To Trust or to Monitor: A Dynamic Analysis

Fali Huang*

School of Economics
Singapore Management University

December 28, 2009

Abstract

In a principal–agent framework, principals can mitigate moral hazard problems using extrinsic incentives, such as monitoring, and relying on agents' intrinsic trustworthiness. This paper suggests that principals, however, may not benefit from hiring more trustworthy agents due to competition and endogeneity of trustworthiness, though they always benefit from cheaper monitoring methods; the opposite is true for agents, whose incomes are higher when they are more trustworthy but lower when monitoring is less costly. Such conflict of interests implies that societies giving employees lower weights in policy determination tend to have higher monitoring intensities and lower employee trustworthiness, which may drive up the overall governance cost.

Keywords Monitoring · Trustworthiness · Trust · Screening · Economic Governance

JEL Classification Numbers D2 · J5 · L2 · M5 · Z13.

1 Introduction

All societies have to deal with moral hazard problems, but each society resolves such problems in different ways; some rely more on the intrinsic trustworthiness of agents, while others depend on heavy use of extrinsic governance and monitoring rules. In the late medieval period, for example, agency relations among Maghribi traders were characterized by the prevalence of trust: “Despite the many opportunities for agents to cheat, only a handful of documents contain allegations of misconduct ...” (Greif 1993). This is, however, not the case in Italy, “where allegations of misconduct are well-reflected in the historical records” (Greif 1993). In current times, labor-management

---

* 90 Stamford Road, Singapore 178903. Email: flhuang@smu.edu.sg. The author gratefully acknowledges the financial support of SMU Research Grant 03-C208-SMU-018.
relations in Japan depend on a high level of trust, while “[t]he twentieth-century American system of industrial labor relations, with its periodic massive layoffs, book-length contracts, and bureaucratic, rule-bound personal interactions, would seem the very model of low-trust social relations” (Fukuyama 1995, p. 218). In a sample of developed economies as shown in Figure 1, the monitoring intensities (measured by the average supervisor-employee ratios) in the manufacturing sector from the 1970s to 1990s are the highest in the UK, US and Canada, with an average over two times that of the rest (Gordon 1994, Vernon 2001).

![Figure 1: The Time Trend of Monitoring Intensity in Manufacturing](image)

Why do societies differ in usage of trust and monitoring? Motivated by this question, this paper examines the decision of principals choosing between monitoring and relying on agent trustworthiness, the agent’s decision of whether or not to invest in trustworthiness and how much, and finally, the government’s policy choice that affects the relative cost of monitoring.

In this paper, trustworthiness is essentially a trait or skill cultivated in childhood that enables one to resist short-run opportunistic temptations. This is consistent with the extensive experimental evidence produced over the past four decades, which shows that internalized trustworthiness is a common phenomenon, is at least in part learned rather than innate, and varies across individuals (Stout and Blair 2001). In a principal-agent context, an agent with higher trustworthiness has stronger intrinsic discipline against shirking, and hence requires less extrinsic incentives to make the required effort. In other words, an agent’s perceived level of intrinsic trustworthiness induces corresponding trust from the principal in the sense that it is optimal for the latter to reduce
monitoring accordingly.\footnote{See Huang (2007) for a more thorough treatment of the relationship between trust, trustworthiness, and institutions (including monitoring and law). For other bases of trust such as social norms or altruism in the context of a principal-agent model, see, for example, Rotemberg (1994), Chami and Fullenkamp (2002), and Casadesus-Masanell (2004).} And so there is substitution between monitoring and relying on agent trustworthiness in mitigating moral hazard problems.

Though the total surplus is higher when agents are more trustworthy, principals may not necessarily capture the rent generated by higher trustworthiness due to two important factors: the competition among principals and the endogeneity of agent trustworthiness. As a result, there is a fundamental conflict of interest between principals and agents in substituting monitoring for trust: principals always prefer cheaper monitoring methods than hiring trustworthy agents, while the opposite is true for agents. This is the key insight of this paper and has not been explored in the literature.

The intuition is as follows. Given that hiring more trustworthy agents can reduce monitoring costs, all principals would want to hire the most trustworthy agents. Such competition among principals will then drive up agents’ wages in proportion to their perceived trustworthiness and the cost of monitoring. This implies that the rent from hiring more trustworthy agents disappears in equilibrium, and principals gain only when \textit{all} agents are more trustworthy.\footnote{The logic is similar in spirit to Becker’s (1962) insight on firms’ reluctance to invest in the general training of employees, anticipating that competing firms will steal them away with higher wages. In contrast, firms may be willing to invest in corporate culture (Rob and Zemsky 2002, Kreps 1997) and employee identity (Akerlof and Kranton 2005), in the same way as they are willing to invest in firm-specific human capital.} But this type of gain can not last long because agents have no incentives to make costly investment in trustworthiness only to bring free windfall to principals.\footnote{Considerable time and resources are usually involved in inculcating moral values in children (Shavell 2002).} So any rent captured by a principal from agent trustworthiness is at risk of being bid away by labor market competition and by endogenous investment of agents. In sharp contrast, the reduction of governance cost due to cheaper monitoring methods is immune to both hazards. Therefore, principals are better off with cheaper monitoring methods but not necessarily so with hiring more trustworthy agents. The opposite is true for agents, since their incomes are higher when they are more trustworthy and when monitoring is more costly.

Since principals and agents have conflicting interests in the usage of trust and monitoring, which side has more weight in resource allocation becomes very important in shaping the relative cost of monitoring and trust and hence their actual usage. This yields the following cross-sectional
variation: *societies giving lower weights to the welfare of workers tend to rely more on monitoring and less on trust*. It is indeed consistent with empirical evidence in Figure 2, where the collective labor power (measured by the collective relations laws index in Botero et al. 2004) is negatively correlated with the supervision intensity among fifteen developed economies (Gordon 1994).\(^4\) For example, in the UK, US and Canada, the monitoring intensities are the highest and labor power indicators are the lowest, while the opposite combination appears in Japan and Germany. A discrete version of this result is that individualistic societies tend to rely more on monitoring than group-oriented ones, if workers in individualistic societies enjoy lower weights in resource-allocating decisions; this is consistent with the distinct experiences of the two medieval trader groups mentioned above.\(^5\)

Another implication of the model is that monitoring intensities are likely to increase over time, if the unit cost of using monitoring technologies decreases faster than the costs of bringing up and screening trustworthy agents, which is plausible, at least recently, given that monitoring technologies are easier to standardize and improve upon than technologies of cultivating and screening.

\(^4\) Also see Esping-Anderson (1990) and Rubery and Grimshaw (2003) for similar evidence.

\(^5\) The “social structure of the Maghribi traders’ group was ‘horizontal,’” as traders functioned as agents and merchants at the same time,” while agency relations were organized “vertically” among the Italian traders in that “merchants and agents constitute two distinct subgroups” (Greif 1993). As a result, the Maghribi traders maintained close social ties to reduce the costs of training and screening trustworthy agents, while the Genoese traders adopted new technologies and institutions to reduce monitoring costs.
and such trend should be more acute in societies with weaker labor power such as in the US, UK, and Canada, which is evident in Figure 1. When monitoring becomes cheaper over time, the average trustworthiness declines because parents anticipate that trustworthiness will be less rewarded. Cheaper monitoring technologies, however, may in the end drive up the overall governance cost by crowding out too much trustworthiness and thus forcing society to rely excessively on extrinsic incentives.

The main contribution of this paper is to provide a novel explanation for how the levels of monitoring intensities and intrinsic trustworthiness differ across societies and evolve over time. It is related to a stream of literature investigating how intrinsic motivation can be crowded in or out by various extrinsic incentives, including high-power incentive schemes (Kreps 1997, Rob and Zemsky 2002, Sliwka 2007), public policies (Bar-Gill and Fershtman 2005), legal institutions (Huck 1998, Bohnet, Bruno and Huck 2001, Huang 2007), and explicit monitoring (Frey 1993). The current paper contributes to this literature by endogenizing both the intrinsic motivation and monitoring intensities and by studying their two-way dynamic interactions, instead of the usual one-way crowding-out effects. More importantly, while most of these studies focus on instant or mechanical feedbacks of extrinsic incentives on agents’ trustworthy behaviors, the current paper emphasizes their long-run effects on agent predisposition through rational human capital investment.

This paper also connects with studies on endogenous social preferences and ethical behaviors. Frank (1987) explores whether an individual wants to choose his own utility function that allows others-regarding elements. Güth and Ockenfels (2005) study the endogeneity of moral preferences using the indirect evolutionary approach, which combines individual rational decision making with the evolutionary approach of preference determination. Kaplow and Shavell (2007) examine how a
social planner would inculcate guilt and virtue in individuals to foster social welfare. Using a similar approach of human capital investment as in the current paper, Huang (2007) studies the formation of social trust in the context of prisoner’s dilemmas. On the empirical front, the effects of applicant screening for better work ethic (which may include individual trustworthiness as one dimension) on the monitoring intensity, wages and other performance related outcomes are estimated by Huang and Cappelli (2010); their results are consistent with theoretical predictions of the principal-agent model in this paper.

This paper proceeds as follows. In Section 2, a simple principal–agent model with monitoring and public observation of agent trustworthiness is introduced, where individual trustworthiness is endogenized through parental investment, and the relative cost of monitoring is affected by government policy. This basic model is then extended to costly screening of agent trustworthiness in Section 3. The final section presents conclusions.

2 The Basic Model

2.1 A Principal–Agent Model with Monitoring

A principal hires an agent to complete a project. The outcome is stochastic: if the agent makes the appropriate effort, the outcome is $h > 0$ with probability $p \in [0, 1]$ and 0 with probability $1 - p$; if the agent shirks, the probability of getting $h$ is $q \in (0, 1)$, where $q < p$, and that of getting 0 is $1 - q$. The cost of making effort is $e$, while shirking involves no cost. $hp - e > hq$ is assumed to be true so that making effort $e$ is the socially optimal choice.

There is a continuum of measure one of agents, who are heterogeneous in predisposition to cooperate. An agent has a degree of trustworthiness $\alpha \geq 0$ that measures the amount of guilt he feels if he shirks, whether caught or not by the principal.\(^8\) The cumulative distribution function of trustworthiness among agents is $F(\cdot)$ on $\mathbb{R}^+$. It characterizes the quality of workforce in this economy. Let $\alpha \geq 0$ denote the lowest trustworthiness among agents in the economy. Agents are risk neutral, and there is a liability constraint such that a negative payment is not feasible for

---

\(^8\)Modeling $\alpha$ as an intrinsic benefit of cooperation does not affect the results. Note that such psychological process typically occurs immediately after one shirks or even simultaneously (because shirking is known to hurt the principal on average), which can be well before the performance outcomes are realized and hence independent of the realization of outcomes.
agents.\footnote{Risk averse agents were assumed in an earlier version of the paper, which yields similar results. Relaxing the liability constraint will no doubt strengthen the extrinsic incentives and thus weaken the role of agent trustworthiness; our results are unlikely to be affected, however, as long as the negative payment can not be too large.}

Presumably, $\alpha$ indicates an agent’s cooperative tendency, which by definition is not necessarily equal to the degree of individual trustworthiness; indeed, the same level of $\alpha$ may lead to different levels of trustworthiness in different situations or games, which is formally shown in Huang (2007). In this paper, $\alpha$ is directly called an agent’s trustworthiness for convenience because the game is fixed so that there is a one-to-one relationship between the two. Though in this specific context the role of $\alpha$ would be similar to some other personal attributes such as obeying order or being loyal to one’s employer (Mulligan 1997), it is much more general than them because it also facilitates cooperation in any prisoner’s dilemmas where there is no authority or employment relationship (Huang 2007). In essence, $\alpha$ indicates the amount of self punishment one is capable of imposing for any deviation from the social welfare maximizing standard, which is similar to the moral standard suggested by Shavell (2002) of doing general good. Individuals with a higher $\alpha$ are thus more trustworthy across various situations.

Principals are identical and of measure one. The reservation utility of agents and the alternative return for principals are normalized to zero. To reduce shirking, a principal may screen job candidates and monitor the agent on the job. In this basic model an agent’s trustworthiness is publicly observed. A more general case is studied in the next section, where a principal can obtain a noisy signal of $\alpha$ through a screening process.\footnote{In practice, virtually all employers use some level of applicant screening. It is a fundamental part of the human resources function in most firms as well as the basis of a substantial consulting industry. Employee selection tests are designed to identify not just skills but dispositional attributes affecting performance like conscientiousness or honesty, which is related to $\alpha$ in our context. Conscientiousness has been found to be a reliable and consistent dimension of personality and also relates strongly to job performance across many types of jobs (e.g., Barrick and Mount 1991).}

The monitoring intensity is denoted by $m \in [0,1]$, which equals the probability that an agent who shirks gets caught by the principal. The total monitoring cost is $mk$, where $k$ measures the unit cost of using monitoring technologies such as video cameras in the workplace.\footnote{The linear format of monitoring cost is used to simplify the exposition and would not affect the qualitative results; a convex cost function should strengthen the incentives for principals to hire trustworthy agents and for agents to invest in trustworthiness, but the fundamental conflict of interests between principals and agents in utilizing trustworthiness versus monitoring still exists.} Monitoring is usually imperfect because effort is difficult to measure; for example, video cameras can record
whether an agent is working, but they do not always enable the principal to tell whether the agent is making a conscientious effort or just daydreaming while working.

The payment to an agent has two components: one is the basic wage $b \geq 0$ that is independent of the agent’s performance, and the other is the incentive payment $w \geq 0$, which will be forsaken if shirking is detected by the principal.\textsuperscript{12} The utility of an agent with $\alpha$ is thus $w + b - e$ when he makes effort, and $(1 - m)w + b - \alpha$ if he shirks.

The time line of the game with publicly observed agent trustworthiness is as follows. Principals announce their incentive packages $(m, w, b)$ as functions of the agent’s perceived trustworthiness $\alpha$. Agents then match with principals. After the matching is finished, agents get the basic wage $b$ and choose whether to make the effort or shirk. Principals then monitor agents with intensity $m$, pay $w$ if no shirking is found, and pay nothing if otherwise.\textsuperscript{13}

Note that an agent with a higher $\alpha$ has a larger intrinsic incentive against shirking, which means that, everything else equal, the principal can afford to spend less in providing extrinsic incentives. In this sense, the principal is induced to use more trust and less monitoring, and hence there is a trade off between them. In other words, trusting more means relying more on agent trustworthiness and less on extrinsic incentives in economic governance. One may argue that this is not pure or unconditional trust, but conditional or calculative trust (Williamson 1993). This is of course the fact; however, it is not easy to imagine why a rational individual should have blind trust on anybody who is not trustworthy. In real life, trust is usually exercised discretely, which cannot be purified from some rational judgement or calculative concerns. And more importantly, such a concept of trust that is conditional on the agent’s perceived trustworthiness is still a useful concept in analysis because it presents a distinct source, apart from monitoring and other types of extrinsic incentives, that generates a certain degree of assurance for the principal that a promised level of effort is to be made by the agent. And as to be shown below, it also affects the incentive

---

\textsuperscript{12}Given that the paper’s main focus is the interactions between trust and monitoring, and that an outcome-contingent wage serves the same purpose as monitoring in deterring shirking, the qualitative results will not be affected by allowing the incentive wage to vary across outcomes; see the Appendix for the alternative proof of Proposition 1 when outcome-contingent wages are used. In general, wages that are not contingent on outcomes may also be adopted when outcomes are not verifiable by the agent, or due to multi-tasking concerns (Holmstrom and Milgrom 1991).

\textsuperscript{13}The qualitative results is unlikely to change if alternative combinations of schemes were used. For example, whatever repeated interactions can do to mitigate the moral hazard problem is either type-revealing or imposing extra extrinsic incentives, both of which are already represented by screening and monitoring in a one-period relationship.
structure and competitive equilibrium as well as the overall governance cost in substantial ways.

2.2 The Competitive Equilibrium

The competitive equilibrium is reached in this game when there is no further changing of partners and, once in a match, nobody has incentives to deviate from their decisions. For positive levels of trust and monitoring to happen, the following condition is assumed:

\[
\frac{e^2}{e - \alpha} \leq k < \frac{k^2(p - q)^2}{4e}.
\]

**Proposition 1** In the competitive equilibrium: (I) The optimal incentive package \((m^*, w^*, b^*)\) for \(\alpha \leq e\) includes

\[
m^*(\alpha, k, \alpha) = \left(\sqrt{e - \alpha} - \sqrt{\alpha - \alpha}\right)/\sqrt{k},
\]

\[
w^*(\alpha, k, \alpha) = \sqrt{k}(\sqrt{e - \alpha} + \sqrt{\alpha - \alpha}),
\]

and \(b^* = 0\), while for \(\alpha > e\) the package contains \(w^* = 2\sqrt{k(e - \alpha)}\), \(m^* = 0\) and \(b^* = 0\). Given this incentive scheme, all agents make the effort. (II) The optimal profits are identical across principals:

\[
Q^*(k, \alpha) = hp - 2\sqrt{k(e - \alpha)},
\]

which is decreasing in \(k\), increasing in \(\alpha\), but independent of \(\alpha\). (III) The monitoring cost is zero for \(\alpha > e\), and

\[
m^*(\alpha, k)k = \sqrt{k}(\sqrt{e - \alpha} - \sqrt{\alpha - \alpha})
\]

for \(\alpha \leq e\), which increases in \(k\) but decreases in \(\alpha\) and \(\alpha\).

**Proof.** We solve the game backwards. Given the incentive package \((w, m, b)\), an agent does not shirk if \(w + b - e \geq (1 - m)w + b - \alpha\). This is simplified to the following no-shirking condition

\[
m w_{\text{extrinsic incentive}} + \alpha w_{\text{intrinsic incentive}} \geq e,
\]

where an agent won’t shirk if the sum of extrinsic and intrinsic incentives is larger than the cost of effort. So the optimal monitoring level should satisfy

\[
m^* = \begin{cases} 
(e - \alpha)/w^* & \text{if } \alpha \leq e \\
0 & \text{if } \alpha \geq e
\end{cases}
\]
which is lower when the agent trustworthiness $\alpha$ is higher for the same wage $w^*$; as a result, all principals would like to hire more trustworthy agents to reduce the monitoring cost.

A natural starting point of the analysis is a principal who hires the least trustworthy agent $\alpha$ in the population and thus faces no competition from other principals. If a positive monitoring level is ever chosen, the principal’s objective function is

$$\max_{w,b} hp - w - \frac{k(e - \alpha)}{w} - b.$$  

Note that $b^* = 0$ must hold, since if not, the profit can always be pushed up by reducing $b$ and increasing $w$ to reduce the monitoring intensity $(e - \alpha)/w$. The optimal solution is $w^*(\alpha, k, \alpha) = \sqrt{k(e - \alpha)}$, which is higher than $e$ under assumption (1) so that the participation constraint $w + b - e \geq 0$ is satisfied. Then we get $m^*(\alpha, k, \alpha) = \sqrt{e - \alpha}/\sqrt{k}$ from (6); note that $w^*(\alpha, k, \alpha)$ and $m^*(\alpha, k, \alpha)$ are consistent with (3) and (2), respectively. Given this incentive scheme, the agent makes the effort, and the principal’s optimal profit is

$$Q^*(k, \alpha) \equiv hp - w^*(\alpha, k, \alpha) - m^*(\alpha, k, \alpha)k = hp - 2\sqrt{k(e - \alpha)}.$$  

The positive monitoring is indeed optimal, since $Q^*(k, \alpha) \geq hq$ holds under assumption (1) (which says $k$ is not too large), where $hq$ arises when the principal does not monitor the agent, who shirks, produces $hq$ on average, and receives the reservation wage $0$.

The principal with an agent $\alpha > \underline{\alpha}$ does not have an unconstrained optimization problem as above because he faces competition from other principals whose agents are less trustworthy than $\alpha$. Such competition will force him to make the same profit as $Q^*(k, \alpha)$, which is the profit received by a principal hiring the least trustworthy agent $\alpha$. The reason is obvious. Suppose he makes a profit $Q$ that is higher than $Q^*(k, \alpha)$; then principals with $\alpha$ would be able to bid away the agent $\alpha$ with a slightly higher wage, as long as the wage increase is smaller than $Q - Q^*(k, \alpha) > 0$. Since in the competitive equilibrium there is no partner-changing, it must be true that all principals make the same level of profit that is equal to $Q^*(k, \alpha)$ and all the rents generated by heterogenous trustworthiness of agents go to the agents as higher wages. The equalization of profits $Q^*(k, \alpha) = Q^*(k, \underline{\alpha})$ implies that

$$w^*(\alpha, k, \alpha) + m^*(\alpha, k, \alpha)k = w^*(\underline{\alpha}, k, \underline{\alpha}) + m^*(\underline{\alpha}, k, \underline{\alpha})k = 2\sqrt{k(e - \underline{\alpha})}$$  

must hold for all $\alpha \geq \underline{\alpha}$, which together with (6) determines $w^*(\alpha, k, \alpha)$ and $m^*(\alpha, k, \alpha)$ for $\alpha \leq e$. To see this, after plugging $m = (e - \alpha)/w$ into (7) we get $w + (e - \alpha)/w = 2\sqrt{k(e - \alpha)}$, which leads to
two solutions \( w = \sqrt{k(e - \alpha)} \pm \sqrt{k(\alpha - \alpha)} \), where only the larger solution \( \sqrt{k(e - \alpha)} + \sqrt{k(\alpha - \alpha)} \) satisfies the competition constraint \( w^*(\alpha, k, \alpha) \geq w^*(\alpha, k, \alpha) \). Plugging (3) into \( m = (e - \alpha)/w^* \) we get (2). For agents with \( \alpha > e \), their intrinsic incentive \( \alpha \) alone is high enough to prevent shirking, so the principal would set \( m^* = 0 \) and \( w^* = 2\sqrt{k(e - \alpha)} \).

Given the incentive package \( (m^*, w^*, b^*) \) and the profit level \( Q^*(k, \alpha) \) defined above, neither principals nor agents have incentives to change partners or to deviate from their optimal choices once being matched. So in the competitive equilibrium with homogenous principals and heterogenous agents, all principals receive the same profit \( Q^*(k, \alpha) \) but incur different monitoring costs \( m^*k \), which decrease with agent trustworthiness \( \alpha \), while agents with higher \( \alpha \) receive higher wages and less monitoring. It is obvious to see that the wage \( w^* \) increases in \( k \) while the monitoring cost \( m^*k \) and the profit \( Q^*(k, \alpha) \) both decrease in \( k \), and \( Q^*(k, \alpha) \) is independent of \( \alpha \), though it increases in \( \alpha \).

The underlying intuition of this proposition is as follows. Given an agent’s trustworthiness, principals adjust their monitoring intensities and incentive pays accordingly to save governance costs. Because of perfect competition between principals, the cost saved is transferred to agents as a rent, leaving principals with a profit that they would have made when hiring the least trustworthy agents in the population. In other words, principals do not benefit from hiring agents with higher trustworthiness in equilibrium, though they do gain when the bottom agents are more trustworthy, since \( Q^*(k, \alpha) \) increases in \( \alpha \). In this sense, the bottom-level trustworthiness \( \alpha \) serves as a public good for all principals.

The results for \( \alpha = 0 \) are illustrated in Figure 1. The agent income in this case is

\[
w^*_0(\alpha, k) = \sqrt{ke} + \sqrt{k\alpha},
\]

for \( \alpha \leq e \), which strictly increases with \( \alpha \), and becomes a constant \( 2\sqrt{ke} \) for \( \alpha \geq e \). The principals’ profit

\[
Q^*_0 = hp - 2\sqrt{ke}
\]

is constant with respect to \( \alpha \). The monitoring cost is

\[
m^*_0(\alpha, k)k = \sqrt{ke} - \sqrt{k\alpha},
\]

for \( \alpha \leq e \), which strictly decreases with \( \alpha \), and becomes 0 for \( \alpha \geq e \).
2.3 Inter-Generational Dynamics: Endogenous Trustworthiness

The distribution of $\alpha$ in society is endogenized in this part. Suppose principals and agents live for one period, each raising a child to replace their role. The underlying technologies remain the same over generations, and all agent children are endowed with an identical productive ability to that of their parents. Their intrinsic trustworthiness is 0 at birth, which can be improved by parental investment during childhood to maximize a child’s lifetime income minus the investment cost.

The sequence of events is as follows. In the beginning of each generation $n = 1, ..., +\infty$, the distribution of $\alpha_n$ is realized. Then the above stage game is played, where the competitive equilibrium derived in Proposition 1 prevails. The initial distribution of $\alpha_1$ is non-degenerate and exogenously given on $\mathbb{R}^+$, where the lowest level is denoted by $\underline{\alpha_1} \geq 0$. In each generation $n \geq 1$, an agent with $\alpha_n$ inculcates trustworthiness $\alpha_{n+1}$ in his child, expecting him to get an equilibrium income $w^*(\alpha_{n+1}, k; \alpha_{n+1})$ when the child becomes an adult, where $\alpha_{n+1}$ denotes the lowest trustworthiness in generation $n + 1$. The inculcation cost $C(\alpha_{n+1}; \alpha_n)$ is increasing and convex in $\alpha_{n+1}$, while it decreases in parental trustworthiness $\alpha_n$; that is, $C_1 > 0$, $C_{11} > 0$ and $C_2 < 0$. We also assume $C_{12} < 0$ so that the marginal cost is lower when the parent is more trustworthy. And it costs nothing to retain the initial zero trustworthiness so that $C(0; \cdot) = 0$. Then generation $n + 1$ replaces the old one, and the sequence of events goes on.

Since agent parents can always get a net return of $\sqrt{ke}$ by not investing in their children’s trustworthiness (i.e., $\alpha_{n+1} = 0$ in (8)), the objective function of a parent in generation $n$ is
\[ \max\{R(\alpha_n), \sqrt{k\epsilon}\}, \]

where

\[ R(\alpha_n) \equiv \max_{\alpha_{n+1}} w^*(\alpha_{n+1}, k, \alpha_{n+1}) - C(\alpha_{n+1}; \alpha_n) \quad (11) \]

is the return if he ever invests in \( \alpha_{n+1} \). Note that \( R(\alpha_n) = \sqrt{k\epsilon} \) when \( \alpha_{n+1} = 0 \). In this setting parents are altruistic; they bear the inculcation cost and enjoy a level of utility gain equal to the child’s future income.\(^{14}\) The exact degree of altruism is not essential to our results, however, as long as parents care about their children’s welfare.

Given that the income of the bottom agents, \( \sqrt{k\epsilon} \), is independent of their trustworthiness \( \alpha_{n+1} \) while investing in any positive trustworthiness is costly, \( \alpha_{n+1} = 0 \) must be true. This implies that the lowest trustworthiness of agents in any generation \( n \geq 2 \) would always be 0.\(^{15}\) As a result, principals would always earn a profit \( Q^*_0 \) in (9) that is independent of \( \alpha_{n+1} \), and thus cannot capture any rent generated by agent trustworthiness from the second generation onwards. Similarly, no agent would have \( \alpha_{n+1} > e \), since doing so yields the same income as having \( \alpha_{n+1} = e \) but incurs larger investment costs. Thus we have proved the following proposition.

**Proposition 2** In any generation \( n \geq 2 \), the lowest trustworthiness is always 0 and the highest is not larger than \( e \). As a result, the profit of all principals is \( Q^*_0 \) in (9).

This proposition suggests that, if principals capture a partial rent \( Q^*(k, \alpha) - Q^*_0 \) in the above static model due to \( \alpha > 0 \), then it has to be transferred back to agents in this dynamic model of endogenous trustworthiness. That is, principals do not benefit from agent trustworthiness once it is costly to cultivate.

The optimal solution to problem (11) with \( \alpha_{n+1} = 0 \) and its comparative statics are stated in the following proposition.

**Proposition 3** (I) There exists a unique optimal solution \( \alpha^*_{n+1} \equiv g^*(\alpha_n; k) \) to problem (11), where \( g^*(\alpha_n; k) \) strictly increases in \( \alpha_n \) and \( k \). (II) In the steady state, agents with \( \alpha_1 = \alpha_1 \) have zero trustworthiness, while others have \( \alpha_{ss} = g^*(\alpha_{ss}; k) > 0 \) when the unit monitoring cost \( k \) is not too small. (III) In all the stable steady states with \( \alpha_{ss} > 0 \), \( \alpha_{ss} \) strictly increases in \( k \), and, contrary to the short-run result in Proposition 1, the total monitoring cost \( m^*(\alpha_{ss}; k)k \) decreases in \( k \) when the elasticity of \( \alpha_{ss} \) over \( k \) is high enough.

\(^{14}\)Note that this is equivalent to maximize the child’s future utility because the income is equal to the sum of child utility \( w^*(\alpha_{n+1}, k, \alpha_{n+1}) \) and the fixed cost of effort \( e \).

\(^{15}\)We ignore the perverse case where \( \alpha \) can be negative.
The uniqueness of the best response function $\alpha_{n+1}^* = g^*(\alpha_n; k)$ in this proposition is a natural result of the concavity of returns from trustworthiness $w^*(\alpha_{n+1}, k, \alpha_{n+1})$ and the convexity of investment cost. In the steady state, the bottom agents (e.g., those with $\alpha_1 = \alpha_1$) have zero trustworthiness, while others have either identical or different levels of positive trustworthiness, depending on the shape of $g^*(\alpha_n; k)$.

The case for a unique steady state positive trustworthiness is illustrated in Figure 4, where the arrows indicate the dynamics over generations. The best response function $\alpha_{n+1}^* = g^*(\alpha_n; k)$ is the same across individual agents and generations, but the range of distribution is shrinking over generations and eventually converges to the unique point $\alpha_{ss}$.

This proposition suggests that, when trustworthiness is endogenously determined, cheaper monitoring technologies may increase the governance cost, which is in stark contrast to the short-run result in Proposition 1. The intuition is as follows. If monitoring is cheaper in the next generation, the lifetime return of trustworthiness $w^*(\alpha_{n+1}, k, \alpha_{n+1})$ is lower, and so agents will invest less in it; this is reflected by the downward shifting of the best response function from $g^*(\alpha_n; k)$ to $g^*(\alpha_n; k')$ in Figure 5, which results in a lower steady state trustworthiness $\alpha_{ss}' < \alpha_{ss}$. But when the levels of trustworthiness are lower, principals have to monitor agents more intensively. When the effect of a higher monitoring intensity outweighs that of a lower unit monitoring cost, the total governance cost goes up; this happens when the elasticity of trustworthiness over $k$ is large enough.\(^{16}\)

\(^{16}\)The mechanism is in some sense similar to the familiar phenomenon that, when demand is elastic, a lower price
Figure 5: Lower Trustworthiness Caused by Cheaper Monitoring

Example. Suppose the cost function is $c(\alpha_{n+1}; \alpha_n) = (\alpha_{n+1})^a(1 + \alpha_n)^{-b}$ where $a > 1$ and $0 < b < a - 1/2$. Then the objective function for agents with $\alpha_1 > \alpha_{n+1}$ is

$$\max_{\alpha_{n+1}} w^*(\alpha_{n+1}, k, 0) - c(\alpha_{n+1}; \alpha_n) = \sqrt{ke} + \sqrt{k\alpha_{n+1}} - (\alpha_{n+1})^a(1 + \alpha_n)^{-b}.$$ 

The unique optimal child trustworthiness is $\alpha_{n+1}^* = g(\alpha_n; k) = \left[\frac{1}{2a}k^{\frac{1}{2}}(1 + \alpha_n)^{\frac{b}{2(2a-1)}}\right]^{\frac{1}{a-1}}$, which is strictly increasing in $k$ and $\alpha_n$, and strictly concave in $\alpha_n$. Thus there exists a unique steady state trustworthiness $\alpha_{ss}^*$ for all non-bottom agents with $\alpha_1 > \alpha_1$. In any generation $n+1$, the monitoring cost is

$$m^*(\alpha_{n+1}^*, k, 0)k = \sqrt{ke} - \sqrt{k\alpha_{n+1}^*} = \sqrt{ke} - k^{\frac{1}{2a}}\left[1 + (1 + \alpha_n)^{b}\right]^{\frac{1}{2a-1}},$$

which strictly decreases in $k$ when $e \leq [(2a)^{2a}(2a-1)^{1-2a}]^{\frac{1}{2a-1}}$.

2.4 Relative Cost of Monitoring Affected by Policies

A social planner, when deciding on how to allocate resources in reducing monitoring and inculcation costs, would take into consideration the dynamic crowding-out and crowding-in effects of monitoring technologies on agent trustworthiness. Individual principals, however, do not necessarily internalize the negative externalities they impose on agent incomes and thus tend to over-invest in monitoring technologies. The reason is that the long-run profit of principals, $Q_n^*$, increases when monitoring is cheaper, but it does not change when inculcation costs are lower; in contrast, the incomes of agents decrease in the former case, but increase in the latter. So principals gain but agents lose when the

may lead to a much higher demand and thus a higher total expenditure.
unit cost of monitoring is lower; agents benefit from lower inculcation costs, whereas principals are indifferent.

This conflict of interests between principals and agents seems fundamental in determining the basic incentive structure of a society’s resource allocation choices between reducing monitoring costs versus reducing inculcation costs, and hence may shape the long-term trends and cross-sectional variations of trust and monitoring intensity.

To fix ideas, suppose the inculcation cost of trustworthiness remains constant, while the government has to determine a policy that affects the relative cost of monitoring before the game starts, anticipating the policy’s effect on monitoring intensities and agent trustworthiness in the steady state. The government represents the interests of both principals and agents but with different weights, where the weight assigned to agent interests, $\gamma$, indicates the strength of labor power in policy determination. Note that when $\gamma = 0.5$, all individuals are treated equally, and hence the government’s choice coincides with the socially optimal result. The policy changes the unit cost of monitoring to $k(1 + t)$, where $t \in [-\tau, \tau] \subset (-1, 1)$ is the policy instrument that can be either tax or subsidy, and $\tau > 0$ is an exogenously determined constant. To simplify analysis, we assume that potential tax revenues or subsidies are absorbed/drawn outside the model. In this model, a positive tax rate will increase the relative cost of monitoring since the inculcation cost of trustworthiness is kept constant; alternatively it can be interpreted as keeping the monitoring cost unchanged but subsidizing the inculcation cost.

The objective function of the government is thus

$$\max_{t \in [-\tau, \tau]} \gamma \left( \sqrt{(1 + t)k\bar{e} e} + \sqrt{(1 + t)k\alpha_{ss} - e} \right) + (1 - \gamma) \left( hp - 2\sqrt{(1 + t)k\bar{e}} \right),$$

where the first term is the weighted aggregate utility of agents at the steady state, and the second term is the aggregate profit of principals. Note that $\alpha_{ss}$ increases in the unit cost of monitoring and thus also in $t$. It is straightforward to see that agents’ utility strictly increases in $t$, while the profit of principals strictly decreases in $t$. The FOC w.r.t. $t$ leads to

$$(3\gamma - 2)\sqrt{e} + \gamma\sqrt{\alpha_{ss}} + \gamma(1 + t^*)\alpha_{ss}^{-\frac{3}{2}} \frac{\partial\alpha_{ss}}{\partial t} \begin{cases} > 0 & \text{if } t^* = \tau \\ = 0 & \text{if } t^* \in (-\tau, \tau) \\ < 0 & \text{if } t^* = -\tau. \end{cases}$$

Note that, in the left hand side of the above condition, the sum of the first two items is positive when $\gamma \geq 0.5$, since $(3\gamma - 2)\sqrt{e} + \gamma\sqrt{\alpha_{ss}} > (4\gamma - 2)\sqrt{\alpha_{ss}}$ holds due to $e > \alpha_{ss}$, and the last item is
always positive. So as long as the weight on agent welfare is not lower than half, i.e., when $\gamma \geq 0.5$ holds, the marginal benefit of $t$ is always positive so that $t^* = \tau$. In other words, if the weight of agents is at least as high as that of principals, it is optimal for the government to increase the relative cost of monitoring by subsidizing the cultivation of trustworthiness. On the other hand, when $\gamma = 0$, the marginal benefit of $t$ is negative so that $t^* = -\tau$ holds; this is also true when $\gamma$ is small enough; that is, when the labor power is so low that government policies reflect mostly the welfare of principals, monitoring may very well be subsidized. For the interior solutions, we get

$$\frac{\partial t^*}{\partial \gamma} = \frac{1}{2} \left( \frac{k}{1 + t^*} \right)^{\frac{3}{2}} \frac{3\sqrt{\epsilon} + \sqrt{\alpha_{ss}} + (1 + t^*)\alpha_{ss}^{-\frac{1}{2}} \frac{\partial \alpha_{ss}}{\partial t}}{-SOC} > 0$$

as long as the second order condition $SOC < 0$ holds. That is, the optimal tax rate $t^*$ on monitoring cost strictly increases with $\gamma$. This implies that lower labor power $\gamma$ is associated with higher monitoring intensities, which is indeed consistent with Figure 2. These results are summarized in the following proposition.

**Proposition 4** The optimal tax rate $t^*$ on the monitoring cost $k$ increases with the weight $\gamma$ assigned to the welfare of agents. This implies that, across otherwise identical societies, those with lower labor power have higher monitoring intensities and lower agent trustworthiness, and thus rely more on monitoring and less on trust in comparison with others.

### 3 Costly Screening

An extension to the basic model is studied in this section, where an agent’s trustworthiness $\alpha$ is not publicly observed. A principal may pay some screening cost to observe a signal $z = \alpha + \varepsilon$, where $\varepsilon$ is a random variable with cdf $V(\cdot)$ and support $(-E, E)$ with $E > 0$.

The following analysis shows that the costly screening induces market friction and thus reduces competition among principals; this allows them to partially capture some rent generated by agent trustworthiness and thus to earn higher profits than in the basic model, which necessarily implies that agent wages would be lower than before. In the long run, however, the endogeneity of agent trustworthiness will again drive away any rent captured by principals. Apart from the new insight that agents are always worse off with higher screening costs but principals are either indifferent or better off, the main results of the basic model are robust to this extension; that is, there is still
a fundamental conflict of interests between principals and agents in the usage of monitoring and trustworthiness together with screening.

Note that screening enables principals to reduce governance cost by hiring more trustworthy agents and using more suitable incentive packages. Screening, however, is costly so that it may not be worthwhile to hire agents with low signals. Thus let $z_l$ denote the threshold signal, below which an agent is not worthwhile to be hired by a screening principal. Since the maximum reduction of the governance cost is $\sqrt{k\epsilon}$, a necessary condition for the existence of a positive mass of screened agents is $S < \sqrt{k\epsilon}$, where $S$ is the expected screening cost of a successful hire.\(^{17}\)

The time line of this screening game is similar to that in the basic model. Principals first decide whether to screen or not. Those who choose to screen announce their selection criterion $z_l$ and the incentive package $(m_s, w_s, b_s)$ as functions of $z$, hire the first agent with $z > z_l$ and reject others. Principals who do not screen would hire whoever comes first and adopt a single incentive package $(m_r, w_r, b_r)$, since all agents look the same to them. Agents then decide where to apply for jobs. If an agent is screened but turned down by a principal, it is publicly observed, though the signal $z$ is not; then this agent can only work for a non-screening principal, since a screening principal is better off by screening a fresh agent than him. As a result, no agents are ever screened more than once, which is meant to prevent endless searching and help principals to form expectations of their agents’ true trustworthiness. After matching is finished, agents get the basic wage, if any, and then choose whether to make the effort or shirk. Principals monitor agents, pay the incentive wages if shirking is not detected, and pay nothing if detected. The competitive equilibrium is reached when all principals stick to their screening choices, there is no partner-changing, and, once in a match, nobody wants to deviate from their decisions.

Again we solve the game backwards. Given the incentive package $(m_s, w_s, b_s)$ offered by a screening principal, the probability that the agent with a signal $z$ will make the effort once hired is

$$\Pr(\alpha > e - m_s w_s | z) = \Pr(\varepsilon < z - e + m_s w_s) = V(z - e + m_s w_s),$$

as implied by the non-shirking condition (5). So with probability $V(z - e + m_s w_s)$ the agent makes the effort, produces an expected output $hp$, and receives wage $w_s$, while with probability

\(^{17}\)This assumption is reasonable since a screening test can be purchased once but used for unlimited times, and in real life the staff in charge of recruiting are often paid a fixed salary that is not dependent on the number of job candidates interviewed. Allowing a positive marginal screening cost complicates the model without offering new insights.
1 - V(z - e + m_s w_s) he shirks, produces a lower expected output $hq$, gets caught with probability $m_s$ and loses the incentive wage $w_s$. A screening principal’s expected profit from hiring an agent with signal $z$ is thus

$$Q_s(z, z_l) = V(z - e + m_s w_s)(hp - w_s) + (1 - V(z - e + m_s w_s))(hq - (1 - m_s)w_s) - m_s k - b_s - S. \quad (12)$$

A non-screening principal does not incur any screening cost and thus observes no signals of an agent’s trustworthiness; she knows, however, that an agent not hired by a screening principal must have a signal $z < z_l$. Let $\Phi(\cdot)$ denote the cdf of the distribution of $z$, which is determined by the distributions of $\alpha$ and $\varepsilon$. Then given the incentive package $(m_r, w_r, b_r)$ offered by a non-screening principal, the probability that an agent shirks is $\Pr(\alpha < e - m_r w_r | z \leq z_l) = \frac{\Phi(e - m_r w_r)}{\Phi(z_l)}$.

A non-screening principal’s expected profit from hiring an agent is thus

$$Q_r(z_l) = (1 - \frac{\Phi(e - m_r w_r)}{\Phi(z_l)})(hp - w_r) + \frac{\Phi(e - m_r w_r)}{\Phi(z_l)}(hq - (1 - m_r)w_r) - m_r k - b_r. \quad (13)$$

When the labor market clears, the proportion of agents working for non-screening principals must be equal to that rejected by screening ones, which is $\Phi(z_l)$; this is also the proportion of principals who choose not to screen. Competition among screening principals equalizes their profits, which, in equilibrium, would also be the same as the optimal profit $Q^*_r(z^*_l)$ made by all non-screening principals due to competition pressure from them. That is, just as in the basic model, all principals earn the same expected profit $Q^*_r(z^*_l)$ in equilibrium, regardless of their screening choices and their agents’ trustworthiness. This is formally proved in the following proposition.

**Proposition 5** In the competitive equilibrium of this screening game: (I) There exists a unique signal $z^*_l$ such that agents with $z > z^*_l$ work for screening principals and others for non-screening principals, where $\partial z^*_l / \partial S > 0$ and $\partial z^*_l / \partial k < 0$. (II) All principals make the same profit $Q^*_r(z^*_l)$ that is larger than that in the basic model, where $\partial Q^*_r / \partial S > 0$, $\partial Q^*_r / \partial k < 0$, and $Q^*_r > Q^*(k, \alpha)$. (III) The optimal incentive packages include $b^*_s = b^*_r = 0$, $w^*_r = e$, $w^*_s \geq e$, and $w^*_s(z^*_l) = e$, while $m^*_s$ and $m^*_r$ maximize $Q_s$ and $Q_r$, respectively. $w^*_s$ is increasing in $k$, decreasing in $S$, and increasing and concave in $z$ and $\alpha$.

**Proof.** In the Appendix. ■

Note that the screening cost increases the profit of principals $Q^*_r(z^*_l)$ but reduces agent incomes in comparison to the basic model. The expected income of an agent is lower not only because it is
the agent that ultimately pays the screening cost $S$, but also because principals earn a higher profit $Q^*_r(z^*_f)$ than before. The intuition is that the positive screening cost is a form of market friction that reduces competition among principals and hence enables them to capture some rent generated by agents. Again, as in the basic model, principals do not gain from hiring an agent with higher trustworthiness, though they do gain when all agents are more trustworthy.

The endogenization of $\alpha$ with costly screening is similar to that in the basic model, except that in each generation the stage game with costly screening is played, and the competitive equilibrium derived in Proposition 5 prevails. Since agents on average receive lower wages when screening is costly and imperfect, they will have less incentives to cultivate trustworthiness than in the basic model. Actually, the imperfect screening may lead to complete extinction of trustworthiness and trust in the economy.

**Proposition 6** When screening for agent trustworthiness is costly, there is always a positive mass of agents that have zero trustworthiness in the steady state. When the signal $\varepsilon$ is too noisy, the only steady state is zero trustworthiness for all agents; steady states with positive trustworthiness $\alpha^{ss}_s$ occur only when signals obtained from screening are informative enough, in which case the properties of $\alpha^{ss}_s$ are the same as those in the basic model. Regardless of the distribution of $\varepsilon$, the endogeneity of trustworthiness reduces principals’ profits to $Q^*_0$.

**Proof.** In the Appendix.

The intuition is similar to that presented previously in the basic model, except for the new insight that cheaper monitoring not only reduces the amount of investment in each child’s trustworthiness, but also induces fewer agents to invest in it, the more so when the screening cost is higher.

Figure 6 illustrates the situation where the signal $\varepsilon$ is informative enough for some agents to have a unique positive trustworthiness $\alpha^{ss}_s$ in the steady state, while agents whose ancestors have lower trustworthiness (i.e., if $\alpha_1 \leq \pi$) will choose to have zero trustworthiness from the second period onwards. Though not shown explicitly, it is not difficult to see from Figure 6 that, if the unit monitoring cost $k$ is too low, the transition function $\alpha^{ss}_{n+1} = g^*_n(\alpha_n; k)$ may fall below the 45° line so that the only steady state is the origin. In other words, cheap monitoring may completely eliminate the incentives to cultivate trustworthiness, even though agent trustworthiness can be observed perfectly after screening.
In summary, costly screening makes the agents strictly worse off, but does not affect the profit level of principals. It thus reduces the incentives to cultivate trustworthiness and pushes principals further towards relying on monitoring. The conflict of interests between principals and agents in terms of using trustworthiness and monitoring remains the same as in the basic model except for the new insight that agents are better off with lower screening costs while principals are not. If the government has to choose an optimal tax rate on screening cost, a straightforward corollary to Proposition 4 would be that societies with lower labor power tend to have lower subsidies on screening costs and thus higher monitoring intensities and lower trustworthiness.

4 Conclusions

This paper analyzes the dynamic relationship between trust and monitoring in reducing moral hazard problems in a principal–agent setting. Agent trustworthiness and monitoring intensities are both determined by fundamental forces in society such as the costs of monitoring and screening agents and the cost of inculcating trustworthiness; their long-term trends and cross-sectional variation are thus shaped by how these relevant costs change. While acknowledging the influence of exogenous technical features on the cost-reduction process, we argue that an important role is also played by the inherent conflict of interests between principals and agents in equilibrium: principals benefit from lower monitoring costs, but not necessarily from lower screening and inculcation costs, whereas the opposite is true for agents. When monitoring becomes relatively cheaper, agent trust-
worthiness declines and monitoring intensities increase over time; they may do so at faster rates in societies where the interests of agents are given a lower weight in the choice of monitoring schemes. The overall governance costs, however, may be driven up by cheaper monitoring technologies, which crowd out intrinsic incentives and induce society to rely too much on extrinsic ones. These results are indeed consistent with preliminary empirical evidence, though more rigorous tests are needed in future research.

The main insights of this paper also apply to situations with general market frictions that give principals certain monopsony powers. Principals may capture a rent from agent trustworthiness when labor market frictions exist; the rent, however, is limited by the degree of frictions and, more importantly, it again disappears once trustworthiness becomes endogenous. Since labor market frictions increase principals’ profits but reduce agent incomes, principals have less incentive to eliminate them, while the opposite is again true for agents. These results can be readily obtained with similar arguments as in the text.

This paper can be extended in various directions to get a more thorough understanding of the relevant issues. For example, the resource allocation decisions on improving various governance modes can be explicitly modeled in a bargaining or political economy environment. The screening process can be fleshed out and repeated interactions between principals and agents may be added to better address potential problems associated with screening in a diverse and mobile society. The identical production ability of agents assumed in this paper can also be relaxed to study the trade-off or complementarity between investment in cognitive and non-cognitive skills from the perspective of aggregate welfare. For instance, if the difficulty in monitoring increases when higher cognitive abilities are involved, then the thoroughness of screening and the combination of governance modes should vary across jobs in some systematic way.

Appendix

Proof of Proposition 3. The objective function for agents with \( \alpha_1 > \underline{\alpha}_1 \) is
\[
\max_{\alpha_{n+1}>0} \sqrt{ke} + \sqrt{k\alpha_{n+1}} - C(\alpha_{n+1}, \alpha_n). 
\]
The FOC for an interior solution in the interval \( (0, e) \) is
\[
\frac{1}{2} \sqrt{k\alpha_{n+1}^{\star-1/2}} - C_1(\alpha_{n+1}^{\star}, \alpha_n) = 0. \tag{14}
\]
It yields the unique optimal choice \( \alpha_{n+1}^{\star} = g(\alpha_n; k) \) in each generation \( n \) since the second order
condition \(-\frac{1}{4}\sqrt{k}\alpha_{n+1}^{-3/2} - C_{11} < 0\) always holds. \(\alpha_{n+1}^*\) increases in \(\alpha_n\) and \(k\) because

\[
\frac{\partial \alpha_{n+1}^*}{\partial \alpha_n} = \frac{-C_{12}}{\frac{1}{4}\sqrt{k}\alpha_{n+1}^{-3/2} + C_{11}} > 0, \quad \frac{\partial \alpha_{n+1}^*}{\partial k} = \frac{\frac{1}{4}k^{-1/2}\alpha_{n+1}^{-1/2}}{\frac{1}{4}\sqrt{k}\alpha_{n+1}^{-3/2} + C_{11}} > 0.
\]

Note that the left-hand side of (14) is strictly increasing in \(k\); this implies that all parents except those with the lowest level of \(\alpha_n\) will invest in a positive \(\alpha_{n+1}^*\) since \(\lim_{\alpha_{n+1}^* \to 0} \frac{1}{4}\sqrt{k}\alpha_{n+1}^{-3/2} - C_{11} > 0\) holds. Furthermore, parents with \(\alpha_n > e\) will invest in \(\alpha_{n+1}^* \leq e\), and \(g(\alpha_n; k)\) is continuous and increasing in \(\alpha_n\). These three conditions suggest that there must exist at least one stable steady state \(\alpha_{ss} \in (0, e)\) such that \(g(\alpha_{ss}; k) = \alpha_{ss}\) holds. And \(\alpha_{ss}\) is unique if \(g(\alpha_n; k)\) is concave; this happens if

\[
g''(\alpha_n; k) = \frac{-C_{11}C_{122} + C_{12}C_{112}}{(\frac{1}{4}\sqrt{k}\alpha_{n+1}^{-3/2} + C_{11})^2} \leq 0.
\]

Note that \(\partial \alpha_{ss}/\partial k > 0\) holds because a higher \(k\) shifts up \(g(\alpha_n; k)\) due to \(\partial \alpha_{n+1}^*/\partial k > 0\).

The monitoring cost at the steady state, \(m_0^*(\alpha_{ss}, k) = \sqrt{ke} - \sqrt{k\alpha_{ss}}\), may actually increase when \(k\) is lower, since

\[
\frac{\partial m_0^*(\alpha_{ss}, k)}{\partial k} = \frac{\sqrt{e} - \sqrt{\alpha_{ss}}}{2\sqrt{k}} \quad \left\{\begin{array}{l}
\text{(+)} \quad k's \text{ direct effect on } \text{governance cost} \\
\text{(-)} \quad k's \text{ indirect effect via agent trustworthiness}
\end{array}\right.
\]

holds if the elasticity of \(\alpha_{ss}\) over \(k\) is high enough, i.e., if \(\frac{\partial \alpha_{ss}}{\partial k} \frac{k}{\alpha_{ss}} > \frac{e}{\alpha_{ss}} - 1\). In fact, as long as \(\alpha_{n+1}^*\) is endogenous, the monitoring cost \(m_0^*(\alpha_{n+1}^*, k)\) may be higher when monitoring is cheaper. Similar arguments suggest that, when there are multiple steady states, the trustworthiness levels and the corresponding governance costs in the stable states will exhibit the same properties with respect to \(k\).

Proof of Proposition 5. Similar arguments as in the proof of Proposition 1 suggest that \(b_s^* = b_r^* = 0\). After re-arranging terms of profit function (13) for a non-screening principal, we get

\[
Q_r(z_l) = hp - w_r - m_r k - \frac{F(e - m_r w_r)}{\Phi(z_l)} (h(p - q) - m_r w_r).
\]

Maximizing \(Q_r\) with respect to \(w_r\) and \(m_r\) subject to the participation constraint \(w_r - e \geq 0\) leads
to the first order conditions

\[
\frac{\partial Q_r}{\partial w_r} = -1 + \frac{F'(e - m_rw_r)m_r}{\Phi(z_l)}(h(p - q) - m_rw_r) + \frac{F(e - m_rw_r)m_r}{\Phi(z_l)} + \lambda_r = 0,
\]
\[
\frac{\partial Q_r}{\partial m_r} = -k + \frac{F'(e - m_rw_r)w_r}{\Phi(z_l)}(h(p - q) - m_rw_r) + \frac{F(e - m_rw_r)w_r}{\Phi(z_l)} = 0,
\]
\[
w_r - e \geq 0, \quad \lambda_r \geq 0, \quad \lambda_r(w_r - e) = 0.
\]

After rearranging the first two conditions, we get \(\lambda_r - 1 + km_r^*/w_r^* = 0\). If \(\lambda_r = 0\), we have \(w_r^* = km_r^*\), which cannot be true since wages and monitoring are substitutes in reducing shirking. If \(\lambda_s > 0\), we have \(w_r^* = e\), and then \(m_r^*\) is uniquely determined by \(\partial Q_r/\partial m_r = 0\), which is equivalent to

\[
F'(e - m_r^*e)(h(p - q) - m_r^*e) + F(e - m_r^*e) - \Phi(z_l)k/e = 0. \quad (15)
\]

By the envelope theorem we get

\[
\frac{\partial Q_s^*(z_l)}{\partial k} = -m_r^* < 0, \quad \frac{\partial Q_s^*(z_l)}{\partial z_l} = \Phi(z_l)^{-2}F'(e - m_r^*e)(h(p - q) - m_r^*e) > 0.
\]

So the non-screening principal’s profit is increasing in \(z_l\).

After re-arranging terms, the profit function (12) for a screening principal becomes

\[
Q_s(z, z_l) = hp - w_s - m_sk - (1 - V(z - e + m_sw_s))(h(p - q) - m_sw_s) - S.
\]

Maximizing \(Q_s\) w.r.t \(w_s\) and \(m_s\) subject to the wage constraint \(w_s \geq w_r^*\) leads to

\[
-1 + V'(z - e + m_sw_s)(h(p - q) - m_sw_s)m_s + (1 - V(z - e + m_sw_s))m_s + \lambda_s = 0,
\]
\[
-k + V'(z - e + m_sw_s)(h(p - q) - m_sw_s)w_s + (1 - V(z - e + m_sw_s))w_s = 0, \quad (16)
\]
\[
w_s - w_r^* \geq 0, \quad \lambda_s \geq 0, \quad \lambda_s(w_s - w_r^*) = 0.
\]

After rearranging the first two conditions, we get \(\lambda_s - 1 + km_s^*/w_s^* = 0\). Similar as above, the case for \(\lambda_s = 0\) is ruled out. When \(\lambda_s > 0\), we have \(w_s = w_r^* = e\), but then it will lead to the same monitoring level and thus different profits across screening principals, which cannot hold in the competitive equilibrium because non-screening principals would have incentives to change their decision and adopt screening. This suggests that the competition constraint from non-screening principals,

\[
Q_s^*(z, z_l) = Q_r^*(z_l), \quad (17)
\]
is binding and should be taken into consideration in the optimization. The optimal incentive package 
\((m^*_s, w^*_s)\) is thus determined as follows. Taking \(w_s\) as given, the FOC (16) yields \(m^*_s = \mu(w_s)\) as a function of \(w_s\), which can then be plugged in (17) to pin down \(w^*_s\). Note that based on (16) we get

\[
\frac{\partial m^*_s}{\partial w_s} = -\frac{\frac{\partial}{\partial w_s} \left( h(p-q) - m_s w_s \right)(V'' m_s w_s - V') - 2V' m_s w_s - V \right)}{(h(p-q) - m_s w_s)w_s^2 V'' - 2V' w_s^2} < 0,
\]

which is not surprising. Then the following condition

\[
Q^*_s(z, z_l) = h p - w^*_s - \mu(w^*_s)k - (1 - V(z - e + w^*_s \mu(w^*_s)))(h(p-q) - w^*_s \mu(w^*_s)) - S = Q^*_r(z_l) \tag{18}
\]

uniquely determines \(w^*_s\), which as in the basic model is the largest solution when multiple solutions exist, and at this point \(\partial Q^*_s / \partial w_s < 0\) and \(\partial^2 Q^*_s / \partial w_s^2 \leq 0\) must be true. Based on (18) we find that \(w^*_s\) increases in \(k\), decreases in \(S\) and \(z_l\), and increasing and concave in \(z\) and \(\alpha\) :

\[
\frac{\partial w^*_s}{\partial k} = \frac{-m^*_s + e}{-Q^*_s / \partial w_s} > 0, \quad \frac{\partial w^*_s}{\partial S} = \frac{-1}{-Q^*_s / \partial w_s} < 0, \quad \frac{\partial w^*_s}{\partial z_l} = \frac{-Q^*_r / \partial w_s}{-Q^*_s / \partial w_s} < 0,
\]

\[
\frac{\partial^2 w^*_s}{\partial \alpha^2} = \frac{\partial^2 w^*_s}{\partial z^2} = \frac{\partial w^*_s}{\partial z} = \frac{V'(h(p-q) - w^*_s \mu(w^*_s))}{-\partial Q^*_s / \partial w_s} \geq 0,
\]

\[
\frac{\partial^2 w^*_s}{\partial \alpha^2} = \frac{\partial^2 w^*_s}{\partial z^2} = \left\{ \frac{\partial Q^*_s}{\partial w_s}(1 + (\mu + \mu' w_s) \frac{\partial w^*_s}{\partial z}) + \frac{\partial^2 Q^*_s \partial w^*_s}{\partial w_s^2} V' \right\} \frac{\partial w^*_s}{\partial z} \frac{\partial w^*_s}{\partial w_s} - 2 < 0.
\]

Another constraint that has to be satisfied in the competitive equilibrium is \(w^*_s(z, z_l | z \geq z_l) \geq w^*_r\), which guarantees that agents with \(z > z_l\) have no incentives to match with non-screening principals, while agents with \(z = z_l\) are indifferent between the two kinds of principals. Since \(w^*_s\) is strictly increasing in \(z\) while \(w^*_r = e\) is a constant for any given \(z_l\), there exists a unique \(z^*_b(z_l)\) satisfying \(w^*_s(z^*_b(z_l), z_l) = e\), where \(z^*_b(z_l)\) is a strictly increasing function of \(z_l\) because \(z^*_b'(z_l) = \frac{\partial w^*_s}{\partial z_l} / \partial w_s / \partial z > 0\); this implies that there is a unique \(z^*_l\) satisfying \(z^*_b(z^*_l) = z^*_l\) or equivalently \(w^*_s(z^*_l, z^*_l) = e\).

Based on (19) we get

\[
\frac{\partial z^*_l}{\partial S} = \frac{-\partial w^*_s}{\partial w_s / \partial z - \partial w^*_s / \partial z_l} > 0, \quad \frac{\partial z^*_l}{\partial k} = \frac{-\partial w^*_s}{\partial w_s / \partial z - \partial w^*_s / \partial z_l} < 0.
\]

As in the basic model, competition between screening and non-screening principals implies that a rent has to be passed to an agent who has a signal \(z > z^*_l\) and works for a screening principal.
So all principals earn an identical profit equal to $Q_r^*(z_l^*)$, which decreases in $k$ and increases in $S$ because $Q_r''(z_l^*) \partial z_l^* / \partial k + \partial Q_r^* / \partial k < 0$ and $Q_r''(z_l^*) \partial z_l^* / \partial S > 0$. In general, principals make a higher profit when screening is available than in the basic model, that is, $Q_r^* \geq Q^*(k, \alpha)$ always holds, since non-screening principals can always get $Q^*(k, \alpha)$ by treating their agents as all having the lowest trustworthiness level $\alpha$.

**Proof of Proposition 6.** We first focus on two extreme cases and then show that the other cases have similar steady states.

**Case 1: No-Information.** The first case is when the signal $z$ is too noisy to contain any information about the true trustworthiness $\alpha$; for example, the noise variable $\varepsilon$ may take only two values $-\infty$ and $+\infty$ with equal probability, which blurs any differences among $\alpha$ and makes it impossible to distinguish between any agents in terms of their trustworthiness. We thus call this case the No-Information Case. In this case, since there is no way for trustworthiness to be recognized and rewarded by principals, being trustworthy carries no explicit returns but incurs a positive inculcation cost; an immediate implication is that no one has incentives to inculcate trustworthiness in children and thus all agents end up with zero trustworthiness: $\alpha_{n+1} = 0$ for any $n \geq 1$. As a result, no principals screen agents, the optimal incentive package is $(m_0^*, w_0^*) \equiv (\sqrt{e/k}, \sqrt{ke})$ as in the basic model for $\alpha = \alpha = 0$, and all principals make the same profit $Q_0^*$ in (9).

**Case 2: Full-Information.** The second case is almost the opposite, where the distribution of the noise $\varepsilon$ is degenerate, always having a fixed value, such that once the screening cost $S$ is paid, the signal $z$ observed by a screening principal reveals the true agent trustworthiness $\alpha$. This case is thus called Full-Information Case. The optimization problem for non-screening principals is the same as before so that $w_r^* = e$ and the optimal profit is denoted by $Q_r^*$, while the objective function for screening principals is

$$Q_s^* = \max_{w_s^* \geq w_r^*} hp - w_s^* - k(e - \alpha)/w_s^* - S,$$

subject to the competition constraint $Q_s^* = Q_r^*$ and the efficient monitoring condition $m_s^* = (e - \alpha)/w_s^*$. The optimal solution is thus $w_s^*(\alpha; Q_r^*) = \sqrt{k\alpha + \frac{1}{4}(hp - Q_r^* - S)^2 - ke + \frac{1}{2}(hp - Q_r^* - S)}$ for any agent with $\alpha > \alpha_t^*$, and $w_s^*(\alpha_t^*; Q_r^*) = e$ as in Proposition 5, where $\alpha_t^*$ is the selection criterion for being hired by a screening principal.

Since the pooling wage $w_r^*$ offered by non-screening principals is independent of an individual’s
trustworthiness $\alpha$ while the inculcation cost increases with $\alpha$, there is no gain for a pooling agent to acquire any $\alpha > 0$. Hence, from the second generation onwards, any agents not screened must have zero trustworthiness. An obvious implication is that the lowest level of trustworthiness among future agents would be 0, that is, $\alpha_{n+1} = 0$ in generation $n \geq 1$. Accordingly, non-screening principals will offer an incentive package $(m_0^*, w_0^*)$ and make a profit $Q_0^*$ in (9). Then due to competition screening principals also make $Q_0^*$, and the extra gain $Q_r^* - Q_0^*$ captured in the above static screening model is transferred back to agents. So principals do not benefit from agent trustworthiness once it becomes endogenous. These results are the same as in the basic model. Thus we have proved the following lemma.

**Lemma 1** For any generation $n + 1 \geq 2$, agents working for non-screening principals have zero trustworthiness and receive an incentive package $(m_0^*, w_0^*)$, while agents working for screening principals receive

\[
\begin{align*}
  w_0^*(\alpha; Q_0^*) &= \sqrt{\alpha} - \frac{1}{2} S + \sqrt{\alpha - S \sqrt{\alpha} + \frac{1}{4} S^2}, \\
  m_0^*(\alpha; Q_0^*) &= (e - \alpha)/w_0^*(\alpha; Q_0^*).
\end{align*}
\]

All principals get the same profit $Q_0^*$ as in the basic model.

The objective function for a parent in generation $n = 1, 2, \ldots$ is $\max\{R_s(\alpha_n), w_0^*\}$, where

\[
R_s(\alpha_n) \equiv \max_{\alpha_{n+1} > 0} w_0^*(\alpha_{n+1}; Q_0^*) - C(\alpha_{n+1}; \alpha_n)
\]

is the return if he ever invests in $\alpha_{n+1} > 0$ and his child is to be hired by a screening principal, and $w_0^*$ is the return of his best alternative, which is no investment so that his child with $\alpha_{n+1} = 0$ will work for a non-screening principal. Intuitively, if a parent $\alpha_n$ chooses to invest in his child, the resulting $\alpha_{n+1}$ must be high enough to enable the child to be hired after being screened and earn an income of at least $C(\alpha_{n+1}; \alpha_n) + w_0^*$ to justify the investment. Since $R_s(\alpha_n)$ strictly increases in $\alpha_n$, there exists a unique threshold trustworthiness $\bar{\alpha}$ for every generation $n$ such that

\[
R_s(\bar{\alpha}) = w_0^*.
\]

So descendants of families with $\alpha_1 \leq \bar{\alpha}$ would have $\alpha_{n+1}^* = 0$ for all generations $n + 1 \geq 2$ and work for non-screening principals. Only those with $\alpha_{ni} > \bar{\alpha}$ would ever invest in their children, whose optimal choices are stated in the following lemma, which is similar to Proposition 3 in the basic model.
Lemma 2  (I) There exists a unique optimal solution $\alpha_{n+1}^* \equiv g_s^*(\alpha_n; k)$ to problem (22), where $g_s^*(\alpha_n; k)$ strictly increases in $\alpha_n$ and $k$. (II) When $k$ is not too low, there exists at least one stable steady state with $\alpha_{ss}^* = g_s^*(\alpha_{ss}^*; k) > 0$ for all agents from families with $\alpha_1 \geq \bar{\alpha}$ and 0 for others; that is, a proportion $\pi \equiv 1 - F(\bar{\alpha})$ of agents have $\alpha_{ss}^*$, where $\pi$ increases in $k$ but decreases in $S$. (III) In all the stable steady states with $\alpha_{ss}^* > 0$, $\alpha_{ss}^*$ strictly increases in $k$, and, contrary to the short-run result in Proposition 5, the monitoring cost $m_s^*(\alpha_{ss}^*, k)$ decreases in $k$ when the elasticity of $\alpha_{ss}^*$ over $k$ is high enough.

Proof. The objective function is

$$\max_{\alpha_{n+1} > 0} \sqrt{k e} + \sqrt{k \alpha_{n+1} - S \sqrt{k e} + \frac{1}{4} S^2 - \frac{1}{2} S - C(\alpha_{n+1}, \alpha_n)}.$$ 

The first order condition for an interior solution is

$$(k \alpha_{n+1} - S \sqrt{k e} + \frac{1}{4} S^2)^{-1/2} k - C_1(\alpha_{n+1}, \alpha_n) = 0. \quad (24)$$

It yields the unique optimal choice $\alpha_{n+1}^* \equiv g_s(\alpha_n; k)$ in each generation $n$ because the second order condition $SOC_s < 0$ always holds due to concavity of $w_s^*(\alpha_{n+1}; Q_n^0)$ and $C_1 > 0$. $\alpha_{n+1}^*$ is increasing in both $\alpha_n$ and $k$ because $\frac{\partial \alpha_{n+1}^*}{\partial \alpha_n} = -\frac{C_1 e}{SOC_s} > 0$, and $\frac{\partial \alpha_{n+1}^*}{\partial k} = \frac{\partial^2 w_s^* / \partial \alpha_{n+1} \partial k}{-SOC_s} > 0$ holds due to $\partial^2 w_s^* / \partial \alpha_{n+1} \partial k > 0$, which is derived as follows. Based on $hp - w_s^* - k(e - \alpha_{n+1})/w_s^* - S = hp - 2 \sqrt{k e} = Q_0^*,$ we get

$$\frac{\partial w_s^*}{\partial k} = e - \alpha_{n+1}, \quad \frac{\partial^2 w_s^*}{\partial \alpha_{n+1} \partial k} = \frac{w_s^*}{(w_s^* - km_s^*)^2} > 0.$$

So the left-hand side of (24) strictly increases in $k$. This implies that when $k$ is not too small, there exists at least one positive steady state $\alpha_{ss}^* > 0$ such that $g_s(\alpha_{ss}^*; k) = \alpha_{ss}^*$ holds, and $\alpha_{ss}^*$ is unique if $g_s(\alpha_n; k)$ is concave in $\alpha_n$. Note that $\partial \alpha_{ss}^*/\partial k > 0$ holds, since a higher $k$ shifts up the transition function due to $\partial \alpha_{n+1}^*/\partial k > 0$.

The monitoring cost at the steady state $m_s^*(\alpha_{ss}^*, k) = k(e - \alpha_{ss}^*)/w_s^*(\alpha_{ss}^*, k)$ may actually decrease in $k$ if the elasticity of $\alpha_{ss}^*$ over $k$ is high enough:

$$\frac{\partial m_s^*(\alpha_{ss}^*, k)}{\partial k} = m_s^* + k \frac{\partial m_s^*}{\partial k} + \frac{\alpha_{ss}^*}{w_s^*} \left( 1 + \frac{e}{\alpha_{ss}^*} - 1 \right) \frac{\partial w_s^*}{\partial \alpha} \frac{\partial \alpha_{ss}^*}{\partial k} \frac{k}{\alpha_{ss}^*} < 0$$

$k$’s direct effect on monitoring cost

$k$’s indirect effect via agent trustworthiness
holds when the indirect effect is big enough. In fact, as long as $\alpha_{n+1}$ is endogenous, the monitoring cost $m^*_s(\alpha^*_{n+1}; k)$ may be higher when monitoring is cheaper.

Based on $\pi = 1 - F(\bar{\pi})$ and (23) we get

$$\frac{\partial \pi}{\partial S} = -F'(\bar{\pi}) \frac{\partial R_s(\bar{\pi})}{\partial S} = F'(\bar{\pi}) \frac{\partial w^*_s(\alpha_{n+1}; Q_0^*)}{\partial S} < 0,$$

$$\frac{\partial \pi}{\partial k} = -F'(\bar{\pi}) \frac{\partial R_s(\bar{\pi})}{\partial k} = F'(\bar{\pi}) \frac{\partial w^*_s(\alpha_{n+1}; Q_0^*)}{\partial k} > 0.$$ 

So the proportion of agents with a positive trustworthiness decreases in screening cost $S$ and increases in monitoring cost $k$. ■

**Other Cases.** It turns out that the steady state results of these two extreme cases generalize to any distribution of the noisy variable $\varepsilon$. The intuition is as follows. Since screening is costly and imperfect, the barriers to benefit from trustworthiness are higher than in the basic model, which may lead to a non-negligible mass of agents to adopt zero trustworthiness for their children from second generation onwards. If agents with positive trustworthiness cannot be distinguished from them and hence paid the same wage, all agents will eventually adopt zero trustworthiness as in the No-Information case. When the signal is informative enough for agents with zero trustworthiness to form a clear separating group from the rest, it is possible for positive trustworthiness to exist in the steady state as in the Full-Information Case. In all these cases, the endogenous choice of trustworthiness increases competition among principals and decreases their profits to $Q_0^*$ regardless of the distribution of the noise variable $\varepsilon$.

Multiple levels of trustworthiness may occur in the steady state only when the signals are precise enough for principals to distinguish among them. Suppose there are three distinct positive trustworthiness levels $0 < \alpha_L < \alpha_M < \alpha_H$ in the steady state, but the signals are too noisy for principals to distinguish one from another. If principals find it optimal to set their incentive package to target at the lowest trustworthiness level $\alpha_L$, i.e., to guarantee that these agents won’t shirk, the other agents won’t shirk either as they are more trustworthy. Given the decision of principals, however, parents of agents with $\alpha_M$ and $\alpha_H$ have incentives to reduce them, because they receive the same wage as the less trustworthy agents with $\alpha_L$ but need to incur higher cultivation costs. So $\alpha_L < \alpha_M < \alpha_H$ cannot happen in the steady state. Realizing that all agents must have the same trustworthiness level $\alpha_L$, principals have to set the incentive package $(w^*_S, m^*_L)$ as in the Full-Information Case discussed above. But this may or may not be an equilibrium depending on
how informative the signal $z$ is: if the signals are very noisy, agents may find it beneficial to slightly reduce their trustworthiness further below $\alpha_L$ and shirk, which means in the end the steady state contains only zero trustworthiness; if the signals are informative enough to stop such unraveling, then it is a steady state equilibrium. If principals target at the middle or top agents, the same logic suggests that the same two types of steady states will happen: either all agents have zero trustworthiness or all have $\alpha_L > 0$, depending on the noisiness of the signal. Similar results also apply to the case of a continuous distribution of $\alpha$ on the interval $[\alpha_L, \alpha_H]$ where principals cannot distinguish among them.

For a positive trustworthiness $\alpha_{ss}^s > 0$ to exist in the steady state for some agents, a necessary condition is to screen out those with zero trustworthiness, which requires $E < \alpha_{ss}^s - E$ to hold or equivalently $\alpha_{ss}^s > 2E$. Another condition is that agents must not have incentives to reduce their trustworthiness just a little below $\alpha_{ss}^s$, which is satisfied when

$$V'(-E) \left[ \sqrt{2kE - S\sqrt{ke} + \frac{1}{4}S^2 - \frac{1}{2}S} \right] \geq C_1(\alpha_{ss}^s, \alpha_{ss}^s),$$

holds; this is formally proved as follows.

By reducing trustworthiness to $\alpha_{ss}^s - \xi$, the agent gets a signal $z = \alpha_{ss}^s - \xi + \varepsilon$; it is larger than $\alpha_{ss}^s$ with probability $1 - V(\xi - E)$, in which case the agent passes the screening process to get utility $(1 - m_{\alpha L}^s)w_{\alpha L}^s - \alpha_{ss}^s + \xi$, otherwise he is screened out and gets utility $w_0^s - e$. So the agent’s expected utility of investing in $\alpha_{ss}^s - \xi$ is

$$[1 - V(\xi - E)][(1 - m_{\alpha L}^s)w_{\alpha L}^s - \alpha_{ss}^s + \xi] + V(\xi - E)(w_0^s - e) - C(\alpha_{ss}^s - \xi, \alpha_{ss}^s),$$

where $w_0^s = \sqrt{ke}$, $w_{\alpha L}^s$ and $m_{\alpha L}^s$ are from (20) and (21) with $\alpha = \alpha_{ss}^s$. The optimal level of $\xi^*$ must satisfy the FOC

$$V'(\xi^* - E) \left[ \frac{1}{2}S - \sqrt{k\alpha_L - S\sqrt{ke} + \frac{1}{4}S^2 - \xi^*} \right] + 1 - V(\xi^* - E) + C_1(\alpha_{ss}^s - \xi^*, \alpha_{ss}^s) \leq 0,$$

which boils down to $V'(-E) \left[ \sqrt{k\alpha_L - S\sqrt{ke} + \frac{1}{4}S^2 - \frac{1}{2}S} \right] \geq C_1(\alpha_{ss}^s, \alpha_{ss}^s)$ when $\xi^* = 0$ holds. This together with the necessary condition $\alpha_{ss}^s > 2E$ implies (25), which is a sufficient condition for no deviation.

An alternative proof of Proposition 1 with outcome-contingent wages:

Suppose a principal offers outcome-contingent wages $w_h$ and $w_0$ for good and bad outcomes, respectively. Note that the principals has to set up the monitoring intensity before the realization
of outcomes, which means that \( m \) is independent of outcomes. Since we already know that \( b = 0 \) in the optimal choice, in this proof we will directly assume it. So given the incentive package \((w_h, w_0, m)\), an agent does not shirk if

\[
pw_h + (1 - p)w_0 - e \geq (1 - m)(qw_h + (1 - q)w_0) - \alpha
\]

holds, which is the incentive constraint condition.

A natural starting point of the analysis is a principal who hires the least trustworthy agent \( \alpha \) in the population and thus faces no competition from other principals. If a positive monitoring level \( m^* > 0 \) is ever chosen, the agent will make the effort, and thus the principal’s objective function is

\[
\max_{w_h, w_0} \ hp - pw_h - (1 - p)w_0 - km \\
\text{s.t. (IR)} \ pw_h + (1 - p)w_0 - e \geq 0, \\
\text{(IC)} \ pw_h + (1 - p)w_0 - e \geq (1 - m)(qw_h + (1 - q)w_0) - \alpha.
\]

Let \( \lambda \) and \( \mu \) denote the multipliers for IR and IC respectively. Then the FOCS for solutions \((w_h, w_0, m)\) are

\[
w_h : -p + \lambda p + \mu(p - (1 - m)q) = 0, \\
w_0 : -(1 - p) + \lambda(1 - p) + \mu[(1 - p) - (1 - m)(1 - q)] \leq 0, \\
m : -k + \mu[qw_h + (1 - q)w_0] = 0.
\]

We first show that \( w_0^* = 0 \). If not, then the equality holds in the FOC \( w_0 \), but then the first two conditions imply that \( p = q \), which contradicts our assumption \( p > q \). It is also easy to see that \( \lambda = 0 \) must hold so that the IR condition is not binding, because otherwise the incentive package is identical to all agents with heterogeneous \( \alpha \), which is definitely not optimal. Solving the remaining conditions out, we get

\[
w_h^*(\alpha, k, \alpha) = \sqrt{k(e - \alpha)/pq}, \\
m^*(\alpha, k, \alpha) = \sqrt{p(e - \alpha)/kq} - (p - q)/q, \\
w_0^*(\alpha, k, \alpha) = 0.
\]

Using the same arguments as in the proof of Proposition 1 where all principals have to get the
same profit, we get

\[
w^*_i(\alpha, k, \alpha) = \sqrt{p/pq}(\sqrt{e - \alpha} + \sqrt{\alpha - \alpha}),
\]
\[
m^*(\alpha, k, \alpha) = \sqrt{p/kq}(\sqrt{e - \alpha} - \sqrt{\alpha - \alpha}) - (p - q)/q,
\]
\[
w^*_0(\alpha, k, \alpha) = 0.
\]

And the optimal profit for principals is

\[
Q^*(k, \alpha) = hp - 2\sqrt{k(e - \alpha)/pq} - k(p - q)/q.
\]

These results are almost identical with those in the basic model except for some minor adjustments. So allowing outcome-contingent wages will not affect any of the main results of the paper.

References


