, all three single-crossing comparisons remain valid despite the violation of separability. This example illustrates that separability is sufficient, but not necessary. We therefore proceed without separability and take the binding of TC_L as the key condition.

First, we rewrite (47) more formally. Let $SW(x, e_1, \lambda)$ denote the social welfare function explicitly including λ. From the binding TC_L constraint in (45) and HC in (46), define

$$t(x, e_1, \lambda) \equiv D(\mu)\Delta w - c(e_1, \theta_L) \text{ and } h(x, e_1) \equiv x[c(e_1, \theta_L) - c(e_1, \theta_H)].$$

[32]Note that $\frac{de_2}{de_1}\Big|_{HC=0} - \frac{de_2}{de_1}\Big|_{SW=\overline{SW}} = \frac{(1-x)y\alpha(\widehat{e}_1)-(1-y)x\alpha(\widehat{e}_2)}{[(1-x)-(1-y)\alpha(\widehat{e}_2)][\lambda(1-x)+(1-\lambda)(1-y)\alpha(\widehat{e}_2)]}\frac{c_e(\widehat{e}_1,\theta_H)}{c_e(\widehat{e}_2,\theta_H)}$, where $\alpha(e_1) - \alpha(e_2) = \frac{(2-t)(e_1-e_2)}{(1+e_1)(1+e_2)} > 0$.

[33]Similar to $TC_H(x, y, e_1, e_2)$, we denote by $IR_L(x, y, e_1, e_2)$ the function for IR_L in (22). If separability holds, then one can show $\frac{de_2}{de_1}\Big|_{HC(x,y,\widehat{e}_1,\widehat{e}_2)=0} - \frac{de_2}{de_1}\Big|_{IR_L(x,y,\widehat{e}_1,\widehat{e}_2)=0} > 0$.

[34]See the part following the proof of Lemma 5 in Online Appendix D.

Then, (47) becomes

$$\max_{x\in[0,1],e_1\in\mathbb{R}_+} SW(x,e_1,\lambda)$$
$$\text{subject to } t(x,e_1,\lambda)=0, \quad k=h(x,e_1).$$

By eliminating one of the three variables, one can obtain an induced relationship between the remaining two variables, namely between x and e_1, x and k, or e_1 and k. For example, eliminating k yields

$$t\big(x,e_1,G(h(x,e_1))\big)=0.$$

Although this equation directly characterizes the relationship between x and e_1, its qualitative properties are difficult to establish because $h(x,e_1)$ depends on both variables. We instead first derive the strictly increasing relationship between x and k, write its inverse as $k=k(x)$, and then use $t\big(x,e_1,G(k(x))\big)=0$ to obtain the remaining relationships.

To facilitate the analysis, define

$$\phi(e_1)\equiv\frac{1}{c(e_1,\theta_L)-c(e_1,\theta_H)}. \tag{84}$$

Then HC in (46) can be written compactly as $\phi(e_1)=\frac{x}{k}$. The function ϕ plays a useful role in both the proof of the lemma and the subsequent reduction of the problem.

Lemma 6 *Suppose that* TC_L *binds. Then the induced pairwise relationships among* x, e_1, *and* k *have the following properties.*

(i) x *is a strictly increasing implicit function of* k, *and vice versa.*

(ii) e_1 *is an implicit function of* k *(and of* x*), but not a monotone one.*

(iii) x *is not an implicit function of* e_1 *(and neither is* k*).*

What the principal directly chooses are x and e_1, while the cutoff k is induced endogenously through the constraints. Extending Theorem 1, we reformulate the problem as if the principal were choosing k. A reformulation with respect to x is possible, so that using k or x is mathematically equivalent given Lemma 6 (i).[35] Economically, however, the formulation in terms of k is more natural, whereas reformulating the problem in terms of x alone and viewing both k and e_1 as induced objects is less transparent.

Substituting $\phi(e_1)=\frac{x}{k}$ from (84) into the binding TC_L constraint gives

$$\frac{1-G(k)}{1-k\phi(e_1)G(k)}\Delta w=c(e_1,\theta_L), \tag{85}$$

[35] Mathematically, a reduction to a one-dimensional problem in terms of e_1 alone is generally not possible, since the feasibility condition need not define a proper implicit function $x(e_1)$ by Lemma 6 (iii).

where $e_1 = e_1(k)$ by Lemma 6 (ii). Using the same procedure in the proof of Theorem 1, the social welfare in (16) can be written as

$$\widehat{SW}(k) = \mathbb{E}^{\lambda}[W] + [1 - Q(e_1(k))]\,\lambda k - \int_{\underline{\omega}}^{k} \omega\, dG, \tag{86}$$

where recall $Q(e) \equiv \frac{c(e,\theta_L)}{c(e,\theta_L)-c(e,\theta_H)}$ in (15). Differentiating with respect to k gives $SW'(k) = G'(k)\big[\Delta w - \psi(k)\big]$, where the simple virtual cost in (42) is replaced by the generalized virtual cost in (87). Under separability, the cost advantage of high-quality individuals is captured by the constant multiplier α. Without separability, this cost advantage varies with the education level and is captured by $Q(e)$, yielding the generalized virtual cost.

$$\psi(\omega) = Q(e_1(\omega))\omega + \left[\frac{d\big(\omega Q(e_1(\omega))\big)}{d\omega} - 1\right]\frac{G(\omega)}{g(\omega)}, \tag{87}$$

where $e_1 = e_1(\omega)$ is from (85).

By Theorem 2, under separability, a standard monotone hazard rate condition is sufficient to guarantee global single-peakedness of the reduced objective. Without separability, the virtual cost becomes endogenous through the education-dependent term $Q(e)$, making a comparably simple global condition difficult to obtain. Nevertheless, a local condition suffices for characterization: It is enough that the generalized virtual cost be increasing at every critical point of the reduced objective. Accordingly, Assumption 2 is replaced by $\Delta w \in \big(Q(\bar{e}_1)\underline{\omega},\ Q(\bar{e}_1)\overline{\omega}\big)$, where recall that $\bar{e}_1$ is the higher education level under the fully informative signal.

Theorem 3 *Suppose $\Delta w \in \big(Q(\bar{e}_1)\underline{\omega}, Q(\bar{e}_1)\overline{\omega}\big)$ and that TC_L binds. If, for every k^* satisfying $k^* = \psi^{-1}(\Delta w)$, we have*

$$\psi'(k^*) > 0,$$

then there exists a unique $k^* \in (\underline{\omega}, k_I)$ satisfying

$$k^* = \psi^{-1}(\Delta w),$$

which characterizes the optimal education mechanism Γ, where $x^* \in (0,1)$, $y^* = 0$, and $e_1^* > e_2^* = 0$, with (x^*, e_1^*) being the unique solution to the system of simultaneous equations in (50). Moreover, the partially informative mechanism strictly dominates the fully informative mechanism.

Unlike the separable case, the characterization in Theorem 3 does not by itself imply that the partially informative education mechanism dominates the fully uninformative benchmark. Under separability, this dominance result follows automatically without any additional restrictions. The following corollary provides a sufficient condition under which the partially informative education mechanism strictly dominates the fully uninformative education system.

Corollary 1 *Suppose the conditions of Theorem 3. Then if $Q(e_U) > \lambda_U Q(e_1(k_U))$ whenever $k_U \leq k_I$, the fully uninformative education system is strictly dominated by the partially informative education mechanism.*

We revisit Example 1 (24). As discussed above, when $t > 2$, the separability property is violated, but the single-crossing condition continues to hold. When $t \in (1, 2)$, even the single-crossing condition may fail. Nevertheless, whenever TC_L remains binding, the example satisfies the local condition in Theorem 3 (see Online Appendix D). Hence, the optimal education mechanism continues to be characterized by the unique threshold satisfying $k^* = \psi^{-1}(\Delta w)$.

D Additional Results

D.1 Two-education-level representation without separability

Corollary 2 *The original signal structure*

$$\pi : \Omega \to \Delta(\mathbb{R}_+)$$

is payoff equivalent to a two-education-level signal structure.

Proof. By the preceding proposition, there exists $T^* \in (0, 1)$ such that the set of wage-matching pairs is

$$\mathcal{L} = \{(x, y) \in [0, 1]^2 : T(x, y) = T^*\}.$$

It remains to select a pair in $\mathcal{L}$ and two education levels that reproduce the original expected cost moments. Let

$$C_H \equiv \int \pi_H(e) c(e, \theta_H)\, de, \qquad C_L \equiv \int \pi_L(e) c(e, \theta_L)\, de,$$

and write

$$c_H(e) \equiv c(e, \theta_H), \qquad c_L(e) \equiv c(e, \theta_L).$$

Since c_H is continuous, strictly increasing, satisfies $c_H(0) = 0$, and is unbounded, it maps $[0, \infty)$ onto $[0, \infty)$. Hence, c_H^{-1} is defined on all of $[0, \infty)$. Define

$$h(u) \equiv c_L\big(c_H^{-1}(u)\big).$$

Because c_L has the same monotonicity and boundary properties, h is continuous, strictly increasing, satisfies $h(0) = 0$, and is unbounded.

Suppose first that $C_H = 0$. Since $T^* \in (0, 1)$, there exists $y_0 \in (0, 1)$ satisfying $T(0, y_0) = T^*$. Set $v = 0$ and choose $u = h^{-1}\left(\frac{C_L}{y_0}\right)$. Then

$$0 \cdot u + (1 - 0)v = 0 = C_H \text{ and } y_0 h(u) + (1 - y_0) h(v) = y_0 \frac{C_L}{y_0} = C_L.$$

Thus, both cost moments are matched. Henceforth, suppose $C_H > 0$.

Now, define

$$x_1 \equiv \frac{T^*(1-\lambda)}{1-\lambda T^*}.$$

Then $x_1 \in (0,1)$ and $T(x_1, 1) = T^*$, so $(x_1, 1) \in \mathcal{L}$. We first show that every low-type cost moment in $\left[0, h\left(\frac{C_H}{x_1}\right)\right]$ can be matched. At the wage-matching pair $(x_1, 1)$, choose $u \in [0, C_H/x_1]$ and set $v = \frac{C_H - x_1 u}{1-x_1}$. Then $u, v \geq 0$ and

$$x_1 u + (1-x_1)v = C_H.$$

Because $y = 1$, the induced low-type cost moment is $h(u)$. As u ranges over $[0, C_H/x_1]$, continuity and strict monotonicity of h imply that $h(u)$ ranges over $\left[0, h\left(\frac{C_H}{x_1}\right)\right]$.

It remains to match values of C_L above this interval. For each $x \in (0, x_1)$, there exists a unique $y(x) \in (x, 1)$ such that $T(x, y(x)) = T^*$. Indeed, $T(x,x) = 1 > T^*$ and $T(x,1) < T^*$ for $x < x_1$, and $T(x,y)$ is continuous in y. Moreover, on the region $y > x$, $\frac{\partial T}{\partial y} < 0$, so the solution is unique. The implicit function theorem also implies that $y(x)$ is continuous on $(0, x_1)$.

Along this branch, set $u = \frac{C_H}{x}$ and $v = 0$. Then

$$xu + (1-x)v = C_H.$$

The corresponding low-type cost moment is

$$F(x) \equiv y(x)h\left(\frac{C_H}{x}\right) + (1-y(x))h(0) = y(x)h\left(\frac{C_H}{x}\right).$$

As $x \uparrow x_1$, we have $y(x) \to 1$, and therefore $F(x) \to h\left(\frac{C_H}{x_1}\right)$. As $x \downarrow 0$, the equation $T(x, y(x)) = T^*$ implies $y(x) \to y_0$, where $y_0 \in (0,1)$ solves $T(0, y_0) = T^*$. Thus, $y(x)$ is bounded away from zero for all sufficiently small x. Since h is unbounded and $C_H/x \to \infty$, it follows that

$$F(x) = y(x)h\left(\frac{C_H}{x}\right) \to \infty.$$

Since F is continuous on the connected interval $(0, x_1)$, its image is an interval. Together with the two endpoint limits, this implies that its image contains $\left[h\left(\frac{C_H}{x_1}\right), \infty\right)$.

Combining the two constructions, every $C_L \geq 0$ can be matched. Therefore, there exist $(\widehat{x}, \widehat{y}) \in \mathcal{L}$ and $u, v \geq 0$ such that

$$\widehat{x}u + (1-\widehat{x})v = C_H \text{ and } \widehat{y}h(u) + (1-\widehat{y})h(v) = C_L.$$

Set $\widehat{e}_1 = c_H^{-1}(u)$ and $\widehat{e}_2 = c_H^{-1}(v)$. Then $c(\widehat{e}_1, \theta_H) = u$ and $c(\widehat{e}_2, \theta_H) = v$, and, by the definition of h,

$$c(\widehat{e}_1, \theta_L) = h(u) \text{ and } c(\widehat{e}_2, \theta_L) = h(v).$$

Hence, the two cost-matching equations (74) and (76) hold. Since $(\widehat{x}, \widehat{y}) \in \mathcal{L}$, the two wage-matching equations (73) and (75) also hold. Therefore, the two-education-level signal structure assigning $\widehat{e}_1$ and $\widehat{e}_2$ with probabilities $(\widehat{x}, \widehat{y})$ is payoff equivalent to the original signal structure. If $u \neq v$, then $\widehat{e}_1 \neq \widehat{e}_2$. If $u = v$, the representation collapses to a single education level, which is a degenerate two-level representation. ■

D.2 Proofs for Online Appendix C

Proof of Lemma 5. The same sign as in (61) can also be obtained when $\alpha(e)$ in (25) is decreasing in e. Since $e_1 > e_2$ by Lemma 3, decreasing α implies $\frac{c_e(e_1,\theta_L)}{c_e(e_1,\theta_H)} \leq \frac{c_e(e_2,\theta_L)}{c_e(e_2,\theta_H)}$. Hence, $c_e(e_1, \theta_L)c_e(e_2, \theta_H) \leq c_e(e_1, \theta_H)c_e(e_2, \theta_L)$. Together with $x(1 - y) > (1 - x)y$, this implies

$$x(1 - y)c_e(e_1, \theta_H)c_e(e_2, \theta_L) > (1 - x)yc_e(e_1, \theta_L)c_e(e_2, \theta_H).$$

Hence the numerator in (61) is strictly positive. Since $(1-x)c_e(e_2, \theta_H)-(1-y)c_e(e_2, \theta_L) < 0$, the expression in (61) is strictly negative. Similarly, as in the proof of Proposition 3, the sign for IR_L holds.

The same sign as in (62) can also be obtained more generally when $\alpha(e)$ in (25) is increasing in e. Since $e_1 > e_2$ by Lemma 3, increasing α implies $c_e(e_1, \theta_L)c_e(e_2, \theta_H) \geq c_e(e_1, \theta_H)c_e(e_2, \theta_L)$. Therefore,

$$yc_e(e_1, \theta_L)c_e(e_2, \theta_H) - yc_e(e_1, \theta_H)c_e(e_2, \theta_L) \geq 0.$$

Substituting into the numerator yields

$$-c_e(e_1, \theta_H)\big[c_e(e_2, \theta_H) - c_e(e_2, \theta_L)\big] + y\Big[c_e(e_1, \theta_L)c_e(e_2, \theta_H) - c_e(e_1, \theta_H)c_e(e_2, \theta_L)\Big].$$

The first term is strictly positive since $c_e(e_2, \theta_H) - c_e(e_2, \theta_L) < 0$, and the second term is nonnegative. Hence the numerator in (62) is strictly positive. Since $(1 - x)c_e(e_2, \theta_H) - (1 - y)c_e(e_2, \theta_L) < 0$, the expression in (62) remains negative. ■

Consider Example 1 (24). For $t > 2$, all the single-crossing conditions hold, despite the violation of separability. The remaining comparison is HC–TC_H. The key term is

$$c_e(e_1, \theta_L)c_e(e_2, \theta_H) - c_e(e_1, \theta_H)c_e(e_2, \theta_L).$$

Substituting $c(e, \theta_H) = e + \frac{e^2}{2}$ and $c(e, \theta_L) = te + e^2$ yields

$$c_e(e_1, \theta_L)c_e(e_2, \theta_H) - c_e(e_1, \theta_H)c_e(e_2, \theta_L) = (t - 2)(e_2 - e_1) < 0,$$

where the negative sign follows from $e_1 > e_2$ and $t > 2$. Now consider the numerator of the derivative in the HC–TC_H comparison:

$$-c_e(e_1, \theta_H)\big[c_e(e_2, \theta_H) - c_e(e_2, \theta_L)\big] + y[c_e(e_1, \theta_L)c_e(e_2, \theta_H) - c_e(e_1, \theta_H)c_e(e_2, \theta_L)],$$

which becomes

$$\begin{aligned}(1+e_1)(t-1+e_2)+y(t-2)(e_2-e_1) &\geq (1+e_1)(t-1+e_2)-(t-2)(e_1-e_2)\\ &=(t-1)+e_1+(1+e_1)e_2+(t-2)e_2>0,\end{aligned}$$

where $y \leq 1$. Since the denominator is negative, the derivative is negative. Thus, the HC–TC$_H$ comparison also holds.

Proof of Lemma 6. *Part (i).* Since $\phi'(e_1)<0$, we can rewrite HC as $e_1=\phi^{-1}\left(\frac{x}{k}\right)$. Substituting this into the binding TC$_L$ condition gives

$$\Psi(x,k)\equiv\frac{1-G(k)}{1-xG(k)}\Delta w-c\left(\phi^{-1}\left(\frac{x}{k}\right),\theta_L\right)=0.$$

Differentiating with respect to x yields

$$\Psi_x(x,k)=\frac{(1-G(k))G(k)}{(1-xG(k))^2}\Delta w-\frac{1}{k\phi'\left(\phi^{-1}\left(\frac{x}{k}\right)\right)}c_e\left(\phi^{-1}\left(\frac{x}{k}\right),\theta_L\right)>0,$$

because $\phi'<0$ and $c_e(\cdot,\theta_L)>0$. Similarly,

$$\Psi_k(x,k)=-\frac{(1-x)g(k)}{(1-xG(k))^2}\Delta w+\frac{x}{k^2\phi'\left(\phi^{-1}\left(\frac{x}{k}\right)\right)}c_e\left(\phi^{-1}\left(\frac{x}{k}\right),\theta_L\right)<0.$$

Therefore, by the implicit function theorem,

$$\frac{dx}{dk}=-\frac{\Psi_k(x,k)}{\Psi_x(x,k)}>0.$$

Hence x is a strictly increasing implicit function of k, and vice versa.

Part (ii). By Part (i), we may write the induced cutoff as $k=k(x)$ with $k'(x)>0$. Substituting $\lambda=G(k(x))$ into the binding TC$_L$ condition gives

$$\frac{1-G(k(x))}{1-xG(k(x))}\Delta w=c(e_1,\theta_L).$$

Since the left-hand side is continuous in x and $c(\cdot,\theta_L)$ is continuous and strictly increasing, this equation defines e_1 as a continuous implicit function of x, and hence of k by Part (i). As $x\to\underline{x}$, Part (i) implies $k(x)\to\underline{\omega}$ and hence $G(k(x))\to 0$. Therefore,

$$\frac{1-G(k(x))}{1-xG(k(x))}\Delta w\to\Delta w.$$

Similarly, as $x\to 1$,

$$\frac{1-G(k(x))}{1-xG(k(x))}\Delta w\to\Delta w.$$

For every interior $x\in(\underline{x},1)$ with $G(k(x))>0$, $\frac{1-G(k(x))}{1-xG(k(x))}<1$, because $x<1$. Hence, $\frac{1-G(k(x))}{1-xG(k(x))}\Delta w<\Delta w$. Since $c(e_1,\theta_L)$ is strictly increasing in e_1, $e_1(x)$ converges to $c_L^{-1}(\Delta w)$

at both endpoints but is strictly below $c_L^{-1}(\Delta w)$ at every interior point. Therefore, e_1 is an implicit function of x, and hence of k, but it is not monotone.

Part (iii). By Part (ii), $e_1(x)$ is continuous, converges to the same value at both endpoints, and takes a strictly lower value in the interior. Fix any $x_0 \in (\underline{x}, 1)$ and choose $\bar{e}$ such that

$$e_1(x_0) < \bar{e} < c_L^{-1}(\Delta w).$$

Then, by the intermediate value theorem, there exist $x' \in (\underline{x}, x_0)$ and $x'' \in (x_0, 1)$ such that

$$e_1(x') = e_1(x'') = \bar{e}.$$

Thus, the inverse relationship from e_1 to x is not single-valued, so x is not an implicit function of e_1 alone. Since x and k are in one-to-one correspondence by Part (i), the corresponding values $k(x')$ and $k(x'')$ are distinct. Therefore, the inverse relationship from e_1 to k is also not single-valued. ■

Proof of Theorem 3. The proof consists of two parts, the existence and the dominance.

Part 1. Let k^* be an interior critical point. Since $SW'(k^*) = 0$, we have

$$\Delta w = \psi(k^*).$$

Differentiating $SW'(k) = g(k)[\Delta w - \psi(k)]$ gives

$$SW''(k) = g'(k)[\Delta w - \psi(k)] - g(k)\psi'(k).$$

Evaluating at k^*, $SW''(k^*) = -g(k^*)\psi'(k^*)$. Since $g(k^*) > 0$ and $\psi'(k^*) > 0$, $SW''(k^*) < 0$. Hence every interior critical point is a strict local maximum. It remains to show uniqueness. Suppose, toward a contradiction, that there are two distinct interior critical points

$$k_1 < k_2.$$

Since both are strict local maxima, there exists $\varepsilon > 0$ small enough such that

$$SW'(k) > 0 \quad \text{for } k \in (k_1 - \varepsilon, k_1) \text{ and } SW'(k) < 0 \quad \text{for } k \in (k_1, k_1 + \varepsilon),$$

and

$$SW'(k) > 0 \quad \text{for } k \in (k_2 - \varepsilon, k_2) \text{ and } SW'(k) < 0 \quad \text{for } k \in (k_2, k_2 + \varepsilon).$$

In particular, for some points close to k_1 and k_2,

$$SW'(k) < 0 \quad \text{to the right of } k_1 \text{ while } SW'(k) > 0 \quad \text{to the left of } k_2.$$

By continuity of SW', there must exist some $\widetilde{k} \in (k_1, k_2)$ such that $SW'(\widetilde{k}) = 0$ and SW' changes sign from negative to positive at $\widetilde{k}$. Hence $\widetilde{k}$ is an interior local minimum. But the first part of the proof shows that every interior critical point is a strict local maximum. This contradiction proves uniqueness.

Part 2. For this proof, denote

$$\Delta c(e) \equiv c(e, \theta_L) - c(e, \theta_H). \tag{88}$$

Then $Q(e)$ in (15) can be rewritten as $Q(e) = \frac{c(e,\theta_L)}{\Delta c(e)}$. In the generalized fully informative benchmark of (29), $e_1(k_I) = \bar{e}_1$ and

$$k_I = \Delta c(\bar{e}_1) \text{ and } \Delta w = c(\bar{e}_1, \theta_L) = Q(\bar{e}_1)\Delta c(\bar{e}_1) = Q(\bar{e}_1)k_I.$$

Hence, by (87),

$$\psi(k_I) = \Delta w + \left[\frac{d}{dk}\big(kQ(e_1(k))\big)\bigg|_{k=k_I} - 1 \right] \frac{G(k_I)}{g(k_I)}.$$

Since $G(k_I)/g(k_I) > 0$,

$$\psi(k_I) > \Delta w \text{ is equivalent to } \frac{d}{dk}\big(kQ(e_1(k))\big)\bigg|_{k=k_I} > 1. \tag{89}$$

Expanding the derivative gives

$$\frac{d}{dk}\big(kQ(e_1(k))\big)\bigg|_{k=k_I} = Q(\bar{e}_1) + k_I Q'(\bar{e}_1)e_1'(k_I).$$

It remains to show that this expression is greater than one. Differentiating the binding TC_L and HC conditions around k_I yields

$$e_1'(k_I) = \frac{1}{\Delta c'(\bar{e}_1) + \Delta c(\bar{e}_1) c_e(\bar{e}_1, \theta_L) \frac{1-G(k_I)}{G(k_I)c(\bar{e}_1,\theta_L)}}.$$

Since $c_e(\bar{e}_1, \theta_L) > 0$, $G(k_I) \in (0, 1)$, and $c(\bar{e}_1, \theta_L) > 0$, we have $e_1'(k_I) < \frac{1}{\Delta c'(\bar{e}_1)}$. If $Q'(\bar{e}_1) \geq 0$, then

$$Q(\bar{e}_1) + k_I Q'(\bar{e}_1)e_1'(k_I) > 1,$$

because $Q(\bar{e}_1) > 1$. If $Q'(\bar{e}_1) < 0$, then

$$Q(\bar{e}_1) + k_I Q'(\bar{e}_1)e_1'(k_I) > Q(\bar{e}_1) + \frac{\Delta c(\bar{e}_1)Q'(\bar{e}_1)}{\Delta c'(\bar{e}_1)}.$$

The right-hand side is greater than 1 if and only if $\big[Q(\bar{e}_1) - 1\big]\Delta c'(\bar{e}_1) + \Delta c(\bar{e}_1)Q'(\bar{e}_1) > 0$, which holds given $c_e(\bar{e}_1, \theta_H) > 0$. To see this, since $Q(e) = \frac{c(e,\theta_L)}{\Delta c(e)}$, we have $c(e, \theta_H) = c(e, \theta_L) - \Delta c(e) = \big[Q(e) - 1\big]\Delta c(e)$. Differentiating with respect to e gives

$$c_e(e, \theta_H) = \big[Q(e) - 1\big]\Delta c'(e) + \Delta c(e)Q'(e).$$

Thus $\big[Q(\bar{e}_1) - 1\big]\Delta c'(\bar{e}_1) + \Delta c(\bar{e}_1)Q'(\bar{e}_1) > 0$ is equivalent to $c_e(\bar{e}_1, \theta_H) > 0$. Hence,

$$Q(\bar{e}_1) + k_I Q'(\bar{e}_1)e_1'(k_I) > 1.$$

Combining the two cases, $\frac{d}{dk}\big(kQ(e_1(k))\big)\Big|_{k=k_I} > 1$ and by (89), $\psi(k_I) > \Delta w$. Finally,

$$\frac{d\widehat{SW}(k)}{dk}\bigg|_{k=k_I} = g(k_I)\big[\Delta w - \psi(k_I)\big] < 0,$$

which proves the dominance. ■

Proof of Corollary 1. Consider the generalized fully uninformative benchmark, where $k_U \equiv c(e_U, \theta_L) - c(e_U, \theta_H)$. Using $Q(e)$ in (15), we have

$$c(e_U, \theta_L) = k_U Q(e_U).$$

Let $e_1(k)$ denote the education level induced by the binding TC_L and HC at cutoff k. We consider two cases.

Case 1. Suppose $k_U \leq k_I$. Since the principal can implement any $k \in [\underline{\omega}, k_I]$, the partially informative mechanism can implement $k = k_U$. Using the generalized reduced welfare expression in (86),

$$\widehat{SW}(k_U) = \mathbb{E}^{\lambda_U}[W] - \lambda_U k_U Q(e_1(k_U)) + \int_{\underline{\omega}}^{k_U} G(\omega)d\omega.$$

By contrast, the generalized fully uninformative benchmark yields

$$\widehat{SW}_U(k_U) = \mathbb{E}^{\lambda_U}[W] - k_U Q(e_U) + \int_{\underline{\omega}}^{k_U} G(\omega)d\omega.$$

Hence, $\widehat{SW}(k_U) - \widehat{SW}_U(k_U) = k_U\left[Q(e_U) - \lambda_U Q(e_1(k_U))\right]$, which results in

$$\widehat{SW}(k_U) > \widehat{SW}_U(k_U) \text{ whenever } Q(e_U) > \lambda_U Q(e_1(k_U)).$$

Since k^* maximizes $\widehat{SW}$, $\widehat{SW}(k^*) \geq \widehat{SW}(k_U) > \widehat{SW}_U(k_U)$.

Case 2. Suppose $k_U > k_I$. Since $\Delta c(e)$ in (88) is strictly increasing and $k_U = \Delta c(e_U), k_I = \Delta c(\overline{e}_1)$, we have $e_U > \overline{e}_1$. The remaining comparison is identical to the separable proof, except that $k_U\alpha/(\alpha-1)$ is replaced by $c(e_U, \theta_L) = k_U Q(e_U)$. In the generalized fully informative benchmark,

$$SW_I = w_L + \int_{\underline{\omega}}^{k_I} G(\omega)d\omega.$$

Hence,

$$\begin{aligned}
SW_I - \widehat{SW}_U(k_U) =& -\lambda_U \Delta w + c(e_U, \theta_L) - \int_{k_I}^{k_U} G(\omega)d\omega \\
>& -\lambda_U \Delta w + c(e_U, \theta_L) - (k_U - k_I)G(k_U) \\
=& \left[-\Delta w + k_I - k_U\right] G(k_U) + c(e_U, \theta_L) \\
=& \left[-c(\overline{e}_1, \theta_H) - c(e_U, \theta_L) + c(e_U, \theta_H)\right] G(k_U) + c(e_U, \theta_L) \\
=& \, G(k_U)\left[c(e_U, \theta_H) - c(\overline{e}_1, \theta_H)\right] + [1 - G(k_U)]c(e_U, \theta_L),
\end{aligned}$$

where the second-to-last equality follows from $\Delta w = c(\bar{e}_1, \theta_L)$ and $k_I = \Delta w - c(\bar{e}_1, \theta_H)$, and the last equality from $k_U = \Delta c(e_U)$. Since $e_U > \bar{e}_1$, $c(e_U, \theta_H) > c(\bar{e}_1, \theta_H)$. Thus,

$$SW_I > \widehat{SW}_U(k_U).$$

Further, by Theorem 3, the partially informative mechanism strictly dominates the fully informative benchmark. Together,

$$\widehat{SW}(k^*) > SW_I > \widehat{SW}_U(k_U).$$

Combining the two cases establishes the result. ■

Consider Example 1 in (24). We show that even when the single-crossing condition may fail for $t \in (1, 2)$, given the binding TC_L, the example satisfies the local condition in Theorem 3. For uniform G on $[0, 1]$, the binding TC_L in (85) becomes

$$c_L(e_1) - k^2 Q(e_1) = (1 - k)\Delta w. \tag{90}$$

Differentiating with respect to k gives

$$e_1'(k) = \frac{2kQ(e_1) - \Delta w}{c_L'(e_1) - k^2 Q'(e_1)}. \tag{91}$$

At any critical point k^* of SW, the first-order condition is $\Delta w = \psi(k^*)$. For uniform G, writing

$$P(k) \equiv kQ(e_1(k)),$$

we have $\psi(k) = P(k) + k[P'(k) - 1]$. Thus, at k^*, $\Delta w = 2k^* Q(e^*) + (k^*)^2 Q'(e^*) e_1'(k^*) - k^*$, where $e^* \equiv e_1(k^*)$. Substituting this into (91) yields

$$e_1'(k^*)[c_L'(e^*) - (k^*)^2 Q'(e^*)] = k^* - (k^*)^2 Q'(e^*) e_1'(k^*).$$

Hence

$$c_L'(e^*) e_1'(k^*) = k^*. \tag{92}$$

For the quadratic specification, $c_L'(e) = t + 2e$, so

$$e_1'(k^*) = \frac{k^*}{t + 2e^*}. \tag{93}$$

Differentiating (92) gives $c_L''(e^*)[e_1'(k^*)]^2 + c_L'(e^*) e_1''(k^*) = 1$. Since $c_L''(e) = 2$,

$$e_1''(k^*) = \frac{1 - 2[e_1'(k^*)]^2}{t + 2e^*}. \tag{94}$$

Using the chain rule, $P'(k) = Q + kQ'e_1'$ and $P''(k) = 2Q'e_1' + kQ''(e_1')^2 + kQ'e_1''$, which yields

$$\psi'(k) = 2P'(k) - 1 + kP''(k) = 2Q + 4kQ'e_1' - 1 + k^2\left[Q''(e_1')^2 + Q'e_1''\right]. \tag{95}$$

Case 1: $1 < t < 2$. Here $Q'(e) < 0$ and $Q''(e) > 0$. Substituting the expressions for $e_1'(k^*)$ and $e_1''(k^*)$ into (95) yields

$$\psi'(k^*) = 2Q(e^*) - 1 - \frac{10(2-t)(k^*)^2}{(2t-2+e^*)^2(t+2e^*)} + \text{positive terms.}$$

Since feasibility of (90) requires $(k^*)^2 < c_L(e^*) - c_H(e^*) = \frac{e^*(2t-2+e^*)}{2}$, the negative term is bounded below by $-\frac{5(2-t)e^*}{(2t-2+e^*)(t+2e^*)}$. Moreover, $2Q(e^*) - 1 = \frac{2t+3e^*+2}{2t-2+e^*}$. Hence it suffices to show

$$\frac{2t+3e^*+2}{2t-2+e^*} > \frac{5(2-t)e^*}{(2t-2+e^*)(t+2e^*)}.$$

Equivalently, $(t+2e^*)(2t+3e^*+2) > 5(2-t)e^*$. But

$$(t+2e^*)(2t+3e^*+2) - 5(2-t)e^* = 2t(t+1) + (7t-4)e^* + 6(e^*)^2 > 0$$

for all $t > 1$ and $e^* > 0$. Hence $\psi'(k^*) > 0$.

Case 2: $t \geq 2$. Here $Q'(e) \geq 0$ and $Q''(e) \leq 0$. By (92), $e_1'(k^*) = \frac{k^*}{c_L'(e^*)} > 0$. Using (93),

$$e_1'(k^*) = \frac{k^*}{t+2e^*} < \frac{1}{t} \leq \frac{1}{2},$$

so (94) implies $e_1''(k^*) > 0$. Consequently, in (95),

$$4k^*Q'(e^*)e_1'(k^*) \geq 0 \text{ and } (k^*)^2Q'(e^*)e_1''(k^*) \geq 0,$$

and the only potentially negative term is $(k^*)^2Q''(e^*)[e_1'(k^*)]^2$. Using again $(k^*)^2 < \frac{e^*(2t-2+e^*)}{2}$, we obtain

$$\left|(k^*)^2Q''(e^*)[e_1'(k^*)]^2\right| < \frac{t-2}{(2t-2+e^*)(t+2e^*)}.$$

On the other hand,

$$2Q(e^*) - 1 = \frac{2t+3e^*+2}{2t-2+e^*}.$$

Since $(t+2e^*)(2t+3e^*+2) = 2t(t+1) + (7t-4)e^* + 6(e^*)^2 > t-2$, we have

$$2Q(e^*) - 1 > \frac{t-2}{(2t-2+e^*)(t+2e^*)}.$$

Thus the positive term $2Q(e^*)-1$ dominates the only potentially negative term, so $\psi'(k^*) > 0$.