# The Division of Surplus and the Burden of Proof*

Deniz Kattwinkel† Justus Preusser‡

March 14, 2025

**Abstract**

A surplus must be divided between a principal and an agent. Only the agent knows the surplus' true size and decides how much of it to reveal initially. Both parties can exert costly effort to conclusively prove the surplus' true size. The agent's liability is bounded by the revealed surplus. The principal is equipped with additional funds. The principal designs a mechanism that allocates the burden of proof and divides the surplus. In principal-optimal mechanisms, the principal's effort to acquire proof decreases in the revealed surplus. The agent's effort initially decreases, but then the sign of its slope alternates across five intervals. Applications include wealth taxation, corporate finance, and public procurements.

*We thank Christoph Carnehl, Gregorio Curello, Jan Knöpfle, Nenad Kos, Marco Ottaviani, Ludvig Sinander, Roland Strausz and various seminar audiences for helpful comments. Preusser acknowledges financial support by the European Research Council (HEUROPE 2022 ADG, GA No. 101055295 – InfoEcoScience).

†University College London, Department of Economics, *d.kattwinkel@ucl.ac.uk*

‡Bocconi University, Department of Economics and IGIER, *justus.preusser@unibocconi.it*

# 1 Introduction

A principal and an agent have to divide a surplus, but only the agent knows its true size. Both parties can acquire conclusive evidence about the size, but only at a cost. Who bears the burden of acquiring evidence, and how does the burden shape the division of surplus? We analyze the full solution to this problem: principal-optimal mechanisms that simultaneously divide the surplus and assign the burden of proof among the parties.

This problem arises in fundamental applications. The state and a citizen divide the citizen's wealth with a tax, an investor and an entrepreneur divide the returns from an investment, the state and a monopolist divide the costs from providing a public good. The size of the surplus is initially known to one party (agent) but not the other (principal). The citizen privately knows their wealth, the entrepreneur privately observes the returns, and the monopolist privately knows the costs. But both parties can acquire conclusive proof about the size of the surplus. The citizen/entrepreneur/monopolist can generate certified financial statements, or the state/investor can conduct a costly audit. Absent such proof, the state/investor can only seize what the citizen/entrepreneur/monopolist voluntarily advances as payment.

In these applications, the principal (state/investor) can more credibly commit and has more financial resources than the agent (citizen/entrepreneur/monopolist). Therefore, we study which mechanisms the principal optimally offers the agent. The agent's liability is limited by the surplus, whilst the principal has additional funds.

In our model, the agent initially holds the surplus—the agent's *type*. The principal can only seize the revealed part of the surplus. There are two ways of revealing surplus. First, the agent can reveal an amount by advancing it as payment. Second, evidence by either party conclusively reveals the difference between the advance payment and the true surplus. Each party chooses a probability of obtaining evidence—their effort—at an increasing cost. The agent's effort is not verifiable. Both parties maximize their transfer net of effort costs.

The agent can costlessly prove the existence of assets by advancing them. The advance payment by the agent captures any costless revelation of assets that makes them seizable. For example, cash holdings can be transferred to the principal or an intermediary. The ownership of other assets can be revealed via legal documents. In contrast, to prove that they do not own any further assets, the agent has to bear a

cost, e.g. by commissioning an external auditor. Similarly, the principal bears a cost for revealing the assets that the agent owns but did not advance.

Transfers between the agent and the principal must be funded out of the revealed surplus or out of the principal's private funds. We have in mind that the principal's private funds are large.

A mechanism has the following timing. First, the agent advances a payment. Second, the agent exerts effort to acquire evidence and, if successful, decides whether to disclose evidence. Third, the principal exerts effort to acquire evidence, but only if the agent has not already disclosed evidence. Finally, the principal seizes a portion of the revealed surplus, and possibly makes a transfer to the agent out of the principal's funds. Crucially, the division of surplus and the transfers may depend on who provided evidence. As we will show, the agent's and the principal's efforts to acquire evidence will play different roles.

A version of the Revelation Principle applies: the above timing is optimal, and the principal incentivizes the agent to advance the full surplus (the agent's type). When the agent does not advance the full surplus, we say the agent deviates. Whenever there is evidence that the agent deviated, the principal seizes the full surplus. Therefore, the agent will acquire evidence only if they advanced the surplus in the first place.

The principal faces a hidden information and a hidden action problem: the agent must be incentivized to advance the true surplus and to acquire and present evidence. The agent's rent from presenting evidence—the *evidence rent*—strengthens the agent's incentive to advance the surplus in the first place, thereby contributing to solving the hidden information problem.

Our contribution is to solve the entangled problems of assigning the burden of proof and dividing the surplus. We study the canonical model of state verification, the surplus division problem (see seminal work by Townsend (1979) and Border and Sobel (1987)), but allow *both* sides to acquire costly conclusive proof. We uncover rich interactions between agent- and principal-evidence that shape the optimal allocation of proof and surplus.

As a first step, we show that to maximize profit it suffices to consider a subclass of mechanisms that we call *tight*. Roughly, a mechanism is tight if the principal cannot simultaneously extract a higher profit from every type and threaten every type with a higher loss from misreporting. In our set-up, tightness already imposes substantial structure on the mechanism.

We show, as our main result, that all optimal tight mechanisms have the following structure. The mechanism divides the possible surplus levels into five intervals, all with a non-empty interior. The principal acquires evidence with interior probability, except if the agent advances the very highest possible surplus. Further, the principal's effort is decreasing in the advance payment. By contrast, the agent's effort is non-monotone: it is strictly decreasing on the first, third, and fifth interval, but strictly increasing on the second and fourth interval.

We explain this structure via the different roles of the principal's and the agent's efforts. The principal's effort to acquire evidence following some advance payment $y$ deters other types $x$ from wrongfully advancing $y$ since the principal seizes the full surplus upon finding evidence that the agent withheld parts of the surplus. The agent's effort plays two roles. First, whenever the agent presents evidence, the principal does not have to exert costly effort for deterrence. Given the principal's effort, we define the efficient agent-effort as the minimizer of total effort costs. Second, the agent gets the evidence rent from exerting effort only if they advance the full surplus in the first place. Therefore, incentivizing agent-effort in the hidden action problem strengthens the incentive to advance the full surplus in the hidden information problem.

The principal can seize more if the agent has more surplus. Therefore, a lower principal-effort should suffice to deter misreports to higher surplus levels. Indeed, we show that the principal's effort decreases in the advanced surplus. The efficient agent effort co-monotonically decreases with the principal's effort.

We now describe the five intervals in more detail.

(1) For super low surplus (1st interval), the implemented agent-effort is inefficiently low but decreasing. These types have strict incentives to advance their full surplus. Therefore, the role of agent-effort is only to reduce the principal's costs but is not needed to strengthen the incentive to advance the surplus. The principal bears the full evidence rent and distorts the agent-effort below the efficient level. The implemented agent-effort decreases co-monotonically with the principal's effort.

(2) For low surplus (2nd interval), the implemented agent-effort is increasing but inefficiently low. Starting in this interval, types have a binding incentive to advance the full surplus. Therefore, the principal can offset the evidence rent with the incentive benefits for the hidden information problem. The necessary hidden information rent is higher for higher types. Thus, the implemented

agent-effort increases in the type, and, at the top of the interval, reaches the efficient level.

(3) For middle surplus levels (3rd interval), the implemented agent-effort is efficient and decreasing. The principal provides the necessary information rent also via a refund in the event that the principal's evidence proves the agent advanced the full surplus. At the top of the interval, the principal exhausts their private funds in this event.

(4) For high surplus (4th interval), the implemented agent-effort is increasing and inefficiently high. Since the principal now exhausts their private funds following principal-evidence, the evidence rent generated by incentivizing agent-effort is now even more valuable for incentivizing the agent to advance the full surplus.

(5) For super high surplus (5th interval), the implemented agent-effort is decreasing and inefficiently high. Incentivizing the agent to advance their super high surplus becomes increasingly difficult: the principal exhausts their private funds following principal evidence, and the agent's effort is so inefficiently high that it is too expensive to provide a higher evidence rent. Therefore, the principal must offer a refund even if neither the agent nor the principal prove that the agent advanced the full surplus.

This characterization of *optimal* tight mechanisms is a special case of the general characterization of tight mechanisms. The difference is that suboptimal tight mechanism may have two additional intervals: one at the very bottom, where the principal acquires evidence with certainty; the other one at the very top, where the principal never acquires evidence; in-between, the mechanism is characterized by five intervals as above, but the intervals may be empty.

If the principal acquires evidence randomly following at least one possible advance payment, all five intermediate intervals have a non-empty interior. If the principal only acquires evidence deterministically, all five intermediate intervals are empty.

We present a ready-to-apply algorithm that turns an arbitrary given mechanism into a tight mechanism with a profit that is (weakly) higher for every possible surplus level. Tightness is a distribution free concept, and also the algorithm does not require the surplus distribution as an input. To characterize optimal tight mechanisms (via the five intervals) we only assume that the distribution is continuous at the very highest surplus level.

Our algorithm suggests simple ways of improving any given mechanism that can

serve as policy advice. Fixing a type, the tightening algorithm tests possible modifications of the mechanism at this type only—instead of more complex modifications that change the mechanism at multiple types. In the taxation problem, for example, consider a citizen with moderate wealth $y$. To deter wealthy types from understating their wealth, the principal must conduct an audit with some probability if the citizen advances $y$. The algorithm tests, for example, if a policymaker should incentivize type $y$ to provide more evidence when they advance $y$, thereby reducing the audit costs. As another example, for sufficiently wealthy citizens (high $y$), the policymaker optimally induces higher agent-effort than what is explained via cost reduction. Indeed, the policymaker can motivate a wealthy citizen to not evade taxes by promising a substantial refund if the citizen provides evidence for having advanced their full wealth.

Finally, we further characterize optimal tight mechanisms by studying the principal's trade-off across different surplus types. The trade-off is between the principal's effort costs at a type $y$ and the surplus that the principal extracts from higher types who are indifferent to advancing $y$ instead of their full surplus. We first show that, in an optimal tight mechanism, for every type $y$ one can find a strictly higher type $\hat{x}(y)$ such that type $\hat{x}(y)$ has a binding incentive constraint to deviate to $y$ and such that $\hat{x}(y)$ is strictly increasing in $y$. These non-local incentive constraints present a major challenge and render standard techniques from mechanism design inapplicable. In our problem, perturbing the mechanism at $y$ impacts the incentives of a distant type—namely, $\hat{x}(y)$. We characterize a necessary first-order condition on the trade-off between $y$ and $\hat{x}(y)$ in isolation from other types.

An important direct consequence from our optimality characterization is that every type of the agent except the very highest one is incentivized to exert a strictly positive effort and, thus, enjoys a strictly positive evidence-rent. The very highest type enjoys a strictly positive information rent. The prediction that all types enjoy strictly positive rent in the optimal mechanism contrasts standard findings for hidden information or hidden action problems, including the findings of Townsend (1979) and Border and Sobel (1987).

Another consequence is that the principal never acquires evidence with certainty: doing so represents an excessive threat since the agent always enjoys a rent from being truthful. In particular, deterministic mechanisms are always suboptimal. The corporate finance literature (e.g. Tirole (2010)) discusses Townsend (1979) as a rationale for the use of debt contracts—which are deterministic—but notices that debt

contracts can be suboptimal (Border and Sobel (1987), Mookherjee and Png (1989)). In our set-up, tight deterministic mechanisms are debt contracts, modified with a clause that incentivizes agent-effort. In contrast to the set-up of Townsend (1979) and Border and Sobel (1987), where under some circumstances optimal mechanisms can be deterministic, we rule this out under any circumstances.

# 2 Model

## 2.1 Set-up

There is a principal and an agent. The principal has private funds $\tau > 0$. The agent holds a surplus $x \in [\underline{x}, \bar{x}]$, where $0 < \underline{x} < \bar{x} < \infty$. The surplus $x$—the agent's *type*—is the agent's private information. The type distribution is denoted $F$, and the minimum (respectively, maximum) of the support of $F$ is $\underline{x}$ (resp., $\bar{x}$).[1]

The agent can make an *advance payment* $y \in [0, x]$ to the principal. The advance payment is contractible and proves the existence of the advanced portion of the surplus. In applications, the advance payment can be a transfer of cash to the principal or an intermediary, or the provision of documents that prove the existence of some assets.

In addition, after the advance payment has been made, both the agent and the principal can acquire conclusive *evidence* about the difference between the advance payment and the true surplus, and, thus, make the true surplus contractible. The agent's advance payment $y$ reveals that the surplus $x$ is at least $y$, and evidence (regardless of who provides it) reveals $x$ exactly. To obtain evidence with probability $e_A \in [0, 1]$, the agent incurs a cost $c_A(e_A)$. If obtained, the agent can (but need not) present the evidence. Similarly, to obtain evidence with probability $e_P \in [0, 1]$, the principal incurs a cost $c_P(e_P)$. While the agent's evidence acquisition effort $e_A$ is not contractible, the principal's effort $e_P$ is contractible. Each party can attempt to obtain evidence only once.

The set of feasible transfers depends on the agent's advance payment and whether evidence was provided. Specifically, if type $x$ advances $y$ and no party provides evidence, then the transfer $t$ from the agent to the principal must be in $[-\tau, y]$; if someone provides evidence, the transfer $t$ must be in $[-\tau, x]$. Given a transfer $t$ and

[1] In the application of the monopolist providing a public good, the model is interpreted as follows. The monopolist has no initial assets. The state provides $\bar{x}$. The monopolist then privately learns the realized costs $k \leq \bar{x}$ for providing the good. The state seeks to retrieve the unused assets $x = \bar{x} - k$.

efforts $e_A$ and $e_P$, the ex-post payoffs of the agent and the principal, respectively, are $x - t - c_A(e_A)$ and $t - c_P(e_P)$, respectively.

**Mechanisms.** We analyze principal-optimal mechanisms.

**Definition 2.1.** A *tax mechanism* is given by a quintuple $(e_A, e_P, r_A, r_P, r_\emptyset)$ of functions and plays out as follows:

(1) The agent makes an advance payment $y \in [0, \bar{x}]$ to the principal.
(2) The principal recommends an effort $e_A(y)$ to acquire evidence.
(3) The agent (covertly) exerts effort. If the agent obtains evidence, the agent chooses whether to disclose it.
(4) (a) If the agent discloses evidence, the principal does not acquire evidence.
    (b) If the agent's advance payment $y$ is in $[0, \underline{x})$, the principal acquires evidence with probability one.
    (c) Otherwise, the principal exerts effort $e_P(y)$ to acquire evidence.
(5) (a) If there is evidence showing that the agent's advance payment is different from the full surplus, the principal seizes the full surplus.
    (b) Otherwise, the principal seizes the advance payment and pays the following refund to the agent:
        (i) $r_A(y)$ if the agent provided evidence, for a total transfer $y - r_A(y)$ from the agent to the principal;
        (ii) $r_P(y)$ if the principal provided evidence, for a total transfer $y - r_P(y)$;
        (iii) $r_\emptyset(y)$ if neither provided evidence, for a total transfer $y - r_\emptyset(y)$.

A tax mechanism is *feasible* if for all $y \in [0, \bar{x}]$ the efforts $e_A(y)$ and $e_P(y)$ are in $[0, 1]$, and the refunds $r_A(y)$, $r_P(y)$, and $r_\emptyset(y)$ are in $[0, y + \tau]$.

A tax mechanism is *incentive compatible (IC)* if each type $x$ of the agent has a best response to advance the full surplus ($y = x$), exert the recommended effort $e_A(x)$, and, if available, disclose evidence.

A version of the Revelation Principle (Appendix A.1) applies: feasible IC tax mechanisms suffice for maximizing the principal's profit. Henceforth, these are simply called *mechanisms*.

A mechanism has *non-random audits* if the principal's effort $e_P(x)$ is either 0 or 1 for all for all $x \in [\underline{x}, \bar{x}]$. Else, a mechanism has *random audits*.

**Assumptions.** The agent's costs $c_A$ and the principal's costs $c_P$ are thrice differentiable, strictly increasing, strictly convex, and $c_A(0) = c_P(0) = c'_A(0) = c'_P(0) = 0$

holds. We make three further assumptions. The first two require that the principal's funds $\tau$ are sufficiently large.

**Assumption 1.** It holds $\tau \geq c'_A(1) > c_P(1)$.

As the proofs will show, $\tau \geq c'_A(1)$ implies that the principal's funds suffice for incentivizing the agent to acquire evidence with certainty, but $c'_A(1) > c_P(1)$ implies that doing so is not optimal.

**Assumption 2.** For all $e_A \in [0,1]$,

$$\underline{x} + \tau > (1 - 2e_A)c'_P(1) + e_A(1 - e_A)\frac{c'''_A(\tilde{e}_A)}{c''_A(e_A)}.$$

As we explain in the context of our main characterization (Section 3.1), Assumption 2 ensures that a "substitution effect" is not too strong, which we use to show that an auxiliary objective is quasi-concave and that its maximizer is well-behaved.

In the standard mechanism design approach, the principal can commit to unbounded transfers ($\tau = \infty$). Hence, we view as natural to assume that $\tau$ is sufficiently large to meet Assumptions 1 and 2. We discuss unbounded transfers as well as relaxations of Assumptions 1 and 2 in Section 6.

We also impose the following regularity condition on the agent's costs that holds, for example, if the first derivative $c'_A$ is convex.

**Assumption 3.** The function $e_A \mapsto e_A c'_A(e_A)$ is strictly convex in $e_A \in [0,1]$.

## 2.2 Incentives and profit

The principal faces a hidden information and a hidden action problem: the agent must be incentivized to advance the true surplus and to acquire evidence. The two problems are entangled: the refunds that the agent is promised for providing evidence give the agent additional incentives to advance their surplus. However, an agent who concealed some surplus will never acquire evidence (as the mechanism would confiscate the whole surplus as a punishment). Thus, "double deviations" are unimportant. We next describe the agent's incentives and the principal's profit in a fixed mechanism $m$.

**Hidden action: Evidence acquisition.** Since acquiring evidence is costly, the agent only acquires evidence that they plan to disclose. We henceforth take as given that the agent discloses acquired evidence.

When type $x$ truthfully advances $x$ and exerts effort $\tilde{e}_A$, their expected utility is the expected refund net of effort costs:

$$(1-\tilde{e}_A)\cdot(e_P(x)\cdot r_P(x)+(1-e_P(x))\cdot r_\emptyset(x))+\tilde{e}_A\cdot r_A(x)-c_A(\tilde{e}_A).$$

With probability $\tilde{e}_A$, only the agent provides evidence, then the refund is $r_A(x)$. With probability $(1-\tilde{e}_A)\cdot e_P(x)$, only the principal provides evidence, then the refund is $r_P(x)$. Otherwise the refund is $r_\emptyset(x)$.

The agent exerts the recommended effort $e_A(x)$ if and only if

$$e_A(x)\in\operatorname*{argmax}_{\tilde{e}_A\in[0,1]}\tilde{e}_A\cdot(r_A(x)-(e_P(x)\cdot r_P(x)+(1-e_P(x))\cdot r_\emptyset(x)))-c_A(\tilde{e}_A).$$

The *evidence rent* of type $x$ is the extra refund net of effort costs given $e_A(x)$:

$$\begin{aligned}\text{evidence rent of type } x =& e_A(x)\cdot(r_A(x)-(e_P(x)\cdot r_P(x)+(1-e_P(x))\cdot r_\emptyset(x)))\\ &-c_A(e_A(x)).\end{aligned}\tag{1}$$

In Appendix A.2, we show that optimally the evidence rent equals $u_A(e_A(x))$, where

$$u_A(e_A(x))=e_A(x)c_A'(e_A(x))-c_A(e_A(x)).\tag{2}$$

In what follows, we simply refer to $u_A(e_A(x))$ as the evidence rent from $e_A(x)$. Note $u_A(e_A(x))$ is strictly increasing in $e_A(x)$.

The expected utility $U_m(x)$ of type $x$ from advancing the surplus and exerting the recommended effort is thus given by

$$U_m(x)=e_P(x)r_P(x)+(1-e_P(x))r_\emptyset(x)+u_A(e_A(x)).$$

**Hidden Information: Advancing the full surplus.** If type $x$ advances a smaller amount $y\in[0,x)$, their expected utility is

$$e_P(y)\cdot(x-x)+(1-e_P(y))\cdot(x-y+r_\emptyset(y)).$$

With probability $e_P(y)$, the principal acquires evidence and confiscates the whole surplus $x$ as a punishment. Otherwise, the principal seizes $y$ and refunds $r_\emptyset(y)$.

It follows that type $x$ advances the full surplus if and only if

$$U_m(x) \geq \sup_{y \in [0,x)} (1 - e_P(y)) \cdot (x - y + r_\emptyset(y)). \tag{3}$$

For advance payments $y \in [0, \underline{x})$, we recall that the principal acquires evidence with probability one. Inspecting (3), it follows that no type has an incentive to advance a payment in $[0, \underline{x})$, and we henceforth ignore such payments. Similarly, we take the domain of all functions $e_A, e_P, r_A, r_P, r_\emptyset$ to be the type space $[\underline{x}, \bar{x}]$.

**The principal's profit.** The principal's profit $\Pi_m(x)$ from type $x$ is given by

$$\begin{aligned}\Pi_m(x) =& x - (1 - e_A(x)) \cdot (e_P(x) \cdot r_P(x) + (1 - e_P(x)) \cdot r_\emptyset(x)) - e_A(x) \cdot r_A(x) \\ & - (1 - e_A(x)) \cdot c_P(e_P(x)) \\ =& x - (e_P(x) \cdot r_P(x) + (1 - e_P(x)) \cdot r_\emptyset(x)) \\ & - u_A(e_A(x)) - c_A(e_A(x)) - (1 - e_A(x)) \cdot c_P(e_P(x)).\end{aligned}$$

The principal seizes the initial offer $x$ but reimburses the agent via $e_P(x) \cdot r_P(x) + (1 - e_P(x)) \cdot r_\emptyset(x)$ or via the evidence rent $u_A(e_A(x))$. Surplus is lost due to the agent's effort, $c_A(e_A(x))$, and the principal's effort, $(1 - e_A(x)) \cdot c_P(e_P(x))$.

**The role of the agent's evidence.** Fixing a type $x$, the agent's effort $e_A(x)$, on the one hand, decreases the principal's profit through the evidence rent $u_A(e_A(x))$ and through surplus destruction $c_A(e_A(x))$. On the other hand, $e_A(x)$ reduces the principal's on-path costs $(1 - e_A(x))c_P(e_P(x))$ and creates slack in the constraint (3) that type $x$ advances the full surplus. Note that $e_A(x)$ (or $r_A(x)$, the refund for type $x$'s presenting evidence) does not affect the incentives of any type other than $x$. Thus, by incentivizing $e_A(x)$ the principal provides targeted incentives to type $x$.

**Efficient agent-evidence.** Given that type $x$'s effort destroys surplus, $c_A(e_A(x))$, but also reduces the principal's effort costs, $(1-e_A(x))c_P(e_P(x))$, we define an *efficient agent-effort* level $e_A^{\text{eff}}(x)$. This effort $e_A^{\text{eff}}(x)$ minimizes total surplus destruction from evidence acquisition, taken the principal's effort $e_P(x)$ as given, i.e.

$$e_A^{\text{eff}}(x) \in \operatorname*{argmin}_{\tilde{e}_A \in [0,1]} c_A(\tilde{e}_A) + (1 - \tilde{e}_A) \cdot c_P(e_P(x)). \tag{4}$$

The minimizer is unique. The efficient agent-effort $e_A^{\text{eff}}$ is endogenous to the mechanism.

The principal's mechanism may optimally distort $e_A(x)$ away from the efficient level $e_A^{\text{eff}}$. Raising the incentivized agent-effort $e_A(x)$ raises the evidence rent $u_A(e_A(x))$, thereby also raising type $x$'s incentive for advancing the full surplus. Thus, the principal may distort $e_A(x)$ upwards or downwards away from $e_A^{\text{eff}}(x)$ depending on type $x$'s incentive for advancing the full surplus, as we discuss subsequently.

# 3 Tight mechanisms

For maximizing profit, it suffices to consider a subclass of mechanisms that we call *tight*. Roughly, a mechanism is tight if the principal cannot simultaneously extract a higher profit from every type and threaten every type with a higher loss from misreporting.

The loss from misreporting in a mechanism $m$ is captured by the *induced loss function* $\lambda_m \colon [\underline{x}, \bar{x}] \to \mathbb{R}_+$, defined for all $x \in [\underline{x}, \bar{x}]$ by

$$\lambda_m(x) = \inf_{y \in [\underline{x}, x]} e_P(y) \cdot x + (1 - e_P(y)) \cdot (y - r_\emptyset(y)). \tag{5}$$

Here, $\lambda_m(x)$ is the infimal loss for type $x$ from not advancing the full surplus, i.e. their best deviation.[2] For incentive compatibility, the on-path loss of type $x$ must not exceed $\lambda_m(x)$. Formally, the inequality (3) rearranges to:

$$x - (e_P(x) \cdot r_P(x) + (1 - e_P(x)) \cdot r_\emptyset(x)) - u_A(e_A(x)) \leq \lambda_m(x). \tag{6}$$

**Definition 3.1** (Tightness)**.** A mechanism $m^*$ is *tighter than* a mechanism $m$ if $(\Pi_m, \lambda_m) \leq (\Pi_{m^*}, \lambda_{m^*})$.[3] A mechanism $m^*$ is *tight* if $m^*$ is tighter than every mechanism $m$ that is tighter than $m^*$, i.e. $(\Pi_{m^*}, \lambda_{m^*}) \leq (\Pi_m, \lambda_m)$ only if $(\Pi_{m^*}, \lambda_{m^*}) = (\Pi_m, \lambda_m)$. (Note, by convention, both $m^*$ and $m$ mean IC mechanisms.)

The principal finds tight mechanisms optimal, as we show using Zorn's Lemma. In Section 3.3, we give a constructive proof and further intuition.

**Lemma 3.1.** *For all mechanisms $m$ there is a tight mechanism that is tighter than $m$.*

[2] In the infimum in (5) we also consider $y = x$. This is only to avoid taking the infimum over the empty set and has no other significance. Type $x$ does not contemplate deviating to themselves. Indeed, even if $y = x$, note that $e_P(y)x + (1 - e_P(y))(y - r_\emptyset(y))$ is not generally equal to type $x$'s on-path loss $e_P(x)(x - r_P(x)) + (1 - e_P(x))(x - r_\emptyset(x)) - u_A(e_A(x))$.

[3] For real-valued functions $g$ and $g^*$ we write $g \leq g^*$ to mean that $g(x) \leq g^*(x)$ holds for all $x$. Similarly, $(g, h) \leq (g^*, h^*)$ means that both $g \leq g^*$ and $h \leq h^*$ hold.

Before characterizing tight mechanisms, we distinguish tightness from other notions.

**Tightness is not Incentive Compatibility.** In our environment, the principal's ability to acquire evidence is a powerful tool for providing incentives. On the one hand, this tool allows the principal to implement all divisions of surplus by acquiring evidence with certainty; thus, incentive compatibility alone imposes little structure on a given mechanism. On the other hand, by acquiring evidence with high probability about a given type $x$, the principal can deter all deviations to $x$ (by threatening to seize everything) and simultaneously ensure type $x$'s incentives (by promising a large refund $r_P(x)$). In particular, the principal can perturb a given mechanism *only* at type $x$ without upsetting incentives of other types. The notion of tightness leverages these type-by-type perturbations to obtain structure on relevant mechanisms.

**Tightness is not Efficiency.** Border and Sobel (1987) and Chander and Wilde (1998) study *efficient* mechanisms in environments where only the principal can acquire evidence. Roughly, a mechanism is said to be efficient if there is no mechanism that extracts pointwise more surplus while exerting a pointwise lower principal-effort, and such that at least one of these inequalities is strict for at least one type. Efficiency is silent on the agent's effort, motivating a novel notion.

**Tightness is not Dominance.** A mechanism $m^*$ is *undominated* if there does not exist $m$ such that $\Pi_{m^*} \leq \Pi_m$ and $\Pi_{m^*} \neq \Pi_m$. Every undominated mechanism is tight, but there are tight mechanisms that are not undominated; see Appendix E.2.

**Tightness is not difficult.** Tight mechanisms are easy to find via a type-by-type algorithm that we provide in Section 3.3.

## 3.1 The class structure of tight mechanisms

In this subsection, we characterize tight mechanisms with random audits, i.e. the principal's effort $e_P(x)$ is interior for at least one type $x$. Such a mechanism admits a non-monotonic relationship between the principal's and agent's evidence acquisition effort. The mechanism divides types into seven endogenous intervals. In optimal tight mechanisms, we later find that the first interval is empty, the last interval contains only the highest type $\bar{x}$, but the five intermediate intervals all have non-empty interiors. We denote those five interior intervals in the order of their surplus levels by: *SuperLow,*

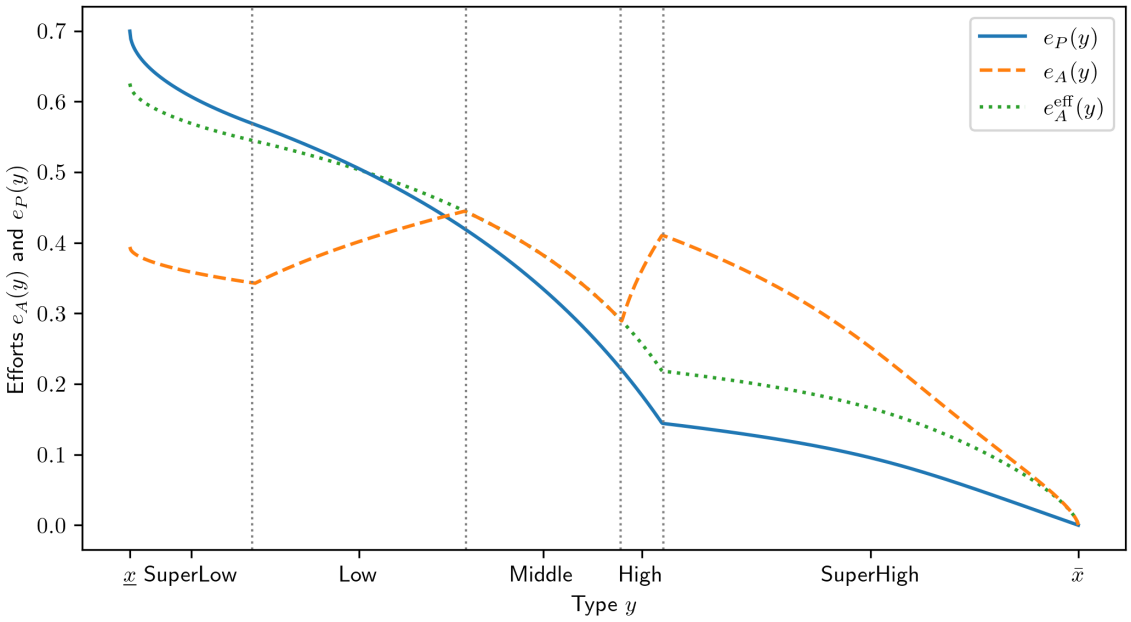


**Figure 1.** The efforts of a tight mechanism in which $e_P$ is interior except at $\bar{x}$. The type space is $[0,1]$, the costs are $c_A(e) = e^4$ and $c_P(e) = 2e^2$ for all $e \in [0,1]$, and the principal's funds are $\tau = 0.2$. Remark 1 at the end of Section 3.3 explains how to compute the efforts.

*Low, Middle, High, SuperHigh.* The intervals differ in the offered refunds and the agent's and principal's effort. Figure 1 illustrates the general behavior of the efforts: while the principal's effort $e_P$ is decreasing, the agent's effort $e_A$ is non-monotone.

The optimality of tight mechanisms is robust to the type distribution: the profit comparison holds type-by-type. Accordingly, the characterization is also robust.

**Theorem 3.1.** *Let $m$ be a tight mechanism with random audits. There exist five consecutive*[4] *intervals of types,* $\text{SuperLow}, \text{Low}, \text{Middle}, \text{High}, \text{SuperHigh}$, *that each have a non-empty interior, that partition the set of types where $e_P$ is interior (i.e. $\{y \in [\underline{x}, \bar{x}] \colon e_P(y) \in (0,1)\}$), and such that all of the following hold:*

*(1)* The principal's effort $e_P$ *is*

- *(a) constantly* 1 *on* $[\underline{x}, \inf(\text{SuperLow}))$*;*
- *(b) continuous except possibly jumping downwards at* $\inf(\text{SuperLow})$*;*
- *(c) decreasing on* $[\underline{x}, \bar{x}]$*;*
- *(d) strictly decreasing on* $\text{SuperLow} \cup \text{Low} \cup \text{Middle} \cup \text{High}$*;*
- *(e) possibly constant on subintervals of* SuperHigh*, but not constant over the*

[4] By "consecutive" we mean $\sup(\text{SuperLow}) = \inf(\text{Low})$ etc.

*whole interval;*

*(f) constantly* 0 *on* $[\sup(\text{SuperHigh}), \bar{x}]$.

*(2)* The agent's effort $e_A$ *is*

*(a) constant on* $[\underline{x}, \inf(\text{SuperLow}))$ *and equal to* $\operatorname{argmin}_{\tilde{e}_A \in [0,1]} u_A(\tilde{e}_A) + c_A(\tilde{e}_A) + (1 - \tilde{e}_A)c_P(1)$.[5]

*(b) continuous except possibly jumping downwards at* $\inf(\text{SuperLow})$*;*

*(c) non monotone; specifically, strictly decreasing on each of* SuperLow, Middle *and* SuperHigh*, but strictly increasing on each of* Low *and* High.

*(d) constantly* 0 *on* $[\sup(\text{SuperHigh}), \bar{x}]$.

*(e) strictly below* $e_A^{\text{eff}}$ *on the interval* $[\underline{x}, \inf(\text{Middle}))$*, equal to* $e_A^{\text{eff}}$ *on* Middle*, strictly above* $e_A^{\text{eff}}$ *on the interval* $(\sup(\text{Middle}), \sup(\text{SuperHigh}))$*, and equal to* $e_A^{\text{eff}}$ *on the interval* $[\sup(\text{SuperHigh}), \bar{x}]$.

*(3) The agent's utility* $U_m$ *is v-shaped: constant on* $[\underline{x}, \inf(\text{SuperLow})]$*, strictly decreasing on* SuperLow*, strictly increasing on* $[\sup(\text{SuperLow}), \bar{x}]$*. Moreover,* $U_m$ *is bounded away from* 0*, i.e.* $\inf_{y \in [\underline{x}, \bar{x}]} U_m(y) > 0$.

*(4) The agent's incentive to advance the full surplus is*

*(a) strict for* $y \in [\underline{x}, \sup(\text{SuperLow}))$*; specifically,* $U_m(y) > y - \lambda_m(y)$.

*(b) binding for* $y \in [\sup(\text{SuperLow}), \bar{x}]$*; specifically,* $U_m(y) = y - \lambda_m(y)$.

*(5) The principal's profit* $\Pi_m$ *is increasing.*

To explain the non-monotone pattern of the efforts, consider a tight mechanism $m$ with random-audits. Assume that $e_P$ is interior on $[\underline{x}, \bar{x})$; i.e., the five intervals SuperLow to SuperHigh cover $[\underline{x}, \bar{x})$, as in Figure 1. This assumption simplifies the exposition and applies to all optimal tight mechanisms, as we show later.

The principal's effort $e_P$ decreases in the type, i.e. the principal exerts less effort to acquire evidence after higher surplus reports; we provide intuition for this important property later. The agent's efficient effort $e_A^{\text{eff}}$, as defined in (4), acts as a complement to the principal's effort $e_P$. Thus, $e_A^{\text{eff}}$ also decreases in the type.

To incentivize the agent to acquire costly evidence (hidden action), the agent must expect a higher refund after providing evidence. Due to limited liability, incentivizing effort generates a strictly positive evidence rent $u_A$.

To incentivize the agent to advance their surplus truthfully (hidden information), the agent must receive sufficient rents from doing so. The evidence rent contributes to the rent from truthfully advancing the surplus. If a type $x$ has a binding incentive to

[5] Assumption 3 implies that this minimizer is unique.

advance the surplus, the information rent is $x - \lambda_m(x)$, i.e. the surplus minus the loss from the best deviation. Importantly, higher types have more to lose—thus, $\lambda_m(x)$ increases in $x$. But since the principal cannot seize more than the surplus, the loss grows more slowly than the surplus—thus, $x - \lambda_m(x)$ increases in $x$.

For super-low surplus types, the *implemented* agent effort $e_A$ is smaller than the efficient effort $e_A^{\text{eff}}$. To gain intuition, consider the lowest type $\underline{x}$. This type has nowhere to lie to and, therefore, does not require any information rent to advance their surplus truthfully. Hence, the principal cannot offset the evidence rent $u_A$ against this information rent. Higher super-low types $x$ also have strict incentives to advance truthfully. As a consequence, the principal incentivizes a lower effort than would be efficient, $e_A(x) < e_A^{\text{eff}}(x)$. Specifically, the implemented effort $e_A(x)$ minimizes surplus destruction plus evidence rent,

$$c_A(\tilde{e}_A) + (1 - \tilde{e}_A)c_P(e_P(x)) + u_A(\tilde{e}_A), \tag{7}$$

across $\tilde{e}_A \in [0, 1]$. For this objective, the optimal effort is co-monotone with $e_P(x)$. Thus, $e_A(x)$ also decreases in $x \in \text{SuperLow}$.

All types from Low onwards have a binding incentive to advance the full surplus. The mechanism must provide these types with rents for advancing their surplus. For types in Low, this information rent can be fully provided through the evidence rent $u_A$, thereby reducing the principal's effective costs of incentivizing agent-effort. Specifically, $u_A(e_A(x)) = x - \lambda_m(x)$ for all $x \in \text{Low}$. The information rent $x - \lambda_m(x)$ increases in $x$, and hence the evidence rent $u_A(e_A(x))$ also increases in $x$. Over the interval Low, the implemented effort $e_A$ thus increases. At the highest type in Low, the implemented effort reaches the efficient level $e_A^{\text{eff}}$.

For $x \in \text{Middle}$, the implemented effort $e_A(x)$ is efficient, i.e. minimizes surplus destruction

$$c_A(\tilde{e}_A) + (1 - \tilde{e}_A)c_P(e_P(x)) \tag{8}$$

across $\tilde{e}_A \in [0, 1]$. Thus, $e_A$ decreases co-monotonically with the principal's effort $e_P$. The increasing information rent $x - \lambda_m(x)$ is now also provided by refunds after a successful principal audit $r_P(x)$; specifically, $e_P(x)r_P(x) = x - \lambda_m(x) - u_A(e_A(x))$. The refund $r_P$ increase on Middle. The highest type in Middle receives the maximal reward after a successful audit, $r_P(x) = x + \tau$, a full refund of the advance payment $x$ and principal's private funds $\tau$.

For types $x$ in High, the principal always pays the maximal reward after a successful audit; i.e. $r_P(x) = x + \tau$. The rest of the information rent is now provided through the evidence rent, meaning $u_A(e_A(x)) = x - \lambda_m(x) - e_P(x)(x + \tau)$. We show that, consequently, the induced evidence effort $e_A(x)$ is increasing in $x$ and above the efficient level. At the highest type in this interval, the implemented evidence effort $e_A(x)$ is so costly to implement that the principal switches to a different instrument for providing the information rent: the no-evidence refund $r_\emptyset(x)$ which is paid if neither the agent nor the principal provide evidence.

For $x \in$ SuperHigh, the principal starts using the no-evidence refund $r_\emptyset(x)$ to contribute to the information rent. With a non-zero refund $r_\emptyset(x)$, there is a new interaction between the agent's and the principal's efforts. Fixing $x \in$ SuperHigh, suppose the principal increases the evidence rent $u_A(e_A(x))$ and reduces $r_\emptyset(x)$ to hold $x$'s on-path utility constant. The reduction of $r_\emptyset(x)$ reduces the incentives of other types to deviate to $x$. Thus, a reduced principal-effort $e_P(x)$ suffices to deter such types. In summary, increasing $e_A(x)$ lets the principal decrease $e_P(x)$. Even taking this additional benefit of incentivizing agent-effort $e_A$ into account, we show $e_A$ decreases on SuperHigh. Specifically, now $e_A(x)$ minimizes

$$c_A(\tilde{e}_A) + (1 - \tilde{e}_A)c_P(\beta(x, \tilde{e}_A)) \tag{9}$$

across $\tilde{e}_A \in [0, 1]$, where

$$\beta(x, \tilde{e}_A) = \sup_{z \in [x, \bar{x}]} \frac{\lambda_m(z) - \lambda_m(x) - u_A(\tilde{e}_A)}{z + \tau}. \tag{10}$$

The objective (9) is supermodular in $(x, \tilde{e}_A)$. Hence, the minimizer $e_A(x)$ *de*creases in $x$. The principal's effort is then given by $e_P(x) = \beta(x, e_A(x))$, interpreted below.

Finally, we explain why the principal's effort is decreasing. For types $x$ in SuperLow, Low, Middle, and High, the no-evidence refund $r_\emptyset(x)$ equals 0. Other types deviating to $x$ expect to lose everything when the principal acquires evidence, and lose $x$ otherwise. The principal's optimal effort choice is then given by $e_P(x) = \alpha(x)$, where

$$\alpha(x) = \sup_{z \in (x, \bar{x}]} \frac{\lambda_m(z) - x}{z - x}. \tag{11}$$

In words, $\alpha(x)$ is the smallest probability that, given $r_\emptyset(x) = 0$, ensures that all higher

types $z$ lose at least $\lambda_m(z)$ when deviating to $x$. The function $\alpha$ is decreasing (since $\lambda_m(z) \leq z$ for all $z$), meaning the principal's effort decreases.

For types $x \in$ SuperHigh, the principal's effort is given by $e_P(x) = \beta(x, e_A(x))$, the behavior of which is more subtle due to the interaction with the agent's effort. In words, $\beta(x, e_A(x))$ is the smallest probability that sustains $\lambda_m$ if $r_\emptyset(x)$ is chosen to ensure the incentives of type $x$ for the given agent-effort $e_A(x)$. More precisely, since $r_P(x) = x + \tau$ for $x \in$ SuperHigh, type $x$'s incentives require

$$x - e_P(x)(x + \tau) - (1 - e_P(x))r_\emptyset(x) - u_A(x) \leq \lambda_m(x).$$

Further, for all higher types $z$ the worst deviation $\lambda_m(z)$ is worse than deviating to $x$:

$$\lambda_m(z) \leq e_P(x)z + (1 - e_P(x))(x - r_\emptyset(x)).$$

These two conditions imply $e_P(x) \geq \beta(x, e_A(x))$. In fact, $e_P(x) = \beta(x, e_A(x))$, as we show. The loss $\lambda_m(x)$ is increasing in $x$, meaning $\beta(x, e_A(x))$ decreases in its first argument. However, since $e_A(x)$ decreases, there is a force pushing $\beta(x, e_A(x))$ up in $x$ (namely, as explained prior to (9), a lower agent-effort $e_A(x)$ necessitates a higher no-evidence refund $r_\emptyset(x)$ and hence a higher principal-effort $e_P(x)$). Assumption 2 ensures that this force is not too strong when $e_A$ is chosen optimally to maximize (9). Thus, $e_P(x)$ also decreases in $x$.[6]

## 3.2 Non-random audits and debt-with-relief

Here, we characterize tight mechanisms with non-random audits—$e_P(x) \in \{0, 1\}$ for all $x$—as *debt-with-relief mechanisms*. Such a mechanism is as in Theorem 3.1, except that the five intervals SuperLow, ..., SuperHigh are all empty.

**Definition 3.2** (Debt-with-relief)**.** A mechanism $m$ is a *debt-with-relief mechanism* if there is a *face value* $y_0 \in [\underline{x}, \bar{x}]$ and a *relief* $\bar{r}_A \in [0, \underline{x} + \tau]$ such that for all $x \in [\underline{x}, \bar{x}]$,

$$\begin{aligned} &e_P(x) = \mathbf{1}_{(x \in [\underline{x}, y_0))}, \\ &r_A(x) = \bar{r}_A \mathbf{1}_{(x \in [\underline{x}, y_0))}, \\ &x - (e_P(x)r_P(x) + (1 - e_P(x))r_\emptyset(x)) = \min(x, y_0). \end{aligned}$$

[6] On SuperHigh, the principal's effort $e_P$ is decreasing, but not necessarily strictly decreasing. See Appendix E.1 for an example.

(Therefore, also $r_P(x) = 0$ if $x < y_0$, and $r_\emptyset(x) = x - y_0$ if $x \geq y_0$.)

Advance payments above $y_0$ are never audited, and the principal seizes $y_0$ from each type above $y_0$. When the agent advances an amount less than $y_0$ and does not provide evidence, the principal audits with certainty and seizes everything. This is the payment structure of a classical "debt" contract, resembling the contracts that Townsend (1979) derived in a model in which the principal audits the agent non-randomly and the agent cannot provide evidence. By auditing with certainty in the case of a default—i.e. when the agent advances less than the face value $y_0$—, the principal deters types above $y_0$ from falsely defaulting. Compared to Townsend's contracts, the debt contract is augmented with a relief clause: in the case of a default, the agent gets a relief $\bar{r}_A$ for providing evidence for having defaulted. Figure 2 illustrates.

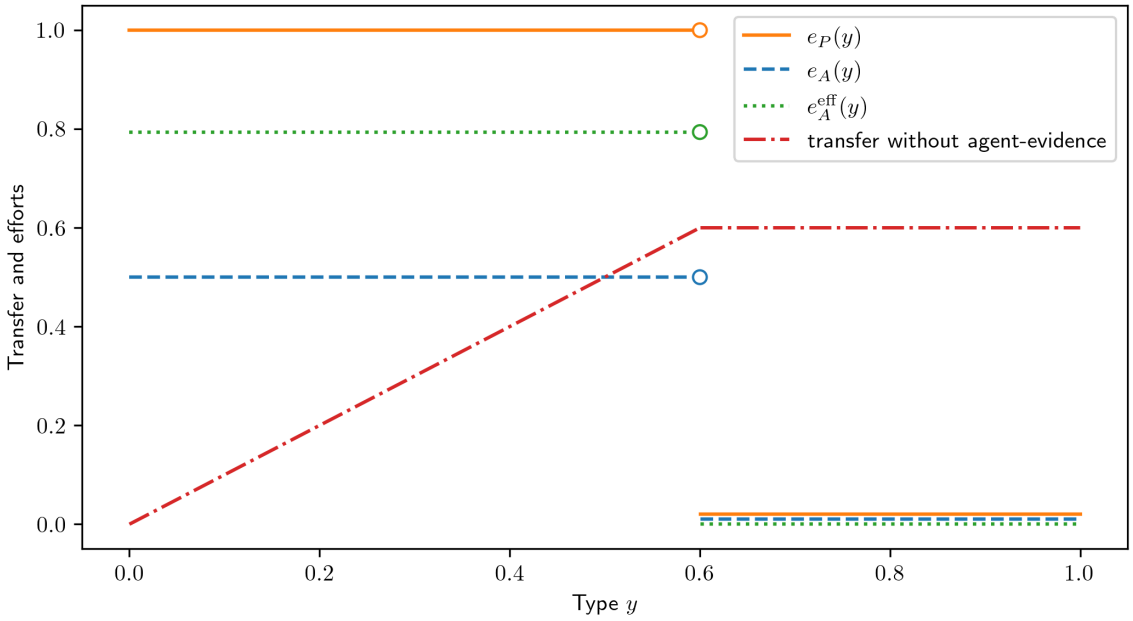


**Figure 2.** The transfer without agent-evidence $(y - e_P(y)r_P(y) - (1 - e_P(y))r_\emptyset(y))$, the efforts $e_A(y)$ and $e_P(y)$, and the efficient agent-effort $e_A^{\text{eff}}(y)$ as a function of the type $y$ in a debt-with-relief mechanism with face value $y_0 = 0.6$. The type space is $[0, 1]$, the costs are $c_A(e) = e^4$ and $c_P(e) = 2e^2$ for all $e \in [0, 1]$, and the principal's funds are $\tau = 0.2$.

To characterize the implemented agent effort, let $\bar{e}_A = \operatorname{argmin}_{\tilde{e}_A \in [0,1]} u_A(\tilde{e}_A) + c_A(\tilde{e}_A) + (1 - \tilde{e}_A)c_P(1)$ minimize the evidence rent plus total surplus destruction when the principal acquires evidence with certainty. Assumption 3 and the assumption $c_A'(0) = 0$ imply that the minimizer is unique and strictly positive.

**Theorem 3.2.** *If $m$ is a tight mechanism with non-random audits, then there is a face value $y_0 \in [\underline{x}, \bar{x}]$ such that $m$ is a debt-with-relief mechanism (with face value $y_0$) with relief given by $\bar{r}_A = c'_A(\bar{e}_A)$ and agent-effort $e_A$ given by $e_A(x) = \bar{e}_A \mathbf{1}(x \in [\underline{x}, y_0))$ for all $x \in [\underline{x}, \bar{x}]$.*

Agent-evidence lets the principal reduce the on-path auditing costs for types in $[\underline{x}, y_0)$ to $(1 - \bar{e}_A)c_P(1)$ compared to $c_P(1)$ in the debt contract of Townsend (1979). The implemented agent-effort $\bar{e}_A$ is inefficiently low; i.e. $\bar{e}_A < e_A^{\text{eff}}(x)$ for all $x \in [\underline{x}, y_0)$. This distortion arises for the same reason as the distortion for the super low types in tight mechanisms with random audits (Section 3.1): types in $[\underline{x}, y_0)$ earn an evidence rent $u_A(\bar{e}_A)$ that the principal cannot recover using other transfers since these types lose everything when not providing evidence.

## 3.3 How to tighten a mechanism?

We present an algorithm that turns an arbitrary given mechanism into a tight mechanism with a type-by-type higher profit. Except for two auxiliary definitions, the steps of the algorithm also proceed type-by-type.

Let $m$ be an arbitrary mechanism. The tightening algorithm proceeds in four steps, illustrated below in an example.

(1) For every type $x \in [\underline{x}, \bar{x}]$, define $\lambda^+(x) = \max(\underline{x}, \sup_{y \in [\underline{x}, x]} \lambda_m(y))$.

(2) For every type $x \in [\underline{x}, \bar{x}]$, choose $(\tilde{r}_A, \tilde{r}_P, \tilde{r}_\emptyset) \in [0, x+\tau]^3$ to minimize the payment $e_P(x)\tilde{r}_P + (1 - e_P(x))\tilde{r}_\emptyset$ subject to:

$$\lambda^+(x) \geq x - (e_P(x)\tilde{r}_P + (1 - e_P(x))\tilde{r}_\emptyset) - u_A(e_A(x)); \tag{12a}$$

$$\forall x' \in [x, \bar{x}], \quad \lambda^+(x') \leq e_P(x)x' + (1 - e_P(x))(x - \tilde{r}_\emptyset); \tag{12b}$$

$$c'_A(e_A(x)) = \tilde{r}_A - (e_P(x)\tilde{r}_P + (1 - e_P(x))\tilde{r}_\emptyset)\,. \tag{12c}$$

Denote the choice by $(\tilde{r}_A(x), \tilde{r}_P(x), \tilde{r}_\emptyset(x))$. In words, (12a) says that type $x$ loses at most $\lambda^+(x)$ from being truthful, (12b) says that higher types $x'$ lose at least $\lambda^+(x')$ from deviating to $x$, and (12c) says that $x$ finds the effort $e_A(x)$ optimal.

(3) Define the *virtual loss function* $\tilde{\lambda}$ for every $x \in [\underline{x}, \bar{x}]$ by:

$$\tilde{\lambda}(x) = \inf_{y \in [\underline{x}, \bar{x}]} e_P(y)x + (1 - e_P(y))(y - \tilde{r}_\emptyset(y)). \tag{13}$$

In words, $\tilde{\lambda}(x)$ is type $x$'s lowest deviation loss when the principal's effort is $e_P$, the no-evidence refund is $\tilde{r}_\emptyset$, and *all* advance payments are feasible for type $x$.

(4) For every type $x \in [\underline{x}, \bar{x}]$, choose $(e_A^*, e_P^*, r_A^*, r_P^*, r_\emptyset^*) \in [0,1]^2 \times [0, x+\tau]^3$ to maximize the profit

$$x - (e_P^* r_P^* + (1-e_P^*) r_\emptyset^*) - u_A(e_A^*) - c_A(e_A^*) - (1-e_A^*) c_P(e_P^*)$$

subject to

$$\tilde{\lambda}(x) \geq x - (e_P^* r_P^* + (1-e_P^*) r_\emptyset^*) - u_A(e_A^*); \tag{14a}$$

$$\forall x' \in [x, \bar{x}], \quad \tilde{\lambda}(x') \leq e_P^* x' + (1-e_P^*)(x - r_\emptyset^*); \tag{14b}$$

$$c_A'(e_A^*) = r_A^* - (e_P^* r_P^* + (1-e_P^*) r_\emptyset^*)\,. \tag{14c}$$

The constraints (14) are interpreted analogously to step (2). Denote the choice by $(e_A^*(x), e_P^*(x), r_A^*(x), r_P^*(x), r_\emptyset^*(x))$. This yields a new mechanism $m^*$.

**Theorem 3.3.** *Let $m$ be a mechanism. If $m^*$ is obtained through steps (1) to (4) of the tightening algorithm applied to $m$, then $m^*$ is tight and tighter than $m$.*

In particular, $m^*$ has a type-by-type higher profit than $m$, and $m^*$ is characterized by Theorem 3.1 or 3.2.

**Example 1.** To illustrate the tightening algorithm, consider the following modified debt contract $m$. For some face value $y_0$, when the agent advances less than $y_0$, the principal acquires evidence with certainty and does not refund anything. When the advanced surplus $x$ exceeds the face value $y_0$, the principal refunds the difference $x - y_0$ and adds $\tau$ as a premium without evidence. The mechanism does not incentivize any agent-effort. Thus, $m$ departs from a debt-with-relief mechanism by failing to incentivize types below $y_0$ to provide evidence, and by giving away $\tau$ to types above $y_0$. Formally, $m$ is given for all types $x$ by

$$m(x) = \begin{cases} e_A(x) = 0,\ e_P(x) = 1,\ r_A(x) = r_P(x) = r_\emptyset(x) = 0, & \text{if } x \leq y_0, \\ e_A(x) = e_P(x) = r_A(x) = r_P(x) = 0, \quad r_\emptyset(x) = x - y_0 + \tau, & \text{if } x > y_0, \end{cases}$$

and the induced loss $\lambda_m(x)$ is given by

$$\lambda_m(x) = \begin{cases} x & \text{if } x \leq y_0; \\ y_0 - \tau & \text{if } x > y_0. \end{cases}$$

The tightening algorithm corrects the departures of $m$ from debt-with-relief. In step (1), for each type $x$ we construct $\lambda^+(x)$ as the running maximum loss across types below $x$. In this example, $\lambda^+(x) = \min(x, y_0)$ for all $x$.

In step (2), at each type $x$ we re-optimize the refunds subject to the constraints that type $x$ loses at most $\lambda^+(x)$ from being truthful, every higher type $x'$ loses at least $\lambda^+(x')$ from deviating to $x$, and type $x$ still finds $e_A(x)$ optimal. In this example, this re-optimization only entails removing the premium if the agent advances more than $y_0$; i.e. for all $x$ above $y_0$ setting $\tilde{r}_\emptyset(x) = x - y_0$ (down from $r_\emptyset(x) = x - y_0 + \tau$).

In step (3), we define the virtual loss function $\tilde{\lambda}$. In this example, $\tilde{\lambda} = \lambda^+$. In general, $\tilde{\lambda}$ is an increasing concave function such that $\tilde{\lambda}(\underline{x}) = \underline{x}$ and $\lambda_m(x) \leq \tilde{\lambda}(x) \leq x$ for all $x$. Step (4) will construct a mechanism $m^*$ that has the virtual loss $\tilde{\lambda}$ as the induced loss $\lambda_{m^*}$. In $m^*$, higher types face higher losses from deviating (i.e. $\tilde{\lambda}$ increases), the agent deviates to minimize their loss (i.e. $\tilde{\lambda}$ is concave), and the lowest type loses everything when deviating (i.e. $\tilde{\lambda}(\underline{x}) = \underline{x}$).

In step (4), we re-optimize all parts of the mechanism type-by-type. For each type $x$, we choose $(e_A^*, e_P^*, r_A^*, r_P^*, r_\emptyset^*)$ to maximize the profit at $x$ subject to: type $x$ loses at most $\tilde{\lambda}(x)$ from being truthful, every higher type $x'$ loses at least $\tilde{\lambda}(x')$ from deviating to $x$, and type $x$ finds $e_A^*$ optimal. In this example, step (4) entails incentivizing agent-effort for types below $y_0$. The result is the debt-with-relief mechanism with face value $y_0$ described by Theorem 3.2.

The proof of Theorem 3.3 involves two steps. First, one can easily check that the algorithm weakly increases the profits and increases the deviation loss type-by-typ; i.e. the algorithm tightens the input. The more difficult second step shows that the output mechanism $m^*$ cannot be further tightened and is therefore tight. The increased induced loss of some type in $m^*$ makes it impossible to increase profits at another type by reapplying the tightening algorithm a second time. The re-optimization of the refunds and the choice of the virtual loss (steps (1) to (3)) rule out this possibility.

**Remark 1.** The tight mechanism from Figure 1 obtains by applying step (4) of the tightening algorithm to $\tilde{\lambda}$ given by $\tilde{\lambda}(x) = (1+x)^{0.7} - 1$ for all $x \in [0, 1]$. Generally, a

tight mechanism $m^*$ obtains by applying step (4) to an arbitrary increasing concave function $\tilde{\lambda}$ such that $\tilde{\lambda}(\underline{x}) = \underline{x}$ and $\tilde{\lambda} \leq \mathrm{id}$; see Theorem C.1 in Appendix C. In this case, $m^*$ has non-random audits if and only if there is $y_0$ such that $\tilde{\lambda}(x) = \min(x, y_0)$ for all $x$; see Theorem B.2 in Appendix B.5.

# 4 Optimal mechanisms

The expected profit of a mechanism $m$ is $\int \Pi_m(x)\,\mathrm{d}F(x)$, where $F$ denotes the type distribution.[7] A mechanism is *optimal* if it maximizes the expected profit across all mechanisms. There is an optimal mechanism that is tight, and all optimal mechanisms are "essentially tight;" see Appendix D.1 for a precise statement.

Optimality sheds light on the trade-off across types. The trade-off is between the principal's effort costs at a type $y$ and the surplus that the principal extracts from types who contemplate deviating to $y$. Tightness is silent on this trade-off since "tighter than" insists that profits increase at all types simultaneously. Before analyzing this trade-off, we show that the principal optimally leaves rents and audits randomly.

## 4.1 Positive rents and random audits are optimal

**Theorem 4.1.** *Let $F$ be continuous at $\bar{x}$.*[8] *Let $m$ be tight and optimal. All types of the agent enjoy a strictly positive utility from advancing the surplus—*$\inf_{x\in[\underline{x},\bar{x}]} U_m(x) > 0$*. The principal's effort $e_P$ is bounded away from one—*$\sup_{x\in[\underline{x},\bar{x}]} e_P(x) < 1$*—, and is strictly positive except at $\bar{x}$.*

Thus, the interval characterization of tight mechanisms (Theorem 3.1) extends to optimal mechanisms, and no debt-with-relief mechanism is optimal. Continuity of $F$ at $\bar{x}$ is a minimal richness condition—there are types strictly below but close to $\bar{x}$.

In the proof, we first show that no type loses everything when advancing the full surplus. If the highest type $\bar{x}$ lost everything, then for all types just below $\bar{x}$ the principal must acquire evidence with certainty to deter type $\bar{x}$ from deviating. But since $F$ is continuous at $\bar{x}$, it is suboptimal to separate $\bar{x}$ from types just below $\bar{x}$

[7] Here, we tacitly restrict attention to mechanisms $m$ with Borel-measurable profit $\Pi_m$. This restriction has no economic substance because for every mechanism (measurable or not) there is a tight one with a type-by-type higher profit (Lemma 3.1), and because all tight mechanisms have a measurable profits (as one may verify using the tightness characterizations, Theorems 3.1 and 3.2).

[8] That is, viewing $F$ as a cummulative distribution function, it holds $\lim_{\varepsilon\searrow 0} F(\bar{x}-\varepsilon) = 1$.

at the cost of acquiring evidence with certainty. Next, for every type $x$ except $\bar{x}$, the assumption $c_P'(0) = 0$ implies that the principal optimally exerts non-zero effort $e_P(x)$ to extract more from higher types who contemplate deviating to $x$. Given $e_P(x) > 0$, the assumption $c_A'(0) = 0$ implies that the principal also incentivizes non-zero agent-effort $e_A(x)$ to reduce the costs $(1 - e_A(x))e_P(x)$. In particular, type $x$ has a non-zero evidence rent $u_A(e_A(x))$. Finally, since no type loses everything when advancing the full surplus, it is excessive for the principal to acquire evidence with certainty and threaten to seize everything when the evidence detects a deviation.

Continuity of $F$ at $\bar{x}$ is important in Theorem 4.1. Suppose instead that types are binary. If the high type $\bar{x}$ is much larger and likelier than the low type, then the principal may optimally seize everything from $\bar{x}$, even at the cost of acquiring evidence about the low type with certainty.

## 4.2 The trade-off across types

How does the principal trade-off effort costs at a type $y$ with the surplus seized from the types $x$ who contemplate deviating to $y$? We begin by defining the binding incentive constraints. For the remainder of this subsection, fix a tight optimal mechanism $m$.

**Binding ICs.** Given types $y$ and $x$ such that $y \leq x$, we say $x$ *is a binding IC type of* $y$ if the best deviation of $x$ is to $y$,[9] i.e. if

$$\lambda_m(x) = e_P(y)x + (1 - e_P(y))(y - r_\emptyset(y)).$$

The upcoming Theorem 4.2 shows that, for every given optimal tight mechanism $m$, there is a continuous strictly increasing function $\hat{x}$ such that for every $y \in (\underline{x}, \bar{x})$ type $\hat{x}(y)$ is a binding IC type of type $y$. We refer to $\hat{x}$ as a *binding IC selection (for* $m$*)*. To gain intuition for why $\hat{x}$ is increasing, recall that the principal seizes everything when detecting a deviation. Since higher types have more to lose, the principal's effort has a screening effect and more effectively deters high types than low types. The principal's effort $e_P(y)$ is decreasing in $y$, and hence $\hat{x}(y)$ is increasing; strict increasingness is more subtle.

Further, the binding incentives are non-local: all types $y \in (\underline{x}, \bar{x})$ satisfy $y < \hat{x}(y)$.

[9] We only consider $x$ and $y$ such that $y \leq x$ since the agent can only advance less than the surplus. In a tight mechanism, no type would advance more than the surplus, even if this were possible.

Intuitively, since the agent risks everything by deviating, worthwhile deviations must be to distant types where the agent saves a lot of surplus if undetected.

Henceforth, we fix such a binding IC selection $\hat{x}$ and describe possible perturbations of the mechanism $m$. To simplify the exposition, let us assume that the binding ICs are *doubly unique*, by which we mean that (i) each type $y \in (\underline{x}, \bar{x})$ has $\hat{x}(y)$ as a unique binding IC type, and that (ii) $y$ is the unique type who has $\hat{x}(y)$ as a binding IC type.[10] Our characterization will not assume double-uniqueness, as explained below.

**Direct and indirect impacts.** To study the principal's trade-offs between a type $y$ and $y$'s binding IC type $\hat{x}(y)$, we perturb the given mechanism $m$ at $y$ and $\hat{x}(y)$. The principal increases $e_A(\hat{x}(y))$ (by increasing the refund $r_A(\hat{x}(y))$), thereby reducing the incentive of $\hat{x}(y)$ to deviate to $y$. Thus, the principal can decrease $e_P(y)$, perturbing some of $e_A(y)$, $r_A(y)$, $r_P(y)$, and $r_\emptyset(y)$ to hold type $y$'s on-path utility constant.

Under the assumption of double-uniqueness, this perturbation captures the trade-off between $y$ and $\hat{x}(y)$ in isolation from other types and their binding IC types. First, only type $\hat{x}(y)$ can enjoy the refund $r_A(\hat{x}(y))$ for agent-evidence. Second, perturbing $e_P(y)$ has no first-order impact on types other than $\hat{x}(y)$ since $\hat{x}(y)$ is $y$'s unique binding IC type. Third, perturbing $\hat{x}(y)$'s on-path payoff has no first-order impact on $\hat{x}(y)$'s incentive to deviate to a type other than $y$ since $y$ is the unique type which has $\hat{x}(y)$ as a binding IC type.

Without the assumption of double-uniqueness, we show that it is possible to construct a binding IC selection $\hat{x}$ such that it is *as if* the principal were solving the trade-off between $y$ and $\hat{x}(y)$ in isolation from other types.

We decompose the overall impact of the perturbation into a *direct* and an *indirect impact* on the burden of proof. The direct and indirect impact, respectively, refer to the change of the total evidence costs at $\hat{x}(y)$ and $y$, respectively. Specifically, for all $y \in (\underline{x}, \bar{x})$, we define the *direct impact* $D(y)$ and the *indirect impact* $I(y)$ as:[11]

$$D(y) = \frac{c'_A(e_A(y)) - c_P(e_P(y))}{u'_A(e_A(y))}, \tag{15a}$$

[10] In fact, since $\hat{x}$ is strictly increasing, (i) implies (ii).

[11] To see that these terms are well-defined, recall from Section 4.1 that in a tight optimal mechanism the five intervals described by Theorem 3.1 cover $[\underline{x}, \bar{x})$. In (15a), we can divide by $u'_A(e_A(y))$ since $u'_A(e_A(y))$ is non-zero if $e_A(y)$ is non-zero, which is the case for $y \in (\underline{x}, \bar{x})$. In (15b), we can divide by $\hat{x}(y) - y$ or $\hat{x}(y) + \tau$ since $y < \hat{x}(y)$ holds for all $y \in (\underline{x}, \bar{x})$, and since $0 < \underline{x} + \tau$ holds by assumption.

$$I(y) = \begin{cases} \dfrac{(1-e_A(y))c_P'(e_P(y))}{\hat{x}(y)-y}, & \text{if } y \in \text{SuperLow} \cup \text{Low} \cup \text{Middle}; \\ \dfrac{(1-e_A(y))c_P'(e_P(y))}{\hat{x}(y)-y} - \dfrac{y+\tau}{\hat{x}(y)-y}D(y), & \text{if } y \in \text{High}; \\ \dfrac{(1-e_A(y))c_P'(e_P(y))}{\hat{x}(y)+\tau}, & \text{if } y \in \text{SuperHigh}. \end{cases} \tag{15b}$$

Both $D$ and $I$ depend on the mechanism $m$ and the binding IC selection $\hat{x}$.

To interpret the impacts, recall that the perturbation entails an increase of $e_A(\hat{x}(y))$ that increases the on-path utility of type $\hat{x}(y)$ by \$1, a decrease of $e_P(y)$, and a perturbation of the other parts of the mechanism at $y$ to fix $y$'s on-path utility.

Increasing $e_A(\hat{x}(y))$ represents \$1 of surplus that falls to $\hat{x}(y)$ plus the impact $D(\hat{x}(y))$, where $D(\hat{x}(y))$ is interpreted as follows. Increasing $e_A(\hat{x}(y))$ destroys some surplus through the agent's effort, $c_A'(e_A(\hat{x}(y))$, but restores some surplus by decreasing on-path at type $\hat{x}(y)$ the principal's effort costs, $-c_P(e_P(\hat{x}(y)))$. The normalization via $u_A'(e_A(\hat{x}(y)))$ identifies the correct rate of change when the decrease of $e_A(\hat{x}(y))$ decreases the on-path utility of $\hat{x}(y)$ by \$1.

The indirect impact $I(y)$ is more delicate and depends on which of the five intervals contains $y$. In all five intervals, the impact entails the costs $(1-e_A(y))c_P'(e_P(y))$ from increasing $e_P(y)$. The normalization, either $\hat{x}(y)-y$ or $\hat{x}(y)+\tau$, identifies the correct change of $e_P(y)$ when the on-path utility of $\hat{x}(y)$ increases by \$1. To hold $y$'s on-path utility constant for $y$ in $\text{SuperLow} \cup \text{Low} \cup \text{Middle}$, the principal perturbs $r_P(y)$. On SuperHigh, we have $r_P(y) = y + \tau$ fixed and the principal instead perturbs $r_\emptyset(y)$; perturbing $r_\emptyset(y)$, of course, impacts the incentives of others to falsely advance $y$, but this is accounted for by the perturbation of $e_P(y)$, which explains why the normalization on SuperHigh is $\hat{x}(y)+\tau$. For $y$ in High, however, the principal holds $y$'s on-path utility constant by perturbing $e_A(y)$ (by perturbing $r_A(y)$). Perturbing $e_A(y)$, in turn, has the impact of $D(y)$ on the evidence production costs at $y$, interpreted just like $D(\hat{x}(y))$ but for type $y$; the factor $\frac{y+\tau}{\hat{x}(y)-y}$ identifies the correct change of $e_A(y)$ that holds $y$'s on-path utility constant.

**Characterization.** The perturbation marginally decreases profits at $\hat{x}(y)$ by $1 + D(\hat{x}(y))$, but marginally increases profits at $y$ by $I(y)$. Optimally, these impacts are balanced, as confirmed by (16a) in the following theorem.

**Theorem 4.2.** *Let the type distribution $F$ be absolutely continuous and have full*

*support. Let $m$ be tight and optimal. There is a continuous strictly increasing binding IC selection $\hat{x}\colon (\underline{x},\bar{x}) \to (\underline{x},\bar{x})$ for $m$ such that, for all $y \in (\underline{x},\bar{x})$, type $\hat{x}(y)$ is a binding IC type of $y$, and $y < \hat{x}(y)$ holds. Defining $D$ and $I$ as in (15) for $m$ and $\hat{x}$, it holds:*

$$\forall [a,b] \subset (\underline{x},\bar{x}), \quad \int_{[a,b]} I(y)\,\mathrm{d}F(y) = \int_{[\hat{x}(a),\hat{x}(b)]} (1+D(x))\,\mathrm{d}F(x); \tag{16a}$$

$$\forall x \in \text{SuperLow}, \quad D(x) = -1; \tag{16b}$$

$$\forall x \in \text{Middle}, \quad D(x) = 0; \tag{16c}$$

$$\forall x \in \text{SuperHigh}, \quad D(x) = I(x). \tag{16d}$$

In addition to the perturbation that draws out the trade-off across types, there are other perturbations affecting only the profit at a single type, yielding (16b), (16c) and (16d). These are the respective first-order conditions of the objectives (7), (8) and (9) explained in the context of tight mechanisms.

We briefly comment on the uniqueness of the binding ICs, and how the binding ICs relate to the principal's effort $e_P$ and the induced loss function $\lambda_m$. For each type $y \in (\underline{x},\bar{x})$, one can show that $\hat{x}(y)$ is type $y$'s unique binding IC type whenever there is no other type $z$ such that $e_P(y) = e_P(z)$. Since $e_P$ is strictly decreasing on the interval $\text{SuperLow} \cup \ldots \cup \text{SuperHigh}$, it follows that the binding ICs are doubly unique on this interval. On SuperHigh, however, $e_P$ may be constant on non-degenerate subintervals $[a,b]$. On such a subinterval $[a,b]$, all types have the same set of the binding IC types, namely an interval $[\hat{a},\hat{b}]$. For types in $[\hat{a},\hat{b}]$, the induced loss function $\lambda_m$ is affine, with slope constantly equal to the constant value of $e_P$ on $[a,b]$. In the proof of Theorem 4.2, we make a particular continuous strictly increasing selection from this set of binding IC types. Specifically, we approximate the optimal tight mechanism $m$ via tight mechanisms in which all types have a unique binding IC types, and then pass to the limit to obtain a candidate selection for $m$.

# 5 Related literature

We study the canonical model of surplus division with verification introduced by Townsend (1979). Among the subsequent literature, the closest to us is Border and Sobel (1987) and Chander and Wilde (1998). Unlike these papers and other subsequent work (e.g., Gale and Hellwig (1985), Monnet and Quintin (2005), Mookherjee and Png

(1989), Popov (2016), Ravikumar and Zhang (2012), Wang (2005)) we assume that both the principal and the agent can acquire evidence. The agent's evidence acquisition is a hidden action problem that is entangled with the original hidden information problem. The agent's hidden action problem changes how the principal acquires evidence in the hidden information problem. For example, we show that optimally the principal never acquires evidence deterministically, whereas Border and Sobel (1987), Chander and Wilde (1998) cannot rule out deterministic optimal mechanisms.[12] Moreover, these papers consider linear evidence technologies. We generalize by introducing a flexible Dye (1985) model in which we identify the success probability with the effort and effort costs are non-linear (and which can well-approximate linear evidence technologies).

The model of Allingham and Sandmo (1972) is the basis of much empirical work on tax compliancy. Our mechanism design approach departs from Allingham-Sandmo, but we make a conceptual contribution: incentivizing evidence acquisition by the agent strengthens the agent's incentive to reveal the surplus. Since the Allingham-Sandmo model tends to over-predict non-compliancy[13] and does not model evidence acquisition by the agent, it is interesting to ask if distinguishing agent- and principal-evidence also enhances compliancy in their framework.[14]

Unlike in classical hidden action problems (Holmström (1979)), our agent's hidden action has no exogenous benefit; its benefit from revealing the surplus is endogenous to the mechanism. Accordingly, our entangled hidden action and hidden information problems cannot obviously be "decoupled" in the sense of Castro-Pires et al. (2024).

Ben-Porath et al. (2023) also study agent-evidence-acquisition as a hidden action problem with multiple agents. In contrast to us, their evidence acquisition is deterministic and the principal cannot acquire their own evidence.

The question of who should acquire costly evidence arises in many problems other than surplus division. One strand of the literature, the literature on verification,

[12] Boyd and Smith (1994) quantify the welfare effects of random versus deterministic principal-evidence acquisition in a setting close to Border and Sobel (1987). Our results suggest that agent-evidence may improve welfare via lower total effort costs, at least for some surplus levels.

[13] For example, Andreoni et al. (1998) write: "The most significant discrepancy that has been documented between the standard economic model of compliance and real-world compliance behavior is that the theoretical model greatly overpredicts noncompliance."

[14] In our set-up, the agent's evidence rent has a similar effect on the agent's incentives as non-monetary rewards from compliance, sometimes referred to as "tax morale" (Luttmer and Singhal, 2014). However, the evidence rent does represent a monetary cost for the principal. Our model better captures tax morale via the agent's costs for acquiring evidence.

focuses on evidence acquisition by a player with commitment power (the principal).[15] Another strand is focused on settings (mechanism design problems and games) in which players without commitment power (agents) can acquire or present evidence at a cost.[16] We combine these two branches to analyze how the burden of proof is optimally shared. There are only few papers that allows counter-parties to provide verifiable information. Bester et al. (2021) study signaling à la Spence (1973) when the firms without commitment can acquire evidence about the worker's type. Stahl and Strausz (2017) compare the outcomes in a lemons market with costly buyer- or seller-evidence-acquisition. Lichtig and Mass (2024) allow the receiver in a disclosure game to publicly design a private signal before the disclosure.

Palonen and Pekkarinen (2022) study taxation with principal-verification. In contrast to us, the agent's effort reduces the chance of being detected by the principal.

Ben-Porath et al. (2019) relate a class of mechanism design problems with costly principal-evidence to Dye-disclosure games. By contrast, we show that in our set-up costly agent- and principal-evidence are not substitutable.

Our advance payment is reminiscent of Celik (2006) and Strausz and Krähmer (2024), who study one-dimensional screening when the agent can misreport in one direction only.

# 6 Discussion

**Funds for incentivizing agent-effort.** The assumption $\underline{x} + \tau > c'_A(1)$ (Assumption 1) guarantees that the principal has sufficient private funds $\tau$ to incentivize agent-effort. To understand the effect of dropping this assumption, consider the maximization problem in step (4) of the tightening algorithm at a given type $y$ and virtual loss function $\tilde{\lambda}$. As our proofs show, the choice of agent-effort $e_A(y)$ in this problem maximizes an auxiliary quasi-concave objective, but is constrained by an

[15] See, e.g. Ahmadzadeh (2024), Ahmadzadeh and Waizmann (2024), Ball and Knoepfle (2024), Ball and Pekkarinen (2024), Ben-Porath et al. (2014), Beshkar and Bond (2017), Brzustowski and Erlanson (2024), Chen et al. (2022), Epitropou and Vohra (2019), Erlanson and Kleiner (2019, 2020, 2024), Halac and Yared (2020), Hu (2024), Kaplow (2011a,b), Kattwinkel and Knoepfle (2023), Khalfan (2023), Khalfan and Vohra (2024), Li (2020, 2021), Li and Libgober (2023), Malenko (2019), Patel and Urgun (2022), Pham (2024), Siegel and Strulovici (2023).

[16] For costly disclosure, see e.g. Bull (2008a,b), Jovanovic (1982), Kartik and Tercieux (2012), Madarasz and Pycia (2023), Perez-Richet and Skreta (2024), Verrecchia (1983). In Asseyer and Weksler (2024), Ben-Porath et al. (2023), Jiang (2024), Pram (2023), Preusser (2022), and Whitmeyer and Zhang (2022) the agents also learn from the evidence about an underlying state.

upper bound $\bar{e}_A(y)$ that increases in $y$. When the principal's funds $\tau$ are too small, this upper bound may bind. As illustrated by Figure 3 in an example, the principal chooses $e_A(y)$ (in the figure, the solid orange line) as the minimum of $\bar{e}_A(y)$ (dashed green) and the unconstrained optimum (i.e., if the principal's funds sufficed to incentive any agent-effort; dotted black).

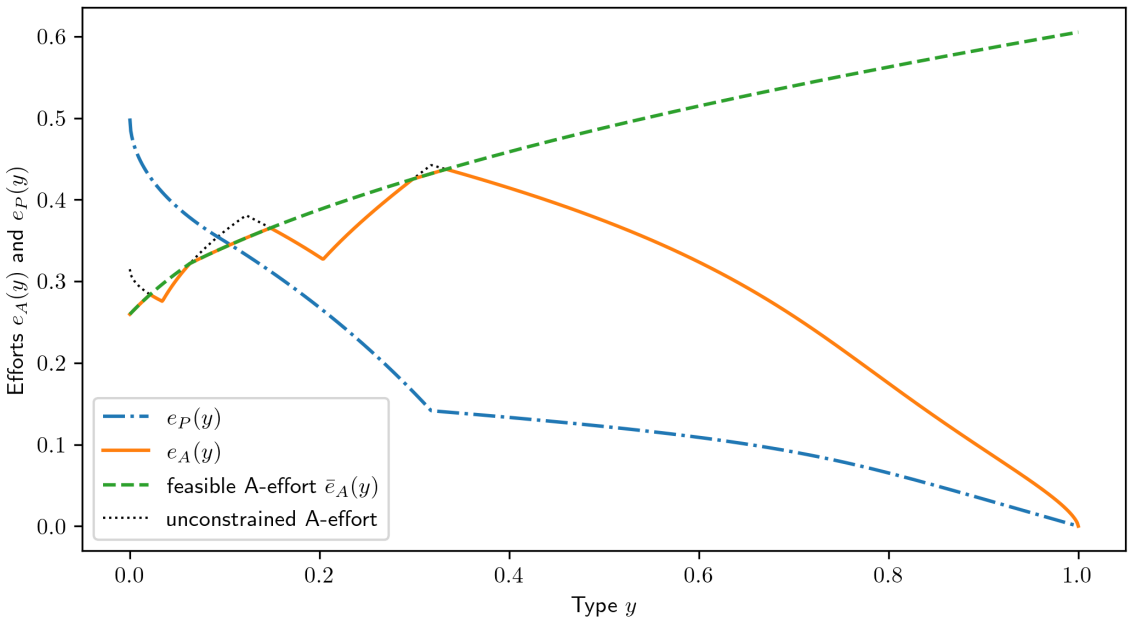


**Figure 3.** Effort levels in a mechanism obtained via step (4) of the tightening algorithm and where the constraint $r_A(y) \leq y + \tau$ binds for some types $y$. The type space is $[0, 1]$, the costs are $c_A(e) = e^4$ and $c_P(e) = 2e^2$ for all $e \in [0, 1]$, and the principal's private funds are $\tau = 0.07$. We applied step (4) to $\tilde{\lambda}$ given by $\tilde{\lambda}(y) = \sqrt{1 + x} - 1$ for all $x \in [0, 1]$.

Assumption 1 also entails $c'_A(1) > c_P(1)$, implying that the principal finds it suboptimal to incentivize the agent to acquire evidence with certainty. Without the assumption $c'_A(1) > c_P(1)$, there may be subintervals where agent-effort is constantly 1 (rather than strictly decreasing or strictly increasing). If agent-effort equals 1, the perturbation of agent-effort that we used to characterize the trade-offs across types is feasible in only one direction. Thus, $c'_A(1) > c_P(1)$ is convenient to assume.

**Regularity assumptions on effort costs.** Assumption 2 has two roles. Perhaps most importantly, we use the assumption to prove that in a tight mechanism with random audits the principal's effort $e_P$ decreases on SuperHigh. The assumption also implies that, for each type $y$, the principal chooses $e_A(y)$ to maximize an auxiliary

*quasi-concave* objective, and that $e_A$ and $e_P$ are continuous whenever $e_P$ is interior. Without Assumption 2, there conceivably are points at which $e_A$ jumps discontinuously downwards and $e_P$ discontinuously upwards. However, Assumption 2 is not needed to prove the single-crossing properties that we use to show that there are exactly five intervals on which the sign of the slope of $e_A$ changes.

**Co-ordinated evidence acquisition.** We have shown that, optimally, the principal tries to acquires evidence only after the agent fails to do so. In some situations, however, it may be natural that the principal and the agent cannot co-ordinate their efforts, resulting in simultaneous effort choices. For example, suppose the principal and the agent only meet twice: when the agent advances a payment, and, later, when the two present their evidence before a court that enforces transfers.

With simultaneous effort choices, the principal's on-path costs from a type $x$ are $c_P(e_P(x))$ (rather than $(1-e_A(x))c_P(e_P(x))$ in the baseline set-up). The profit is

$$x - (e_P(x)r_P(x) + (1-e_P(x))r_\emptyset(x)) - u_A(e_A(x)) - c_A(e_A(x)) - c_P(e_P(x)).$$

In the baseline set-up, there are two motives for incentivizing agent-effort: reducing the on-path costs, and using the evidence rent to solve the hidden information problem. With simultaneous effort, only the second motive remains: we conjecture that the incentivized agent-effort optimally equals 0 for an interval of low types, then increases (as previously on High), and then decreases (as previously on SuperHigh). Figure 4 confirms this conjecture in an example.

**One-sided evidence.** Border and Sobel (1987) and related work study the problem where only the principal can acquire evidence; we discuss the differences in Section 5. If the principal cannot acquire evidence, the problem is uninteresting: the agent advances a surplus of zero without providing evidence; since only the revealed surplus is contractible, the principal must accept this payment.

**Other evidence structures.** Our model assumes that evidence reveals the exact difference between the full surplus $x$ and the advance payment $y$. Keeping this structure of principal-evidence, suppose that the agent's effort stochastically reveals an amount $d \in [0, x-y]$. The contractible surplus is $y+d$ (unless the principal acquires evidence). If the agent advanced the surplus and acquires evidence, i.e. $x = y$, then necessarily $d = 0$, so that the surplus is revealed like in the baseline model. If $x \neq y$,

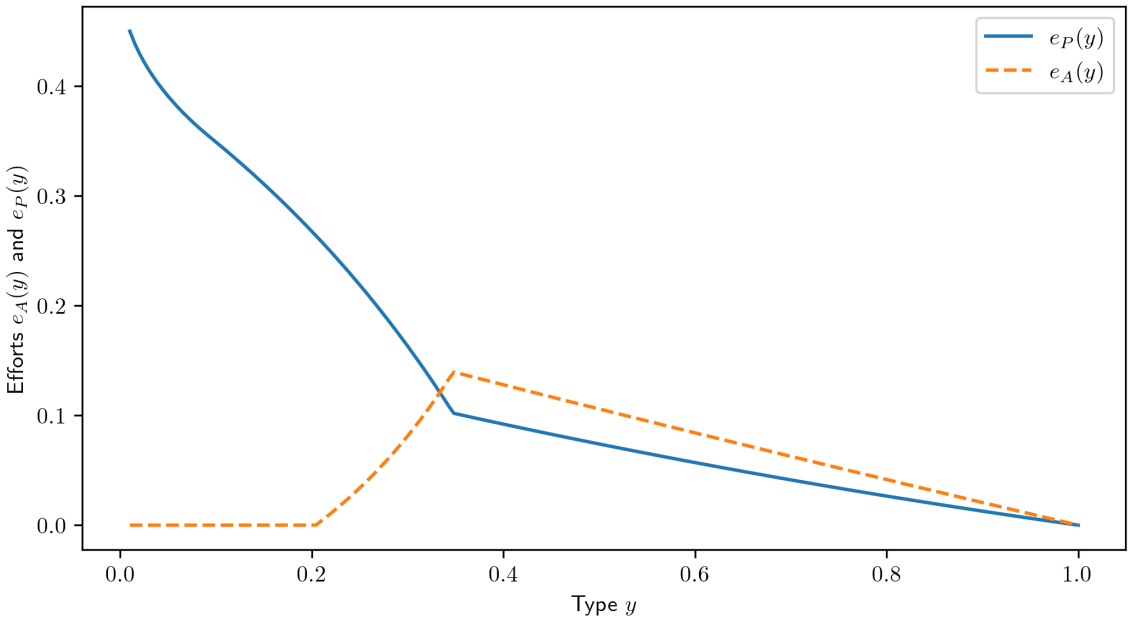


**Figure 4.** A mechanism computed via step (4) of the tightening algorithm with the modified formula for the profit under simultaneous evidence acquisition. The type space is $[0, 1]$, the effort costs are $c_P(e) = e^2$ and $c_A(e) = \exp(e) - 1 - e$ for all $e \in [0, 1]$, the principal's funds are $\tau = 0.2$. We applied step (4) to $\tilde{\lambda}$ given by $\tilde{\lambda}(x) = \sqrt{1 + x} - 1$ for all $x \in [0, 1]$.

suppose $d = 0$ has probability 0, so that the agent's evidence almost surely reveals that the agent deviated. We conjecture that our analysis is unchanged since on-path the agent advances the full surplus, and since, after deviating, the agent will not acquire evidence to prove that they deviated.

Things are less obvious if the principal's evidence uncovers the non-advanced surplus stochastically. We conjecture that optimal mechanisms induce local binding incentive constraints if each type of the agent can pass as a nearby type at a smooth rate. This conjecture follows Ball and Kattwinkel (2025), who analyze probabilistic verification in a general principal-agent problem without agent-evidence.

**Evidence about the advance payment.** In our model, evidence is obtainable only after the agent has advanced a payment; that is, evidence is used to substantiate the agent's claims. We motivate this assumption as follows.

First, suppose evidence reveals the difference between the advance payment and the true surplus. A non-existent advance payment is interpreted as zero. Since the true surplus is non-zero, the principal infers from the presented evidence if the agent

acquired it before or after advancing a payment. The principal will deter the agent from acquiring evidence before advancing a payment (for reasons that become apparent below). Thus, our analysis applies.

Second, the principal may be able to verify *when* the agent exerts effort. For example, a citizen may have to ask the state what exactly constitutes documentation of wealth; else, the citizen cannot begin exerting effort to produce documentation. Similarly, suppose there is a date at which the principal is sure that the agent knows the surplus and could not yet have acquired evidence. For example, suppose the surplus, that is to be divided between an entrepreneur and an investor, is the profit that was realized up to a fixed date. It may take the entrepreneur some time to gather evidence, and the entrepreneur cannot begin doing so before all profits are realized.

But suppose the agent could in fact acquire evidence even before advancing a payment. Evidence reveals the true surplus and makes it contractible, but the agent's evidence does not reveal when it was acquired. Then, the agent would first exert covert effort to acquire evidence, and then choose an advance payment *contingent* on whether the effort is successful. The agent again enjoys a rent from the hidden action of exerting effort, but this rent's effect on the hidden information problem is different to the baseline set-up. If the agent acquires evidence, the agent will advance the surplus and present evidence. However, if the agent fails to acquire evidence, the agent may contemplate deviating. Thus, the incentive constraint of advancing the full surplus binds only in the event that the agent fails to acquire evidence. Consequently, the rents from the hidden action problem do not strengthen the agent's incentive in the hidden information problem. We conjecture that, for all types for which principal-effort is non-zero, the principal optimally implements an inefficiently low agent-effort; specifically, $e_A(x) \in \operatorname{argmin}_{\tilde{e}_A} u_A(\tilde{e}_A) + c_A(\tilde{e}_A) + (1 - \tilde{e}_A)c_P(e_P(x))$ for all $x \in [\underline{x}, \bar{x}]$. The agent's and the principal's efforts are co-monotonically decreasing for all types. There are no subintervals on which the agent's effort increases.

**Unbounded private funds $\tau$.** If the principal has enormous private funds ($\tau \to \infty$), then, in the characterization of tight mechanisms with random audits, the interval High vanishes, because, intuitively, the additional benefit of using the evidence rent to provide rewards for truthtelling is small if the principal can pay a high reward $\tau$. The interval SuperHigh need not vanish, and the principal's effort is strictly positive but close to 0 throughout this interval; this is intuitive from inspecting the formula (10) for the principal's effort on SuperHigh. In economic terms, this characterization shows

that with large private funds the principal scarcely uses the agent-evidence refund $r_A$ to provide bonus rewards for truthtelling (but only uses $r_A$ to incentivize agent-effort). Such bonus rewards are provided only via the principal-evidence refund $r_P$, if at all.

The analysis of the limit where $\tau = \infty$—i.e., the principal can commit to arbitrarily large transfers—is delicate due to existence issues. To certain types $x$, the principal wishes to make a large transfer with vanishing probability after the principal acquires evidence. In the limit, it is ill-defined to make an unbounded transfer with probability zero. Importantly, the principal cannot shift this transfer to the no-evidence refund if type $x$ is in $[\underline{x}, \bar{x})$ because raising the no-evidence refund at $x$ attracts other types to deviate to $x$. For these reasons, an optimal mechanism fails to exist. Similarly, given a mechanism $m$ one cannot generally find a tight mechanisms that is tighter than $m$.

It is well-known that existence issues may arise in models with principal-evidence and unbounded transfers, e.g. Border and Sobel (1987) and Ahmadzadeh and Waizmann (2024). Border and Sobel impose bounds on transfers, and Ahmadzadeh and Waizmann study approximately optimal mechanisms.

# A Preliminaries

## A.1 Revelation Principle

Following Myerson (1982), a general mechanism is of the following form.

(1) The agent reports the type via a cheap talk message.

(2) The principal recommends an advance payment and an agent-effort.

(3) The agent advances a payment, and then covertly exerts effort to acquire evidence (recall that evidence can only be acquired after an advance payment has been made). If the agent obtains evidence, the agent chooses whether to disclose it.

(4) The principal exerts effort to acquire evidence. Then, the principal implements a transfers. The transfer is constrained by the advance payment and the available evidence, as described in Section 2.

Additionally, the agent finds it optimal to report the type truthfully and exert the recommended effort. Note that Myerson's Revelation Principle captures any "grand" mechanism featuring multiple rounds of cheap-talk. Our only assumptions on the evidence technology are that the agent can only acquire evidence after they made the public advance payment, and that both agent and principal can try to acquire

evidence only once. In such a grand mechanism, the agent may exert effort randomly following some randomized cheap talk. The recommended agent-effort following a report $x$ then replicates the ex-ante probability that type $x$ acquires evidence in the grand mechanism in equilibrium, and the principal's effort following report $x$ is the ex-ante probability that the principal acquires evidence when the agent is type $x$.

We now argue that it is without loss for the principal to use a feasible incentive-compatible tax mechanism. First, whenever the agent discloses evidence, the principal optimally does not exert effort to acquire evidence; the reason is that the principal can already condition the transfer on the true surplus; thus, by possibly re-defining transfers, the principal can reduce their own effort costs. Second, rather than sending a cheap-talk message about the type, the principal demands the agent advance their type as payment ("Put your money where your mouth is!"); the principal commits to treating this payment as if the agent had reported the payment as their type; since the agent found it optimal to report truthfully, the agent now finds it optimal to advance the type. Third, if there is evidence that the agent did not advance the full surplus, the principal optimally seizes everything since doing so only strengthens the incentive to advance the full surplus. Similarly, if the agent advances strictly less than the lowest type $\underline{x}$, it follows that the agent did not advance the full surplus, and hence the principal optimally acquires evidence and seizes everything. We obtain a feasible incentive-compatible tax mechanism.

## A.2 Evidence rent

The next lemma shows that, without loss for optimality, the evidence rent as defined in (1) is given by $u_A$, where $u_A(\tilde{e}_A) = \tilde{e}_A c'_A(\tilde{e}_A) - c_A(\tilde{e}_A)$ for all $\tilde{e}_A \in [0,1]$. The proof amounts to showing that the principal does not provide a *strict* incentive for the agent to acquire evidence *with probability one*. Hence, the optimal choice of $e_A(x)$ satisfies a first-order condition.

**Lemma A.1.** *Let $\tilde{m}$ be mechanism. There is a mechanism $m$ such that $\Pi_{\tilde{m}} \leq \Pi_m$ and such that for all $x \in [\underline{x}, \bar{x}]$,*

$$e_A(x) \cdot (r_A(x) - (e_P(x) + (1 - e_P(x))r_\emptyset(x))) - c_A(e_A(x)) = u_A(e_A(x)). \tag{17}$$

*Proof of Lemma A.1.* It suffices to find a mechanism $m$ such that $\Pi_{\tilde{m}} \leq \Pi_m$ and $r_A(x) - (e_P(x)r_P(x) + (1 - e_P(x))r_\emptyset(x) \leq c'_A(1)$ for all $x$. Indeed, the latter inequality

implies that $e_A(x)$ satisfies the first-order condition $c_A'(e_A(x)) = r_A(x) - (e_P(x) + (1 - e_P(x))r_\emptyset(x)$, yielding (17). We may assume $c_A'(1)$ is finite as else there is nothing to prove. Fix $x$ such that $\tilde{r}_A(x) - (\tilde{e}_P(x)\tilde{r}_P(x) + (1 - \tilde{e}_P(x))\tilde{r}_\emptyset(x) > c_A'(1)$. Thus, type $x$ acquires evidence with certainty $\tilde{e}_A(x) = 1$. We modify the mechanism at type $x$ as follows. Given $\tilde{e}_A(x) = 1$, the principal may as well acquire evidence with probability 1 (since the principal expects to never do so on-path) and increase $\tilde{r}_P(x)$ to $r_P(x) = r_A(x) - c_A'(1)$. This increase of $r_P(x)$ is feasible since $\tilde{r}_A(x) - (\tilde{e}_P(x)\tilde{r}_P(x) + (1 - \tilde{e}_P(x))\tilde{r}_\emptyset(x) > c_A'(1)$ holds and since $\tilde{e}_P(x)\tilde{r}_P(x) + (1 - \tilde{e}_P(x))\tilde{r}_\emptyset(x) \in [0, x + \tau]$. All other parts of the mechanism are left unchanged. Increasing $\tilde{e}_P(x)$ to 1 increases the loss from not advancing the full surplus, and both the agent's on-path loss and the principal's profit is unchanged (since, again, $r_P(x)$ is never paid on-path). Thus, by repeating the modification for every type $x$ satisfying $\tilde{r}_A(x) - (\tilde{e}_P(x)\tilde{r}_P(x) + (1 - \tilde{e}_P(x))\tilde{r}_\emptyset(x) > c_A'(1)$, we obtain a mechanism $m$ with the same profit as $\tilde{m}$. By construction, $r_A(x) - (e_P(x)r_P(x) + (1 - e_P(x))r_\emptyset(x) \leq c_A'(1)$ for all $x$. □

# B Tight mechanisms

## B.1 Preliminaries

This part of the appendix develops basic properties of the maximization problem from step (4) of the tightening algorithm (restated in Definition B.1 below). This maximization problem is crucial for characterizing tightness: as made precise by Proposition B.2, we can characterize tight mechanisms via the comparative statics of the maximization problem in the type. Along the way, we prove in Proposition B.1 that the tightening algorithm tightens the input mechanism.

Let $\Lambda_0$ denote the set of functions $\lambda\colon [\underline{x}, \bar{x}] \to [-\tau, \bar{x}]$ such that $\lambda \leq \mathrm{id}$. For every mechanism, the induced loss function is in $\Lambda_0$. For later reference, also define $\Lambda$ as the set of increasing concave $\lambda\colon [\underline{x}, \bar{x}] \to \mathbb{R}$ satisfying $\lambda(\underline{x}) = \underline{x}$ and $\lambda \leq \mathrm{id}$.

**Definition B.1.** Given $y \in [\underline{x}, \bar{x}]$ and $\lambda \in \Lambda_0$, let $M(y, \lambda)$ be the set of tuples $(\tilde{e}_A, \tilde{e}_P, \tilde{r}_A, \tilde{r}_P, \tilde{r}_\emptyset) \in [0,1]^2 \times [0, y + \tau]^3$ such that all of the following hold:

$$\lambda(y) \geq y - (\tilde{e}_P\tilde{r}_P + (1 - \tilde{e}_P)\tilde{r}_\emptyset) - u_A(\tilde{e}_A); \tag{18a}$$

$$\forall x \in [y, \bar{x}], \quad \lambda(x) \leq \tilde{e}_P x + (1 - \tilde{e}_P)(y - \tilde{r}_\emptyset); \tag{18b}$$

$$c_A'(\tilde{e}_A) = \tilde{r}_A - (\tilde{e}_P\tilde{r}_P + (1 - \tilde{e}_P)\tilde{r}_\emptyset)\,. \tag{18c}$$

Denote the profit from such a tuple by

$$\Pi(y, \tilde{e}_A, \tilde{e}_P, \tilde{r}_A, \tilde{r}_P, \tilde{r}_\emptyset) = y - \tilde{e}_P \tilde{r}_P - (1 - \tilde{e}_P)\tilde{r}_\emptyset - u_A(\tilde{e}_A) - c_A(\tilde{e}_A) - (1 - \tilde{e}_A)c_P(\tilde{e}_P).$$

Finally, define $T(y, \lambda) = \operatorname{argmax}_{m(y) \in M(y,\lambda)} \Pi(y, m(y))$.[17] We write $m \in T(\lambda)$ (resp., $m \in M(\lambda)$) to mean $m(y) \in T(y, \lambda)$ (resp. $m(y) \in M(y, \lambda)$) for all $y$.[18]

We first verify that applying $T$ tightens a mechanism.

**Lemma B.1.** *Let $m$ be a mechanism. If $m^* \in T(\lambda_m)$, then $(\Pi_m, \lambda_m) \leq (\Pi_{m^*}, \lambda_{m^*})$.*

*Proof of Lemma B.1.* We first observe $m(y) \in M(y, \lambda_m)$ for all $y$. Indeed, (18a) says that $m$ is IC for type $y$, (18b) follows from the definition $\lambda_m$, and (18c) is simply a restatement of (1) and (2).

Now, if $m^* \in T(\lambda_m)$, then under $m^*$ every type $y$ loses at most $\lambda_m(y)$ from being truthful, but loses at least $\lambda_m(y)$ from deviating. It follows that $m^*$ is feasible, IC, and induces a loss function $\lambda_{m^*}$ which lies pointwise above $\lambda_m$. Further, the profit under $m^*$ is also pointwise higher than the profit under $m$ since $m \in M(\lambda_m)$. □

We relate $T$ to another maximization problem $\hat{T}$ that obtains by guessing-and-verifying the optimal refunds. Fixing type $y$, the refunds for type $y$ should be minimal subject to $y$'s incentive to be truthful. Further, since the no-evidence refund $\tilde{r}_\emptyset$ attracts other types to deviate to $y$, the refund $\tilde{r}_\emptyset$ should be non-zero only if $\tilde{r}_P$ equals its upper bound $y + \tau$, i.e. if $\tilde{r}_P$ cannot be used to control $y$'s incentives. Finally, by Assumption 1, the principal's funds suffice to incentivize any agent-effort via $\tilde{r}_A$.

**Definition B.2.** For all $y \in [\underline{x}, \bar{x}]$ and $\lambda \in \Lambda_0$, define $E(y, \lambda)$ as the set of pairs $(\tilde{e}_A, \tilde{e}_P) \in [0, 1]$ such that,

$$\forall x \in [y, \bar{x}], \quad \lambda(x) \leq \tilde{e}_P x + \min\left((1 - \tilde{e}_P)y, \lambda(y) + u_A(\tilde{e}_A) + \tilde{e}_P \tau\right). \tag{19}$$

For $(\tilde{e}_A, \tilde{e}_P) \in E(y, \lambda)$, define

$$\hat{\Pi}(y, \lambda(y), \tilde{e}_A, \tilde{e}_P) = \min\left(y - u_A(\tilde{e}_A), \lambda(y)\right) - c_A(\tilde{e}_A) - (1 - \tilde{e}_A)c_P(\tilde{e}_P). \tag{20}$$

Finally, define $\hat{T}(y, \lambda) = \operatorname{argmax}_{(\tilde{e}_A, \tilde{e}_P) \in E(y,\lambda)} \hat{\Pi}(y, \lambda(y), \tilde{e}_A, \tilde{e}_P)$.

[17] For all $y$ and $\lambda \in \Lambda_0$, the set $M(y, \lambda)$ is non-empty; e.g., let $\tilde{e}_P = 1$, $\tilde{e}_A = 0$ and $\tilde{r}_P = \tilde{r}_A = y + \tau$ (and $\tilde{r}_\emptyset$ arbitrary). By compactness and continuity, $T(y, \lambda)$ is non-empty.

[18] Formally, $T(\lambda)$ (resp., $M(\lambda)$) is the set of selections from $y \mapsto T(y, \lambda)$ (resp., from $y \mapsto M(y, \lambda)$).

Intuitively, the minimum in the profit (20) is given by $\lambda(y)$ when $y$'s IC binds. The minimum in (19) is given by $\lambda(y) + u_A(\tilde{e}_A) + \tilde{e}_P\tau$ when $r_P(y) = y + \tau$.

The next lemma establishes an equivalence between $\hat{T}$ and $T$.

**Lemma B.2.** *Let $y \in [\underline{x}, \bar{x}]$ and $\lambda \in \Lambda_0$. For all $(\tilde{e}_A, \tilde{e}_P) \in [0,1]^2$, there are $\tilde{r}_A, \tilde{r}_P, \tilde{r}_\emptyset$ such that $(\tilde{e}_A, \tilde{e}_P, \tilde{r}_A, \tilde{r}_P, \tilde{r}_\emptyset) \in T(y, \lambda)$ if and only if $(\tilde{e}_A, \tilde{e}_P) \in \hat{T}(y, \lambda)$. Further, if $m \in T(\lambda)$, then for all $y \in [\underline{x}, \bar{x}]$,*

$$\begin{aligned}
\Pi(y, m(y)) &= \hat{\Pi}(y, \lambda(y), e_A(y), e_P(y));\\
U_m(y) &= \max(u_A(e_A(y)), y - \lambda(y));\\
\lambda_m(y) &= \inf_{z \in [\underline{x}, y]} e_P(z)y + \min\left((1 - e_P(z))z, \lambda(z) + u_A(e_A(z)) + e_P(z)\right).
\end{aligned}$$

*Proof of Lemma B.2.* Fixing $\tilde{e}_A$ and $\tilde{e}_P$, consider $(\tilde{r}_A, \tilde{r}_P, \tilde{r}_\emptyset)$ chosen as follows: the refund $\tilde{r}_\emptyset$ solves $(1 - \tilde{e}_P)(y - \tilde{r}_\emptyset) = \min\left((1 - \tilde{e}_P)y, \lambda(y) + u_A(\tilde{e}_A) + \tilde{e}_P\tau\right)$, the refund $\tilde{r}_P$ solves $\tilde{e}_P\tilde{r}_P + (1 - \tilde{e}_P)(y - \tilde{r}_\emptyset) = \min(y - u_A(\tilde{e}_A), \lambda(y))$, and the refund $\tilde{r}_A$ solves $c_A'(\tilde{e}_A) = \tilde{r}_A - (\tilde{e}_P\tilde{r}_P + (1 - \tilde{e}_P)\tilde{r}_\emptyset)$. Assumption 1 implies $\tilde{r}_A \in [0, y + \tau]$. It is tedious but straightforward to verify that the triple $(\tilde{r}_A, \tilde{r}_P, \tilde{r}_\emptyset)$ maximizes $\Pi(y, \cdot)$ subject to $(\tilde{e}_A, \tilde{e}_P, \tilde{r}_A, \tilde{r}_P, \tilde{r}_\emptyset) \in M(y, \lambda)$ and results in profit $\hat{\Pi}(y, \lambda(y), \tilde{e}_A, \tilde{e}_P)$. Every other maximizing triple agrees with the chosen triple, unless $\tilde{e}_P = 0$ (resp., $\tilde{e}_P = 1$), in which case the choice of $\tilde{r}_P$ (resp., of $\tilde{r}_\emptyset$) is irrelevant. The formulas for the profit, the interim utility and the induced loss also follow from these formulas for the refunds. □

Recall that $\Lambda$ is the set of increasing concave $\lambda\colon [\underline{x}, \bar{x}] \to \mathbb{R}$ satisfying $\lambda(\underline{x}) = \underline{x}$ and $\lambda \leq \mathrm{id}$. In a tight mechanism, the induced loss function is in $\Lambda$, as we will show below. Towards this result, the following lemma is essential.

**Lemma B.3.** *Let $m$ be a mechanism, and let $\lambda_m$ be its induced loss function. For all $x \in [\underline{x}, \bar{x}]$, let $\lambda^+(x) = \max(\underline{x}, \sup_{x' \in [\underline{x}, x]} \lambda_m(x'))$, and*

$$\tilde{\lambda}(x) = \inf_{y \in [x, \bar{x}]} e_P(y)x + \min\left((1 - e_P(y))y, \lambda^+(y) + u_A(e_A(y)) + e_P(y)\tau\right). \tag{21}$$

*Then $\tilde{\lambda} \in \Lambda$ and all $x \in [\underline{x}, \bar{x}]$ satisfy $(e_A(x), e_P(x)) \in E(x, \tilde{\lambda})$, $\lambda_m(x) \leq \tilde{\lambda}(x)$, and $\Pi_m(x) \leq \hat{\Pi}(x, \lambda^+(x), e_A(x), e_P(x)) \leq \hat{\Pi}(x, \tilde{\lambda}(x), e_A(x), e_P(x))$.*

*Proof of Lemma B.3.* First, note that $\lambda_m \leq \lambda^+ \leq \mathrm{id}$ and $\lambda^+(\underline{x}) = \underline{x}$ hold, and that $\lambda^+$ is increasing.

To verify $\tilde{\lambda} \in \Lambda$, note that $\tilde{\lambda}$ is increasing and concave as $\tilde{\lambda}$ is the pointwise infimum of increasing affine functions Finally, $\tilde{\lambda}(\underline{x}) = \underline{x}$ and $\tilde{\lambda} \leq \mathrm{id}$ follow from inspecting (21) and using $\lambda^+ \leq \mathrm{id}$ and $\lambda^+(\underline{x}) = \underline{x}$.

In an intermediate step, we show $\lambda^+(x) \leq \tilde{\lambda}(x)$ for all $x$. By inspection, $\underline{x} \leq \tilde{\lambda}(x)$. Thus, we show $\sup_{x' \in [\underline{x}, x]} \lambda_m(x') \leq \tilde{\lambda}(x)$; i.e. for all $x' \in [\underline{x}, x]$ and all $y$,

$$\lambda_m(x') \leq e_P(y)x + \min\left((1 - e_P(y))y, \lambda^+(y) + u_A(e_A(y)) + e_P(y)\tau\right).$$

First, if $x' \leq y$, then since $\lambda^+$ is increasing and $\lambda_m(x') \leq \lambda^+(x') \leq x'$, it holds

$$\begin{aligned}\lambda_m(x') \leq \lambda^+(x') &\leq e_P(y)x' + \min\left((1 - e_P(y))x', \lambda^+(x')\right) \\ &\leq e_P(y)x + \min\left((1 - e_P(y))y, \lambda^+(y) + u_A(e_A(y)) + e_P(y)\tau\right),\end{aligned}$$

as desired. Second, if $y \leq x'$, then since $(e_A, e_P) \in E(\lambda)$ and $\lambda_m(y) \leq \lambda^+(y)$, it holds

$$\begin{aligned}\lambda_m(x') &\leq e_P(y)x' + \min\left((1 - e_P(y))y, \lambda_m(y) + u_A(e_A(y)) + e_P(y)\tau\right) \\ &\leq e_P(y)x + \min\left((1 - e_P(y))y, \lambda^+(y) + u_A(e_A(y)) + e_P(y)\tau\right),\end{aligned}$$

Thus, $\lambda^+(x) \leq \tilde{\lambda}(x)$ for all $x$.

It now easily follows $(e_A, e_P) \in E(\tilde{\lambda})$; indeed, if $y \leq x$, then (the first inequality follows from the definition of $\tilde{\lambda}$, the second inequality follows from $\lambda^+ \leq \tilde{\lambda}$)

$$\begin{aligned}\tilde{\lambda}(x) &\leq e_P(y)x + \min\left((1 - e_P(y))y, \lambda^+(y) + u_A(e_A(y)) + e_P(y)\tau\right) \\ &\leq e_P(y)x + \min\left((1 - e_P(y))y, \tilde{\lambda}(y) + u_A(e_A(y)) + e_P(y)\tau\right).\end{aligned}$$

Finally, for all $x$, Lemmata B.1 and B.2 imply $\Pi_m(x) \leq \hat{\Pi}(x, \lambda_m(x), e_A(x), e_P(x))$, and $\lambda_m \leq \lambda^+$ implies $\hat{\Pi}(x, \lambda_m(x), e_A(x), e_P(x)) \leq \hat{\Pi}(x, \tilde{\lambda}(x), e_A(x), e_P(x))$. □

The next proposition shows that the tightening algorithm tightens the input mechanism, and that the virtual loss function from the algorithm is in $\Lambda$.

**Proposition B.1.** *Let $m$ be a mechanism. Applying the tightening algorithm to $m$, let $\tilde{\lambda}$ and $m^*$, resp., denote the obtained virtual loss function and mechanism, resp. Then, $\tilde{\lambda} \in \Lambda$, $m^* \in T(\tilde{\lambda})$, and $(\Pi_m, \lambda_m) \leq (\Pi_{m^*}, \tilde{\lambda}) \leq (\Pi_{m^*}, \lambda_{m^*})$.*

*Proof of Proposition B.1.* First, we show the virtual loss function $\tilde{\lambda}$ equals the function given in (21) of Lemma B.3. Let $\lambda^+$ be as Lemma B.3 (and hence as in step (1) of the

algorithm). Fix $x$. For type $x$, step (2) of the algorithm entails choosing $(\tilde{r}_A, \tilde{r}_P, \tilde{r}_\emptyset)$ to maximize $\Pi(x, e_A(x), e_P(x), \tilde{r}_A, \tilde{r}_P, \tilde{r}_\emptyset)$ subject to $(e_A(x), e_P(x), \tilde{r}_A, \tilde{r}_P, \tilde{r}_\emptyset) \in M(x, \lambda^+)$. As in the proof of Lemma B.2, the solution entails $\tilde{r}_\emptyset$ satisfying $(1 - e_P(x))(x - \tilde{r}_\emptyset) = \min((1 - e_P(x))x, \lambda^+(x) + u_A(e_A(x)) + e_P(x)\tau)$. Applying this formula for all $x$, the virtual loss as defined in (13) equals (21).

Given that the virtual loss is given by (21), Lemma B.3 implies $\tilde{\lambda} \in \Lambda$ and $\lambda_m \leq \tilde{\lambda}$. Step (4) of the algorithm entails choosing $m^* \in T(\tilde{\lambda})$. Thus, Lemma B.2 implies $\hat{\Pi}(x, \tilde{\lambda}, e_A(x), e_P(x)) \leq \Pi_{m^*}(x)$ for all $x$. Lemma B.3 further implies $\Pi_m(x) \leq \hat{\Pi}(x, \tilde{\lambda}, e_A(x), e_P(x))$. In sum, $\Pi_m \leq \Pi_{m^*}$. Finally, Lemma B.1 implies $\tilde{\lambda} \leq \lambda_{m^*}$. □

The next proposition is crucial for characterizing tightness. The key claims of the characterization (Sections 3.1 and 3.2) ask how in a tight mechanism $m$ the efforts and the binding constraints change with the agent's type. The next proposition turns these questions into a comparative statics problem: how does the set of maximizers $T(y, \lambda_m)$ change in $y$, given that $m$'s induced loss function $\lambda_m$ is in $\Lambda$?

**Proposition B.2.** *If $m$ is tight, then $\lambda_m \in \Lambda$, $m \in T(\lambda_m)$, and all $x \in [\underline{x}, \bar{x}]$ satisfy*

$$\begin{aligned}
\Pi_m(x) =& \min(x - u_A(e_A(x)), \lambda_m(x)) - u_A(e_A(x)) \\
&- c_A(e_A(x)) - (1 - e_A(x))c_P(e_P(x)), \\
U_m(x) =& \max(u_A(e_A(x)), x - \lambda_m(x)), \\
\lambda_m(x) =& \inf_{y \in [\underline{x}, x]} e_P(y)x + \min\left((1 - e_P(y))y, \lambda_m(y) + u_A(e_A(y)) + e_P(y)\tau\right).
\end{aligned}$$

*Proof of Proposition B.2.* By Proposition B.1, there is $\tilde{\lambda} \in \Lambda$ and $m^* \in T(\tilde{\lambda})$ such that $(\Pi_m, \lambda_m) \leq (\Pi_{m^*}, \tilde{\lambda}) \leq (\Pi_{m^*}, \lambda_{m^*})$ (namely, $m^*$ obtained by applying the tightening algorithm to $m$). Since $m$ is tight, also $(\Pi_m, \lambda_m) = (\Pi_{m^*}, \lambda_{m^*})$. In particular, $\lambda_m = \tilde{\lambda}$, and hence $\lambda_m \in \Lambda$. Further, $\Pi_m = \Pi_{m^*}$ and $m^* \in T(\tilde{\lambda})$ and $\lambda_m = \tilde{\lambda}$ together imply $m \in T(\lambda_m)$. The claims regarding $\Pi_m$, $U_m$, and $\lambda_m$ now follow from Lemma B.2. □

## B.2 Proof of Lemma 3.1

The proof uses definitions from Appendix B.1.

Let $m$ be a mechanism. We show there is a tight mechanism $m^*$ such that $(\Pi_m, \lambda_m) \leq (\Pi_{m^*}, \lambda_{m^*})$. Let $\mathcal{I}$ denote the set of pairs $(\Pi, \lambda)$ of functions from $[\underline{x}, \bar{x}]$ to $\mathbb{R}$ such that $\lambda \in \Lambda_0$, such that $(\Pi_m, \lambda_m) \leq (\Pi, \lambda)$, and such that there is a mechanism

$m' \in T(\lambda)$ whose profit $\Pi_{m'}$ equals $\Pi$. The set $\mathcal{I}$ is partially ordered by $\leq$. The set $\mathcal{I}$ is non-empty; e.g. find $m' \in T(\lambda_m)$, so that Lemma B.1 implies $(\Pi_m, \lambda_m) \leq (\Pi_{m'}, \lambda_m)$, meaning $(\Pi_{m'}, \lambda_m) \in \mathcal{I}$.

Below, we show that $\mathcal{I}$ has a maximal element $(\Pi^*, \lambda^*)$. Before doing so, we find a mechanism with the desired properties, taking $(\Pi^*, \lambda^*)$ as given. By definition of $\mathcal{I}$, we may find a mechanism $m^*$ such that $m^* \in T(\lambda^*)$ and $\Pi^* = \Pi_{m^*}$. Lemma B.1 implies $\lambda^* \leq \lambda_{m^*}$. Thus $(\Pi_m, \lambda_m) \leq (\Pi^*, \lambda^*) \leq (\Pi_{m^*}, \lambda_{m^*})$, and thus it remains to show $m^*$ is tight. To that end, let $m'$ be a mechanism such that $(\Pi_{m^*}, \lambda_{m^*}) \leq (\Pi_{m'}, \lambda_{m'})$. Apply $T$ to $\lambda_{m'}$ to find $m'' \in T(\lambda_{m'})$. Thus $\Pi_{m'} \leq \Pi_{m''}$ (Lemma B.1), and thus $(\Pi_{m''}, \lambda_{m'}) \in \mathcal{I}$ and $(\Pi^*, \lambda^*) \leq (\Pi_{m^*}, \lambda_{m^*}) \leq (\Pi_{m'}, \lambda_{m'}) \leq (\Pi_{m''}, \lambda_{m'})$. Since $(\Pi^*, \lambda^*)$ is maximal in $\mathcal{I}$, we infer $(\Pi_{m'}, \lambda_{m'}) = (\Pi_{m^*}, \lambda_{m^*})$, proving that $m^*$ is tight.

It remains to show that $\mathcal{I}$ admits a maximal element. Using Zorn's Lemma, it suffices to show that every chain $C$ in $\mathcal{I}$ admits an upper bound in $\mathcal{I}$. We denote a typical element $c$ of $C$ by $c = (\Pi^c, \lambda^c)$. Our candidate for $\lambda^*$ is defined for all $y \in [\underline{x}, \bar{x}]$ by $\lambda^*(y) = \sup_{c \in C} \lambda^c(y)$. Next, we find $m^* \in T(\lambda^*)$. (Using that for every $c$ the function $\lambda^c$ is in $\Lambda_0$, one may verify $\lambda^* \in \Lambda_0$; hence it makes sense to apply $T$ to $\lambda^*$.) Let $\Pi^* = \Pi_{m^*}$ be the profit induced by $m^*$. We show $\sup_{c \in C} \Pi^c(y) \leq \Pi^*(y)$ for all $y \in [\underline{x}, \bar{x}]$ (which implies that $(\Pi^*, \lambda^*)$ is an upper bound of $C$ in $\mathcal{I}$ and hence completes the proof). Fix $y$. For every $c \in C$, find a mechanism $m^c$ such that $m^c \in T(\lambda^c)$ and $\Pi^c = \Pi_{m^c}$. We view $(m^c, \Pi^c, \lambda^c)_{c \in C}$ as a net with the order inherited from $C$. Since $C$ is a chain, the net $(\lambda^c)_{c \in C}$ converges pointwise to $\lambda^*$, and $(\Pi^c(y))_{c \in C}$ converges to $\sup_{c \in C} \Pi^c(y)$. Further, for every $c$, we have $m^c(y) \in [0,1]^2 \times [0, y + \tau]^3$. By possibly passing to a subnet, let $m^c(y)$ converge to $m^\dagger(y)$. Since for every $c$ the mechanism $m^c$ induces $\Pi^c$ and $(\Pi^c(y))_{c \in C}$ converges to $\sup_{c \in C} \Pi^c(y)$, it follows that $\sup_{c \in C} \Pi^c(y)$ equals the profit of $m^\dagger$ at $y$. Since also $m^c(y) \in M(y, \lambda^c)$ and $\lambda^c \to \lambda^*$ pointwise, one may verify $m^\dagger(y) \in M(y, \lambda^*)$. Since $m^*(y) \in T(y, \lambda^*)$, the profit of $m^\dagger$ at $y$ is less than the profit of $m^*$ at $y$. Since the profit of $m^\dagger$ at $y$ equals $\sup_{c \in C} \Pi^c(y)$ while the profit induced by $m^*$ at $y$ equals $\Pi^*(y)$, we conclude $\sup_{c \in C} \Pi^c(y) \leq \Pi^*(y)$. □

## B.3 Preparations for the tightness characterization

This section derives auxiliary results to characterize tightness. Perhaps most importantly: Proposition B.3 (Appendix B.3.2) shows how profit is optimally pinned down by the implemented agent-effort, and Propositions B.4 and B.5 (Appendix B.3.3) char-

acterize how the benefits of agent-effort depend on the type and feasibility constraints.

The following definitions and results take as given and depend on a function $\lambda \in \Lambda$. For the sake conciseness, we do not indicate this dependence in the notation. No confusion should arise here since for now we only deal with one fixed function $\lambda \in \Lambda$.

### B.3.1 Increasing information rent

Let $\lambda \in \Lambda$, i.e. $\lambda$ is increasing and concave, and $\lambda(\underline{x}) = \underline{x}$ and $\lambda \leq \mathrm{id}$. An important implication is that $\lambda$ is Lipschitz continuous with constant 1. A second important implication is that the information rent $y - \lambda(y)$ is increasing but the relative loss $\frac{\lambda(y)}{y}$ is decreasing in $y$.

### B.3.2 Minimal principal-effort

Given $\lambda \in \Lambda$ and a type $y$, consider the maximization problem $\hat{T}(y, \lambda)$ (Definition B.2). As confirmed below, the optimal principal-effort equals the minimal feasible value, which equals the maximum of $\alpha$ and $\beta$, defined next. The minimum depends on $\lambda$, the type $y$, and the chosen agent-effort.

**Definition B.3.** Let $\lambda \in \Lambda$. For all $y \in [\underline{x}, \bar{x}]$, $\tilde{e}_A \in [0, 1]$, and $s \in \mathbb{R}$, let

$$\alpha(y) = \begin{cases} \max\left(0, \sup_{x \in (y, \bar{x}]} \frac{\lambda(x) - y}{x - y}\right), & \text{if } y < \bar{x}, \\ 0, & \text{if } y = \bar{x}. \end{cases}$$

$$\beta(y, \tilde{e}_A) = \max\left(0, \max_{x \in [\underline{x}, \bar{x}]} \frac{\lambda(x) - \lambda(y) - u_A(\tilde{e}_A)}{x + \tau}\right);$$

$$\hat{Z}(s) = \operatorname*{argmax}_{x \in [\underline{x}, \bar{x}]} \frac{\lambda(x) - s}{x + \tau}.$$

Let $\hat{z}$ be an increasing selection from $\hat{Z}$.[19]

Since $\lambda$ is continuous, $\beta$ is well-defined and continuous, and $\hat{Z}$ has non-empty compact values. Further, $\lambda \leq \mathrm{id}$ and $0 < \underline{x} + \tau$ imply $\beta(y, \tilde{e}_A) < 1$.

For the analysis, it helps to distinguish multiple regimes depending on whether the principal's effort equals $\alpha(y)$ or $\beta(y, \tilde{e}_A)$, and on whether the minimum in the objective (20) is given by $y - u_A(\tilde{e}_A)$ or $\lambda(y)$. If $y - u_A(\tilde{e}_A) < \lambda(y)$, then the principal's effort

[19] For example, for all $s$ let $\hat{z}(s) = \max \hat{Z}(s)$. Then $\hat{z}$ is well-defined since $\hat{Z}$ has non-empty compact values, and $\hat{z}$ is increasing since the objective in the definition of $\hat{Z}$ is supermodular in $(x, s)$.

must be given by $\alpha(y)$. Intuitively, the inequality $y - u_A(\tilde{e}_A) < \lambda(y)$ indicates that type $y$ has a strict incentive to be truthful. Thus, the principal should not pay type $y$ the maximal reward $r_P(y) = y + \tau$ after the principal acquires evidence, since this would provide excessive incentives.

Thus, there are three intersecting regimes: first, $y - u_A(\tilde{e}_A) \leq \lambda(y)$; second, $y - u_A(\tilde{e}_A) \geq \lambda(y)$ and $\alpha(y) \leq \beta(y, \tilde{e}_A)$; third, $\alpha(y) \geq \beta(y, \tilde{e}_A)$. The profits are:

**Definition B.4.** Let $y \in [\underline{x}, \bar{x}]$, $\lambda \in \Lambda$ and $\tilde{e}_A \in [0, 1]$. Define

$$\begin{aligned}
\pi_1(y, \tilde{e}_A) &= y - u_A(\tilde{e}_A) - c_A(\tilde{e}_A) - (1 - \tilde{e}_A)c_P(\alpha(y)) \\
\pi_2(y, \tilde{e}_A) &= \lambda(y) - c_A(\tilde{e}_A) - (1 - \tilde{e}_A)c_P(\alpha(y)) \\
\pi_3(y, \tilde{e}_A) &= \lambda(y) - c_A(\tilde{e}_A) - (1 - \tilde{e}_A)c_P(\beta(y, \tilde{e}_A)) \\
\underline{\pi}(y, \tilde{e}_A) &= \min\left(\pi_1(y, \tilde{e}_A), \pi_2(y, \tilde{e}_A), \pi_3(y, \tilde{e}_A)\right).
\end{aligned}$$

For $i \in \{1, 2\}$, we denote the derivative of $\pi_i$ with respect to its second argument by $\partial_2 \pi_i$. Turning to $\pi_3$, Theorem 2 of Milgrom and Segal (2002) implies that $\tilde{e}_A \mapsto \pi_3(y, \tilde{e}_A)$ is absolutely continuous for all $y \in [\underline{x}, \bar{x}]$.[20] We denote the (essentially unique) derivative of $\pi_3(y, \cdot)$ by $\partial_2 \pi_3(y, \cdot)$. Specifically, for all $y \in [\underline{x}, \bar{x}]$ and $\tilde{e}_A \in [0, 1]$,

$$\partial_2 \pi_3(y, \tilde{e}_A) = c_P(\beta(y, \tilde{e}_A)) - c'_A(\tilde{e}_A) + (1 - \tilde{e}_A)\frac{c'_P(\beta(y, \tilde{e}_A))u'_A(\tilde{e}_A)}{\hat{z}(\lambda(y) + u_A(\tilde{e}_A))) + \tau}\mathbf{1}_{(\beta(y,\tilde{e}_A))>0)},$$

where $\alpha$, $\beta$ and $\hat{z}$ are as in Definition B.3.

**Proposition B.3.** *Let $\lambda \in \Lambda$. If $m \in T(\lambda)$, then all $y \in [\underline{x}, \bar{x}]$ satisfy $e_A(y) < 1$ and*

$$\begin{aligned}
\Pi_m(y) = \underline{\pi}(y, e_A(y)) &= \max_{\tilde{e}_A \in [0,1]} \underline{\pi}(y, \tilde{e}_A), \\
e_P(y) &= \max\left(\alpha(y), \beta(y, e_A(y))\right).
\end{aligned}$$

*Proof of Proposition B.3.* Let $y \in [\underline{x}, \bar{x}]$. Lemma B.2 implies $(e_A(y), e_P(y) \in \hat{T}(y, \lambda)$. Thus $(e_A(y), e_P(y))$ maximizes (20) subject to (19). The objective (20) is decreasing in $\tilde{e}_P$, strictly so if $\tilde{e}_A < 1$. Fixing $\tilde{e}_A$, the minimal value of $\tilde{e}_P$ satisfying (19) at $(y, \tilde{e}_A)$ is $\max\left(\alpha(y), \beta(y, \tilde{e}_A)\right)$. Thus, to prove Proposition B.3, it suffices to show that $e_A(y) = 1$ cannot be optimal for maximizing (20) subject to (19). By inspection,

[20] Theorem 2 of Milgrom and Segal (2002) applies since the assumption $0 < \underline{x} + \tau$ implies that the objective in the definition of $\beta$ has a bounded derivative with respect to $\tilde{e}_A$ almost everywhere.

for all $i \in \{1,2,3\}$, the derivative of $\pi_i(y, \tilde{e}_A)$ with respect to $\tilde{e}_A$ at $\tilde{e}_A = 1$ equals $-u'_A(1)\mathbf{1}_{i=1} - c'_A(1) + c_P(1)$. Since $c'_A(1) > c_P(1)$ (Assumption 1) and $u_A$ is increasing, this derivative is strictly negative. Thus, $e_A(y) = 1$ cannot be optimal. □

### B.3.3 Quasi concavity and quasi submodularity

Next, we establish single-crossing properties of $\pi_1$, $\pi_2$, $\pi_3$, and $\underline{\pi}$, which are crucial for deriving the order described by Theorem 3.1. The next lemma uses Assumption 2.

**Lemma B.4.** *Let $\lambda \in \Lambda$, $y, y' \in X$, $e_A, e'_A \in [0,1]$. and $\lambda(y)+u_A(e_A) \leq \lambda(y')+u_A(e'_A)$.*
*If $e_A \leq e'_A$, then $\partial_2\pi_3(y, e_A) \geq \partial_2\pi_3(y', e'_A)$.*
*If $e_A < e'_A$, then $\partial_2\pi_3(y, e_A) > \partial_2\pi_3(y', e'_A)$.*

*Proof of Lemma B.4.* Abbreviate $q = \beta(y, e_A)$ and $q' = \beta(y', e'_A)$. Since $\lambda(y) + u_A(e_A) \leq \lambda(y') + u_A(e'_A)$ it holds $q \geq q'$ (by inspecting the definition of $\beta$) and $\hat{z}(\lambda(y) + u_A(e_A)) \leq \hat{z}(\lambda(y') + u_A(e'_A))$ (since $\hat{z}$ is increasing). Thus,

$$\begin{aligned}\partial_2\pi_3(y, e_A) =& c_P(q) - c'_A(e_A) + (1 - e_A)\frac{c'_P(q)u'_A(e_A)}{\hat{z}(\lambda(y) + u_A(e_A)) + \tau}\mathbf{1}_{(q>0)} \\ \geq& c_P(q') - c'_A(e_A) + (1 - e_A)\frac{c'_P(q')u'_A(e_A)}{\hat{z}(\lambda(y') + u_A(e'_A)) + \tau}\mathbf{1}_{(q'>0)},\end{aligned}$$

whereas

$$\partial_2\pi_3(y', e'_A) = c_P(q') - c'_A(e'_A) + (1 - e'_A)\frac{c'_P(q')u'_A(e'_A)}{\hat{z}(\lambda(y') + u_A(e'_A)) + \tau}\mathbf{1}_{(q'>0)}.$$

If $e_A = e'_A$, then clearly $\partial_2\pi_3(y, e_A) \geq \partial_2\pi_3(y', e'_A)$. It remains to show that if $e_A < e'_A$, then $\partial_2\pi_3(y, e_A) > \partial_2\pi_3(y', e'_A)$ To that end, it suffices to show that the map

$$\tilde{e}_A \mapsto -c'_A(\tilde{e}_A) + (1 - \tilde{e}_A)\frac{c'_P(q')u'_A(\tilde{e}_A)}{\hat{z}(\lambda(y') + u_A(e'_A)) + \tau}\mathbf{1}_{(q'>0)}$$

is strictly decreasing. If $q' = 0$, strict decreasingness is immediate from the strict convexity of $c_A$. Thus let $q' > 0$. By direct computation, the derivative with respect to $\tilde{e}_A$ is

$$-c''_A(\tilde{e}_A) - \frac{c'_P(q')}{\hat{z}(\lambda(y') + u_A(e'_A)) + \tau}\left(\tilde{e}_A c''_A(\tilde{e}_A) - (1 - \tilde{e}_A)(c''_A(\tilde{e}_A) + \tilde{e}_A c'''_A(\tilde{e}_A))\right).$$

If the term in the large parentheses is positive, we are done. So suppose said term is negative. Denote $\rho = \frac{\bar{x}}{\bar{x}+\tau}$, and note the bound $q' = \beta(y', e'_A) \leq \frac{\hat{z}(\lambda(y')+u_A(e'_A)))}{\hat{z}(\lambda(y')+u_A(e'_A)))+\tau} \leq \rho$. Thus it suffices to show

$$0 < c''_A(\tilde{e}_A) + \frac{c'_P(\rho)}{\underline{x}+\tau}\left(\tilde{e}_A c''_A(\tilde{e}_A) - (1-\tilde{e}_A)(c''_A(\tilde{e}_A) + \tilde{e}_A c'''_A(\tilde{e}_A))\right).$$

By rearranging, we find that this inequality is implied by Assumption 2. □

**Proposition B.4.** *Let $\lambda \in \Lambda$. For all $i \in \{1,2,3\}$, the function $\pi_i$ is submodular, and, for all $y \in [\underline{x}, \bar{x}]$, the function $\pi_i(y,\cdot)$ is strictly quasiconcave. For $y \in [\underline{x}, \bar{x}]$, the function $\underline{\pi}(y,\cdot)$ is strictly quasiconcave.*

*Proof of Proposition B.4.* Assumption 3 implies that $u_A + c_A$ is strictly quasiconvex, implying that $\pi_1(y,\cdot)$ is strictly quasiconcave for all $y$. Clearly, $\pi_2(y,\cdot)$ is strictly concave for all $y$. Finally, $\pi_1$ and $\pi_2$ are submodular since $\alpha$ decreases.

The second claim in Lemma B.4 implies that $\tilde{e}_A \mapsto \partial_2\pi_3(y,\tilde{e}_A)$ is strictly decreasing for all $y$; thus, $\pi_3(y,\cdot)$ is strictly quasiconcave. Next, recall that $\lambda$ is increasing. Hence, for all $\tilde{e}_A \in [0,1]$ and $y, y' \in X$, if $y < y'$, then $\lambda(y) + u_A(e_A) \leq \lambda(y') + u_A(e_A)$, and hence Lemma B.4 implies $\partial_2\pi_3(y,e_A) \geq \partial_2\pi_3(y',e_A)$. Hence, $\pi_3$ is submodular.

Finally, the function $\underline{\pi}(y,\cdot)$ is strictly quasiconcave for each $y$ since it is the pointwise-minimum of strictly quasiconcave functions. □

Propositions B.3 and B.4 and Lemma B.2 have the following corollary.

**Corollary B.1.** *Let $\lambda \in \Lambda$. Let $x \in [\underline{x}, \bar{x}]$. If $m(x)$ and $m^*(x)$ are both in $T(x,\lambda)$, then $(e_A(x), e_P(x)) = (e^*_A(x), e^*_P(x))$. Further, $\hat{T}(\lambda)$ is a singleton.*

The next lemma essentially says that the benefit from incentivizing agent-effort is highest when profit is given by $\pi_3$ and lowest when profit is given by $\pi_1$.

**Proposition B.5.** *Let $\lambda \in \Lambda$. Let $y \in [\underline{x}, \bar{x}]$ and $0 \leq e_A < e_A \leq 1$. Then*

$$\pi_1(y, e'_A) \geq \pi_1(y, e_A) \quad \Rightarrow \quad \pi_2(y, e'_A) \geq \pi_2(y, e_A); \tag{22}$$

*Further, if $\alpha(y) \leq \beta(y, e'_A)$, then*

$$\pi_2(y, e'_A) \geq \pi_2(y, e_A) \quad \Rightarrow \quad \pi_3(y, e'_A) \geq \pi_3(y, e_A). \tag{23}$$

*Proof of Proposition B.5.* For all $\tilde{e}_A \in [0,1]$, direct computation shows $\partial_2\pi_1(y,\tilde{e}_A) = c_P(\alpha(y)) - c'_A(\tilde{e}_A) - u'_A(\tilde{e}_A) \leq c_P(\alpha(y)) - c'_A(\tilde{e}_A) = \partial_2\pi_2(y,\tilde{e}_A)$. Thus (22).

Next, let $\alpha(y) \leq \beta(y, e'_A)$. Thus also $\alpha(y) \leq \beta(y,\tilde{e}_A)$ for all $\tilde{e}_A \in [e_A, e'_A]$ since $\beta$ is decreasing in its second argument. To establish (23), we now observe:

$$\begin{aligned}\partial_2\pi_3(y,\tilde{e}_A) =& c_P(\beta(y,\tilde{e}_A)) - c'_A(\tilde{e}_A) + (1-\tilde{e}_A)\frac{c'_P(\beta(y,\tilde{e}_A))u'_A(\tilde{e}_A)}{\hat{z}(\lambda(y)+u_A(\tilde{e}_A))}\mathbf{1}_{(\beta(y,\tilde{e}_A)>0)}\\ \geq& c_P(\beta(y,\tilde{e}_A)) - c'_A(\tilde{e}_A)\\ \geq& c_P(\alpha(y)) - c'_A(\tilde{e}_A) = \partial_2\pi_2(y,\tilde{e}_A).\end{aligned}$$

□

### B.3.4 Continuity

We next deal with continuity of $\alpha(y)$ in $y$. The supremum in the definition of $\alpha(y)$ is taken over the half-open interval $(y,\bar{x}]$, which can lead to a discontinuity. This discontinuity is important for delineating mechanisms with non-random audits from ones with random audits.

Given $\lambda \in \Lambda$, define $\underline{y} = \max\{y \in [\underline{x},\bar{x}] \colon \lambda(y) = y\}$. The set $\{y \in [\underline{x},\bar{x}] \colon \lambda(y) = y\}$ contains $\underline{x}$ and is a closed interval. The next lemma shows that, among other things, if $\alpha$ is interior at least once, then $\alpha$, $e_A$ and $e_P$ are continuous except possibly at $\underline{y}$.

**Lemma B.5.** *Let $\lambda \in \Lambda$ and $m \in T(\lambda)$. If $\{y \in [\underline{x},\bar{x}] \colon \alpha(y) \in (0,1)\}$ is non-empty, then $\underline{y} < \lambda(\bar{x})$ and for all $y \in [\underline{x},\bar{x}]$,*

*(1) if $y \in [\underline{x},\underline{y})$, then $\alpha(y) = 1$;*

*(2) if $y \in (\underline{y},\bar{x}]$, then $\alpha(y) < 1$, and $\alpha$, $e_A$ and $e_P$ are continuous at $y$;*

*(3) $\alpha$, $e_A$, and $e_P$ are right-continuous at $\underline{y}$;*

*(4) $\alpha$ is strictly decreasing on $(\underline{y},\lambda(\bar{x}))$, and constantly 0 on $[\lambda(\bar{x}),\bar{x}]$.*

*Proof of Lemma B.5.* First, we show $\underline{y} < \lambda(\bar{x})$. By assumption, there is $y$ such that $\alpha(y) \in (0,1)$. Since $\alpha(\underline{x}) = 0$ definitionally, we have $y < \bar{x}$. The definition of $\alpha$ and $\alpha(y) < 1$ thus require $(\lambda(\bar{x}) - y)/(\bar{x} - y) < 1$, implying $\lambda(\bar{x}) < \bar{x}$. Thus, $\underline{y} < \lambda(\bar{x})$ by definition of $\underline{y}$.

For $y \in [\underline{x},\underline{y})$, the definition of $\alpha$ requires $\alpha(y) \geq (\lambda(\underline{y}) - y)/(\underline{y} - y)$. Thus $\alpha(y) = 1$ since $\lambda(\underline{y}) = \underline{y}$.

Next, let $y \in (\underline{y},\bar{x}]$. We show $\alpha(y) < 1$ and that $\alpha$ is continuous at $y$; then, we show $\alpha$ is strictly decreasing on a neighborhood of $y$ under the additional assumption

$y < \lambda(\bar{x})$. By definition of $\underline{y}$, it holds $\lambda(y) < y$. Since $\lambda$ is continuous, there is $\varepsilon > 0$ and a neighborhood $U$ of $y$ such that all $z \in U$ satisfy $\alpha(z) = \max\left(0, \max_{x \in [z+\varepsilon, \bar{x}]} \frac{\lambda(x)-z}{x-z}\right)$. Now Berge's Maximum theorem implies that $\alpha$ is continuous at $y$. Further, since $\lambda(x) < x$ for all $x \in (\underline{y}, \bar{x}]$, we have $\alpha(y) < 1$. Now suppose, additionally, $y < \lambda(\bar{x})$. Then clearly $\alpha(y) > 0$. Since also $\lambda(x) < x$ for all $x \in (\underline{y}, \bar{x}]$, the ratio $\frac{\lambda(x)-z}{x-z}$ is strictly decreasing in $z \in U$ for all fixed $x$. Thus $\alpha$ strictly decreases on $U$.

We next show continuity of $e_A$ and $e_P$ on $(\underline{y}, \bar{x}]$. Since $\alpha$ is continuous on $(\underline{y}, \bar{x}]$, the profit $\underline{\pi}(y, \tilde{e}_A)$ is continuous in $(y, \tilde{e}_A)$ for $y \in (\underline{y}, \bar{x}]$. Since $e_A(y)$ maximizes $\underline{\pi}(y, \cdot)$ (Proposition B.3) and $\underline{\pi}(y, \cdot)$ is strictly quasiconcave (Proposition B.4), Berge's Maximum Theorem implies that $e_A$ is continuous. Proposition B.3 now implies that $e_P$ is continuous on $(\underline{y}, \bar{x}]$.

We next show right-continuity at $\underline{y}$. Since $\alpha$ is decreasing, it holds $\alpha(\underline{y}) \geq \lim_{\varepsilon \searrow 0} \alpha(\underline{y} + \varepsilon)$. It is easily verified that $\lim_{\varepsilon \searrow 0} \alpha(\underline{y} + \varepsilon) \geq (\lambda(x) - \underline{y})/(x - \underline{y})$ holds for all $x$ such that $x > \underline{y}$. Since $\alpha(\underline{y}) = \sup_{x \in (\underline{y}, \bar{x}]} (\lambda(x) - \underline{y})/(x - \underline{y})$, we get $\alpha(\underline{y}) = \lim_{\varepsilon \searrow 0} \alpha(\underline{y} + \varepsilon)$. Thus $\alpha$ is right-continuous at $\underline{y}$. Right-continuity of $e_A$ and $e_P$ at $\underline{y}$ now follow from the same argument as in the previous paragraph. □

## B.4 Random audits: Proof of Theorem 3.1

Here, we characterize mechanisms $m$ with random audits for which there exists $\lambda \in \Lambda$ such that $m \in T(\lambda)$. This characterization will imply Theorem 3.1.

Fix $\lambda \in \Lambda$, $m \in T(\lambda)$. To distinguish whether the profit $\underline{\pi}$ is given by $\pi_1$, $\pi_2$ or $\pi_3$, it is useful to define, for all $y \in [\underline{x}, \bar{x}]$ and $\tilde{e}_A \in [0, 1]$,

$$\begin{aligned} d_1(y, \tilde{e}_A) &= y - u_A(\tilde{e}_A) - \lambda(y) \\ d_2(y, \tilde{e}_A) &= (1 - \alpha(y))y - u_A(\tilde{e}_A) - \lambda(y) - \alpha(y)\tau. \end{aligned}$$

Both $d_1(y, \tilde{e}_A)$ and $d_2(y, \tilde{e}_A)$ are decreasing in $\tilde{e}_A$ but increasing in $y$ (since $y - \lambda(y)$ is increasing in $y$, while $\alpha$ is decreasing). Further $d_1(y, \tilde{e}_A) \geq d_2(y, \tilde{e}_A)$ and

$$\begin{aligned} d_1(y, \tilde{e}_A) \leq 0 \quad &\Leftrightarrow \quad \underline{\pi}(y, \tilde{e}_A) = \pi_1(y, \tilde{e}_A); \\ d_2(y, \tilde{e}_A) \leq 0 \leq d_1(y, \tilde{e}_A) \quad &\Leftrightarrow \quad \underline{\pi}(y, \tilde{e}_A) = \pi_2(y, \tilde{e}_A); \\ 0 \leq d_2(y, \tilde{e}_A) \quad &\Leftrightarrow \quad \underline{\pi}(y, \tilde{e}_A) = \pi_3(y, \tilde{e}_A) \quad \Leftrightarrow \quad \alpha(y) \leq \beta(y, \tilde{e}_A). \end{aligned} \tag{24}$$

Next, define (continuing to note the dependence on $\lambda$ and $m$)

$$Y^\circ = \{y \in [\underline{x}, \bar{x}] \colon 0 < e_P(y) < 1\}, \tag{25a}$$
$$Y_0 = \{y \in [\underline{x}, \bar{x}] \colon e_P(y) = 1\}, \tag{25b}$$
$$\text{SuperLow} = \{y \in Y^\circ \colon d_1(y, e_A(y)) < 0\}, \tag{25c}$$
$$\text{Low} = \{y \in Y^\circ \colon d_2(y, e_A(y)) < 0 = d_1(y, e_A(y))\}, \tag{25d}$$
$$\text{Middle} = \{y \in Y^\circ \colon d_2(y, e_A(y)) < 0 < d_1(y, e_A(y))\}, \tag{25e}$$
$$\text{High} = \{y \in Y^\circ \colon d_2(y, e_A(y)) = 0 < d_1(y, e_A(y))\}, \tag{25f}$$
$$\text{SuperHigh} = \{y \in Y^\circ \colon 0 < d_2(y, e_A(y))\}, \tag{25g}$$
$$Y_6 = \{y \in [\underline{x}, \bar{x}] \colon e_P(y) = 0\}, \tag{25h}$$
$$\underline{y} = \max\{y \in [\underline{x}, \bar{x}] \colon \lambda(y) = y\} \tag{25i}$$
$$\bar{y} = \min\{y \in [\underline{x}, \bar{x}] \colon \lambda(y) = \max \lambda\}. \tag{25j}$$

The sets $Y_0, \text{SuperLow}, \ldots, \text{SuperHigh}, Y_6$ partition $[\underline{x}, \bar{x}]$ since $d_2 \leq d_1$.

For disjoint subsets $A$ and $B$ of $\mathbb{R}$, we write $A < B$ to mean that $a < b$ holds for all $a \in A$ and $b \in B$. We write $A < B < C$ to mean $A < B$ and $B < C$ and $A < C$.[21]

**Theorem B.1.** *Let $\lambda \in \Lambda$ and $m \in T(\lambda)$. If $m$ has non-random audits, then*

*(1) It holds $Y_0 <$ SuperLow $<$ Low $<$ Middle $<$ High $<$ SuperHigh $< Y_6$. Each of* SuperLow, …, SuperHigh *has a non-empty interior, and $Y_6$ contains $\bar{x}$. Finally, $\underline{y} = \inf(Y^\circ) = \inf(\text{SuperLow})$ and $\bar{y} = \sup(\text{SuperHigh})$.*[22]

*(2)* The principal's effort $e_P$ *is*

- *(a) constantly* 1 *on* $[\underline{x}, \underline{y})$*;*
- *(b) continuous except possibly at $\underline{y}$, and $e_P$ is continuous from the right at $\underline{y}$;*
- *(c) decreasing on* $[\underline{x}, \bar{x}]$*;*
- *(d) strictly decreasing on* SuperLow $\cup$ Low $\cup$ Middle $\cup$ High*;*
- *(e) constantly* 0 *on* $[\bar{y}, \bar{x}]$.

*(3)* The agent's effort $e_A$ *is*

- *(a) constant on* $[\underline{x}, \inf(\text{SuperLow}))$ *and equal to* $\operatorname{argmin}_{\tilde{e}_A \in [0,1]} u_A(\tilde{e}_A) + c_A(\tilde{e}_A) + (1 - \tilde{e}_A)c_P(1)$.
- *(b) continuous except possibly jumping downwards at* $\inf(\text{SuperLow})$*;*

[21] Note $A < B$ holds vaccuously if $A$ or $B$ is empty. Hence, $A < B$ and $B < C$ need not imply $A < C$ as $B$ could be empty.

[22] The point $\underline{y}$ may be either in $Y_0$ or in SuperLow.

(c) *strictly decreasing on each of* SuperLow, Middle *and* SuperHigh, *but strictly increasing on each of* Low *and* High.

(d) *constantly* 0 *on* $[\sup(\text{SuperHigh}), \bar{x}]$.

(e) *strictly below* $e_A^{\text{eff}}$ *on* $[\underline{x}, \inf(\text{Middle}))$, *equal to* $e_A^{\text{eff}}$ *on* Middle, *strictly above* $e_A^{\text{eff}}$ *on* $(\sup(\text{Middle}), \sup(\text{SuperHigh}))$, *and equal to* $e_A^{\text{eff}}$ *on* $Y_6$.

*(4) The agent's utility* $U_m$ *is v-shaped; specifically, constant on* $[\underline{x}, \inf(\text{SuperLow}))$, *strictly decreasing on* SuperLow, *and strictly increasing on* $[\sup(\text{SuperLow}), \bar{x}]$. *Moreover,* $U_m$ *is bounded away from* 0, *i.e.* $\inf_{y \in [\underline{x}, \bar{x}]} U_m(y) > 0$.

*(5) All* $y \in Y_0 \cup \text{SuperLow}$ *satisfy* $U_m(y) = u_A(e_A(y)) > y - \lambda_m(y)$.
*All* $y \in \text{Low} \cup \text{Middle} \cup \text{High} \cup \text{SuperHigh} \cup Y_6$ *satisfy* $U_m(y) = y - \lambda_m(y)$.

*(6) The principal's profit* $\Pi_m(y)$ *is increasing in* $y$.

Before proving Theorem B.1, we prove Theorem 3.1 from the main text.

*Proof of Theorem 3.1.* If $m$ is tight and has non-random audits, Proposition B.2 implies $\lambda_m \in \Lambda$ and $m \in T(\lambda_m)$. Now define $Y_0, \text{SuperLow}, \ldots, \text{SuperHigh}, Y_6$ as in (25), and apply Theorem 3.1. For the claims regarding the binding incentive constraints, note the constraint binds for a type $y$ if and only if $U_m(y) = y - \lambda_m(y)$. □

*Proof of Theorem B.1.* We proceed in several steps.

**Step 1.** SuperLow $< \ldots <$ SuperHigh

*Proof.* We show $\text{SuperLow} \cup \text{Low} \cup \text{Middle} \cup \text{High} < \text{SuperHigh}$, the other cases being similar. Towards a contradiction, let $z \in \text{SuperLow} \cup \text{Low} \cup \text{Middle} \cup \text{High} <$ SuperHigh and $y \in \text{SuperHigh}$ be such that $y < z$. We first provide the argument assuming $z \in \text{SuperLow}$, later explaining how to adapt the argument to the cases $z \in \text{Low} \cup \text{Middle}$ and $z \in \text{High}$. Since $z \in \text{SuperLow}$ and $y \in \text{SuperHigh}$, it holds $d_1(z, e_A(z)) < 0 < d_2(y, e_A(y))$, meaning the profit at $(z, e_A(z))$ is given by $\pi_1(z, e_A(z))$, and the profit at $(y, e_A(y))$ is given by $\pi_3(y, e_A(y))$. Since both $d_1$ and $d_2$ are increasing in the first argument but decreasing in the second argument, we have $e_A(y) < e_A(z)$. Find $\varepsilon > 0$ sufficiently close to 0 such that $e_A(y) < e_A(y) + \varepsilon < e_A(z) - \varepsilon < e_A(z)$ and $d_1(z, e_A(z) - \varepsilon) < 0 < d_2(y, e_A(y) + \varepsilon)$. Thus, $\underline{\pi}(z, e_A(z) - \varepsilon) = \pi_1(z, e_A(z) - \varepsilon)$ and $\underline{\pi}(y, e_A(y) + \varepsilon) = \pi_3(y, e_A(y) + \varepsilon)$. Since for every type $x$ the effort $e_A(x)$ maximizes $\underline{\pi}(x, \cdot)$ (Proposition B.3), we have $\pi_1(z, e_A(z)) \geq \pi_1(z, e_A(z) - \varepsilon)$ and $\pi_3(y, e_A(y)) \geq \pi_3(y, e_A(y) + \varepsilon)$.

We next claim $\pi_1(y, e_A(y)) \geq \pi_1(y, e_A(y)+\varepsilon)$. Since $y \in \text{SuperHigh}$, we have $\alpha(y) \leq \beta(y, e_A(y))$. Since $\pi_3(y, e_A(y)) \geq \pi_3(y, e_A(y)+\varepsilon)$, invoking (23) from Proposition B.5 yields $\pi_2(y, e_A(y)) \geq \pi_2(y, e_A(y)+\varepsilon)$, and then invoking (22) from Proposition B.5 yields $\pi_1(y, e_A(y)) \geq \pi_1(y, e_A(y)+\varepsilon)$, as claimed.

Since $\pi_1(y, e_A(y)) \geq \pi_1(y, e_A(y)+\varepsilon)$ and $y < z$, submodularity of $\pi_1$ implies $\pi_1(z, e_A(y)) \geq \pi_1(z, e_A(y)+\varepsilon)$. In summary, $\pi_1(z, e_A(y)) \geq \pi_1(z, e_A(y)+\varepsilon)$ and $\pi_1(z, e_A(z)) \geq \pi_1(z, e_A(z)-\varepsilon)$ hold. These two inequalities contradict the strict quasiconcavity of $\pi_1$ (Proposition B.4) since $e_A(y) < e_A(y)+\varepsilon < e_A(z)-\varepsilon < e_A(z)$.

We now explain how to adapt the arguments to the cases $z \in \text{Low} \cup \text{Middle}$ and $z \in \text{High}$. In both cases, the assumptions $z < y$ and $y \in \text{SuperHigh}$ imply $e_A(y) < e_A(z)$. If $z \in \text{Low} \cup \text{Middle}$, then for $\varepsilon > 0$ sufficiently small we have $\underline{\pi}(z, e_A(z)) = \pi_2(z, e_A(z))$ and $\underline{\pi}(z, e_A(z)-\varepsilon) = \pi_2(z, e_A(z)-\varepsilon)$. We can establish $\pi_2(y, e_A(y)) \geq \pi_2(y, e_A(y)+\varepsilon)$ via the same argument as above, and then obtain a contradiction to the strict quasiconcavity of $\pi_2$. Finally, if $z \in \text{High}$, then a simpler argument yields a contradiction to the strict quasiconcavity of $\pi_3$. □

**Step 2.** *On* $\text{SuperLow} \cup \ldots \cup \text{High}$*, the principal's effort is strictly decreasing. On* SuperHigh*, the principal's effort is decreasing.*

*Proof.* We know from (24) that $\alpha(y) \geq \beta(y, e_A(y))$ holds for all $y \in \text{SuperLow} \cup \ldots \cup \text{High}$. Thus $e_P = \alpha$ on $\text{SuperLow} \cup \ldots \cup \text{High}$ (Proposition B.3). Thus Lemma B.5 implies that $e_P$ strictly decreases on $\text{SuperLow} \cup \ldots \cup \text{High}$.

Since $d_2(y, e_A(y)) > 0$ for all $y \in \text{SuperHigh}$, using (24) and Proposition B.3 we conclude $e_P(y) = \beta(y, e_A(y))$ for all $y \in \text{SuperHigh}$. To show that $e_P$ decreases on SuperHigh, it suffices to show the following: *if* $z, y \in X$ *are such that* $z < y$, $d_2(y, e_A(y)) > 0$, *and* $d_2(z, e_A(z)) > 0$, *then* $\beta(y, e_A(y)) \leq \beta(z, e_A(z))$. Towards a contradiction, let $\beta(y, e_A(y)) > \beta(z, e_A(z))$. Note $\lambda(z) \leq \lambda(y)$ since $\lambda$ is increasing. Thus $\lambda(y)+u_A(e_A(y)) < \lambda(z)+u_A(e_A(z))$. This strict inequality requires $e_A(y) < e_A(z)$. Thus, $e_A(y)+\varepsilon < e_A(z)-\varepsilon$ and $\lambda(y)+u_A(e_A(y)+\varepsilon) < \lambda(z)+u_A(e_A(z)-\varepsilon)$ for all sufficiently small $\varepsilon > 0$. According to Lemma B.4, therefore, all such $\varepsilon$ satisfy $\partial_2\pi_3(y, e_A(y)+\varepsilon) < \partial_2\pi_3(z, e_A(z)-\varepsilon)$. Hence, all sufficiently small $\varepsilon > 0$ satisfy

$$\pi_3(y, e_A(y)+\varepsilon) - \pi_3(y, e_A(y)) > \pi_3(z, e_A(z)) - \pi_3(z, e_A(z)-\varepsilon).$$

Since $d_2(y, e_A(y)) > 0$ and $d_2(z, e_A(z)) > 0$, also $d_2(y, e_A(y)+\varepsilon) > 0$ and $d_2(z, e_A(z)-$

$\varepsilon) > 0$ for sufficiently small $\varepsilon$. Thus

$$\begin{aligned} 0 \geq \underline{\pi}(y, e_A(y) + \varepsilon) - \underline{\pi}(y, e_A(y)) &= \pi_3(y, e_A(y) + \varepsilon) - \pi_3(y, e_A(y)) \\ &> \pi_3(z, e_A(z)) - \pi_3(z, e_A(z) - \varepsilon) \\ &= \underline{\pi}(z, e_A(z)) - \underline{\pi}(z, e_A(z) - \varepsilon). \end{aligned}$$

But Proposition B.3 asserts that $e_A(z)$ maximizes $\underline{\pi}(z, \cdot)$; contradiction. □

**Step 3.** *The set* $\{y \in [\underline{x}, \bar{x}] \colon \alpha(y) \in (0,1)\}$ *is non-empty.*

This step uses the assumption that $m$ has random audits, i.e. $Y^\circ$ is non-empty.

*Proof.* Towards a contradiction, let $\{y \in [\underline{x}, \bar{x}] \colon \alpha(y) \in (0,1)\}$ be empty. Inspecting the definition of $\alpha$, one may verify that $\lambda(y) = \min(y, \underline{y})$ holds for all $y \in [\underline{x}, \bar{x}]$. By inspection, we thus have $\alpha(y) = \mathbf{1}(y < \underline{y})$ and $\beta(y, \tilde{e}_A) = 0$ for all $y \in [\underline{y}, \bar{x}]$ and all $\tilde{e}_A \in [0,1]$. Since $e_P(y) = \max(\alpha(y), \beta(y, e_A(y)))$ for all $y$ (Proposition B.3), we infer that $e_P$ maps to $\{0,1\}$, meaning $Y^\circ$ is empty; contradiction. □

**Step 4.** *It holds* $\alpha(y) = 1$ *if and only if* $y \in Y_0$. *Further,* $\alpha$, $e_A$, *and* $e_P$ *are all continuous at each point in* $[\underline{x}, \bar{x}] \setminus Y_0$, *and right-continuous at* $\underline{y}$. *Further,* $\underline{y} = \inf Y^\circ = \inf \text{SuperLow}$.

*Proof.* Recall $e_P(y) = \max(\alpha(y), \beta(y, e_A(y)))$ (Proposition B.3) and $\beta < 1$ hold. Thus, $\alpha$ and $e_P$ are both constantly 1 on $Y_0$, while on $Y_0$ both $\alpha$ and $e_P$ are interior. Thus the continuity claims follow from Lemma B.5. Lemma B.5 also asserts that $\alpha$ equals 1 on $[\underline{x}, \underline{y})$ and is interior on $(\underline{y}, \bar{x}]$. Thus $\underline{y} = \inf Y^\circ$. Since SuperLow $< \ldots <$ SuperHigh, also $\inf \text{SuperLow} = \inf Y^\circ = \underline{y}$. □

Since $\alpha$ is decreasing and SuperLow $< \ldots <$ SuperHigh holds, we also find $Y_0 < \ldots < Y_6$ and that each of $Y_0, \text{SuperLow}, \ldots, \text{SuperHigh}, Y_6$ is an (empty or non-empty) interval.

**Step 5.** *For all* $y \in [\underline{x}, \bar{x}]$, *it holds* $e_P(y) > 0$ *if and only if* $e_A(y) > 0$.

The argument is expected and uses the assumption $c_A'(0) = c_P'(0) = 0$.

*Proof.* If $0 < e_P(y)$ and $e_A(y) = 0$ (resp. if $0 = e_P(y)$ and $e_A(y) > 0$), then for all $i \in \{1,2,3\}$ the derivative $\partial_2 \pi_i(y, \tilde{e}_A)$ is strictly positive (resp. negative) for all $\tilde{e}_A$ sufficiently close to $e_A(y)$, contradicting that $e_A(y)$ maximizes $\underline{\pi}(y, \cdot)$. □

**Step 6.** *Each of the sets* SuperLow, ..., SuperHigh *is non-empty. Moreover,* $\lambda(\bar{x}) \in$ SuperHigh $\cup Y_6$.

*Proof.* Since $\alpha$ is strictly positive on $(\underline{y}, \lambda(\bar{x})]$, it holds $d_1(y, e_A(y)) > d_2(y, e_A(y))$ for all $y \in (\underline{y}, \lambda(\bar{x}))$. Further, $\alpha$ and $e_A$ are continuous on $(\underline{y}, \lambda(\bar{x}))$. Thus, it will follow that SuperLow $\cap(\underline{y}, \lambda(\bar{x})), \ldots,$ SuperHigh $\cap(\underline{y}, \lambda(\bar{x}))$ are all non-empty if we can show $d_1(\underline{y}, e_A(\underline{y})) < 0$ and $d_2(\lambda(\bar{x}), e_A(\lambda(\bar{x}))) > 0$.

Consider $\underline{y}$. Since $\alpha(\underline{y}) > 0$, we have $e_P(\underline{y}) > 0$. By a previous step, hence $e_A(y) > 0$. Since $\lambda(\underline{y}) = \underline{y}$, we conclude $\underline{y} < \lambda(\underline{y}) + u_A(e_A(\underline{y}))$, i.e. $d_1(\underline{y}, e_A(\underline{y})) < 0$.

Consider $\lambda(\bar{x})$. Towards a contradiction, let $d_2(\lambda(\bar{x}), e_A(\lambda(\bar{x}))) \leq 0$, i.e. $(1 - \alpha(\lambda(\bar{x})))\lambda(\bar{x}) \leq \lambda(\lambda(\bar{x})) + u_A(e_A(\lambda(\bar{x}))) + \alpha(\lambda(\bar{x}))\tau$. Thus $\alpha(\lambda(\bar{x})) \geq \beta(\lambda(\bar{x}), e_A(\lambda(\bar{x})))$, and thus $e_P(\lambda(\bar{x})) = \alpha(\lambda(\bar{x}))$. Lemma B.5 implies $\lambda(\lambda(\bar{x})) < \lambda(\bar{x})$ and $\alpha(\lambda(\bar{x})) = 0$, and hence an earlier step implies $e_A(\lambda(\bar{x})) = 0$. Thus $(1 - \alpha(\lambda(\bar{x})))\lambda(\bar{x}) > \lambda(\lambda(\bar{x})) + u_A(e_A(\lambda(\bar{x}))) + \alpha(\lambda(\bar{x}))\tau$; contradiction.

Finally, the previous paragraph also establishes $d_2(\lambda(\bar{x}), e_A(\lambda(\bar{x}))) > 0$, meaning $\lambda(\bar{x}) \in$ SuperHigh $\cup Y_6$. □

**Step 7.** *The agent's effort* $e_A$ *is strictly decreasing on each of* SuperLow *and* Middle*, decreasing on* SuperHigh*, but strictly increasing on each of* Low *and* High*. Further, the agent's effort is constant on* $Y_0$*, and constantly* 0 *on* $Y_6$.

A later step establishes strict monotonicity of $e_A$ on SuperHigh.

*Proof.* Since $e_P$ is constantly 0 on $Y_6$, also $e_A$ is constantly 0 on $Y_6$.

Consider SuperLow. Since $d_1$ and $d_2$ are continuous, for all $y \in$ SuperLow if $\varepsilon > 0$ is sufficiently close to 0 then $\underline{\pi}(y, e_A(y) \pm \varepsilon) = \pi_1(y, e_A(y) \pm \varepsilon)$. We know $e_A(y) < 1$ (from Proposition B.3) and $0 < e_A(y)$ (from a previous step, since $0 < e_P(y)$). Thus $e_A(y)$ satisfies the first-order condition $u_A'(e_A(y)) + c_A'(e_A(y)) = c_P(\alpha(y))$. In an earlier claim we established that $\alpha$ is strictly decreasing on SuperLow. Assumption 2 implies that $u_A' + c_A'$ is strictly increasing. Thus $e_A$ is strictly decreasing on SuperLow.

A similar argument shows that $e_A$ is strictly decreasing on Middle (the first order condition is now $c_A'(e_A(y)) = c_P(\alpha(y))$).

Consider SuperHigh. Since $e_A$ is continuous on SuperHigh, it suffices to show that for all $y \in$ SuperHigh there is a neighborhood $U$ of $y$ such that $e_A$ is decreasing on $U$. To that end, since $d_1$, $d_2$, and $e_A$ are continuous (in their respective arguments), there is a neighborhood $U$ of $y$ such that $d_2(x, e_A(z)) \neq 0$ for all $x$ and $z$ in $U$.

Thus (24) implies that $\underline{\pi}(x, e_A(z)) = \pi_3(x, e_A(z)) \leq \pi_3(x, e_A(x)) = \underline{\pi}(x, e_A(x))$ for all $x, z \in U$. Now let $x$ and $z$ be in $U$ and such that $z < x$. We show $e_A(z) \geq e_A(x)$. Towards a contradiction, let $e_A(z) < e_A(x)$. We know $\pi_3(z, e_A(z)) \geq \pi_3(z, e_A(x))$ and $\pi_3(x, e_A(x)) \geq \pi_3(x, e_A(z))$. Since $\pi_3$ is submodular, also $\pi_3(x, e_A(z)) \geq \pi_3(x, e_A(x))$. Thus $\pi_3(x, e_A(x)) = \pi_3(x, e_A(z))$. Since $e_A(z) < e_A(x)$, this equation contradicts the strict quasiconcavity of $\pi_3(x, \cdot)$.

Next, consider Low. By definition of Low, all $y \in \text{Low}$ satisfy $u_A(e_A(y)) = y - \lambda(y)$. Recall also $\lambda(\underline{x}) = \underline{x}$ (since $\lambda \in \Lambda$) while $\lambda(y) < y$ holds on Low (Lemma B.5). Since $y - \lambda(y)$ is increasing in $y$ and $\lambda$ is concave, it follows that $y - \lambda(y)$ is strictly increasing on Low. Since $u_A(e_A(y)) = y - \lambda(y)$ for all $y \in \text{Low}$, agent-effort $e_A$ is also strictly increasing on Low.

Finally, consider High. By definition of High, all $y \in \text{High}$ satisfy $\frac{u_A(e_A(y))}{y} = 1 - \alpha(y) - \frac{\alpha(y)\tau}{y} - \frac{\lambda(y)}{y}$. We know that $\lambda(y)/y$ is decreasing (Appendix B.3.1). Hence, it will follow that $e_A$ is strictly increasing on High if we can show that $\alpha$ is strictly decreasing on High. From a previous step, we know $\lambda(\bar{x}) \in \text{SuperHigh} \cup Y_6$, and we know $\text{High} < \text{SuperHigh} \cup Y_6$. From a different step we also $\underline{y} = \inf(\text{SuperLow})$ and $\text{SuperLow} < \text{High}$. In particular, $\underline{y} < y < \lambda(\bar{x})$ for all $y \in \text{High}$. Lemma B.5 thus implies $\alpha$ is strictly decreasing on High. □

**Step 8.** *Each of the sets* SuperLow, ..., SuperHigh *has a non-empty interior.*

*Proof.* Since $e_A$ and $\alpha$ are continuous on $\text{SuperLow} \cup \ldots \cup Y_6$, it follows that each of SuperLow, Middle, and SuperHigh is open in $[\underline{x}, \bar{x}]$, and we already know that these sets are non-empty. It remains to show that Low and High admit non-empty interiors. We do so for High, the argument for Low being similar.

Since High is a non-empty interval, it suffices to show High is not a singleton. Towards a contradiction, let High be a singleton $x$. Since Middle and SuperHigh, there is a sequence $(z_n)_n$ in Middle and a sequence $(y_n)_n$ in SuperHigh such that both sequences converge to $x$. As in a previous step, for all $n$ the effort $e_A(z_n)$ satisfies the first-order condition:

$$0 = c_P(\alpha(z_n)) - c'_A(e_A(z_n)) = c_P(e_P(z_n)) - c'_A(e_A(z_n)),$$

where we used that $\alpha$ and $e_P$ agree on Middle. We next derive a similar first-order condition for $y_n$. Indeed, it holds $\underline{\pi}(y_n, e_A(y_n)+\varepsilon) = \pi_3(y_n, e_A(y_n)+\varepsilon)$ for all sufficiently

small $\varepsilon > 0$. Since $e_A(y_n)$ maximizes $\underline{\pi}(y_n, \cdot)$, there is a sequence $(\varepsilon_{n,k})_{k\in\mathbb{N}}$ converging to 0 from above and such that $0 \leq \partial_2\pi_3(y_n, e_A(y_n) + \varepsilon_{n,k})$ holds for all $k$. Therefore

$$\begin{aligned}0 \geq& \partial_2\pi_3(y_n, e_A(y_n) + \varepsilon_{n,k})\\ \geq& c_P(\beta(y_n, e_A(y_n) + \varepsilon_{n,k})) - c'_A(e_A(y_n) + \varepsilon_{n,k})\\ &+ (1 - e_A(z_n))\frac{c'_P(\beta(y_n, e_A(y_n) + \varepsilon_{n,k}))u'_A(e_A(y_n) + \varepsilon_{n,k})}{\bar{x} + \tau}.\end{aligned}$$

We now take $k \to \infty$ and recall $e_P(y_n) = \beta(y_n, e_A(y_n))$ to find

$$0 \geq c_P(e_P(y_n)) - c'_A(e_A(y_n)) + (1 - e_A(y_n))\frac{c'_P(e_P(y_n))u'_A(e_A(y_n))}{\bar{x} + \tau},$$

Both $e_P$ and $e_A$ are continuous at $x$. Thus:

$$0 \geq (1 - e_A(x))c'_P(e_P(x))u'_A(e_A(x)).$$

Thus $e_A(x) = 1$ or $e_A(x) = 0$ or $e_P(x) = 0$. However, we know $e_P(x) > 0$ (since $x \in$ High). Hence also $e_A(x) > 0$ (by a previous step). From $e_A(x) > 0$ we also get $u'_A(e_A(x)) > 0$. Finally, Proposition B.3 asserts $e_A(x) < 1$. Contradiction. □

**Step 9.** *The function $\lambda$ is strictly increasing on $Y_0 \cup \ldots \cup$ SuperHigh, and constantly equal to $\max\lambda$ on $Y_6$.*

*Proof.* Denote $x = \sup$ SuperHigh. Since $\lambda$ is increasing and concave, it suffices to show that $x$ is the smallest type satisfying $\lambda(x) = \max\lambda$. By continuity, $e_P(x) = 0$, and thus $e_A(x) = 0$. Since $e_P$ agrees with $y \mapsto \beta(y, e_A(y))$ for $y \in$ SuperHigh, we get $\beta(x, 0) = 0$, and thus $\lambda(x) = \max\lambda$ (by inspecting the definition of $\beta$). Conversely, if $x'$ is in SuperHigh but strictly less than $x$, then $\beta(x', e_A(x')) = e_P(x') > 0$, and hence $\lambda(x') < \max\lambda$. □

**Step 10.** *It holds $e_P(\bar{x}) = 0$. On* SuperHigh*, the principal's effort $e_P$ is non-constant and the agent's effort $e_A$ is strictly decreasing.*

*Proof.* First, $e_P(\bar{x}) = 0$ follows easily by inspecting the definitions of $\alpha$ and $\beta$, and then invoking Proposition B.3. Thus $\bar{x} \in Y_6$. The principal's effort $e_P$ is continuous on the interval SuperHigh $\cup Y_6$, strictly positive on SuperHigh, and 0 on $Y_6$. Thus, $e_P$ is non-constant on SuperHigh.

Now consider the claim regarding $e_A$. We already know that $e_A$ is decreasing on SuperHigh. Towards a contradiction, let $[z_0, z_1]$ be a non-degenerate subinterval in the interior of SuperHigh on which $e_A$ constantly equals $e_A(z_0)$. We already know that $e_A$ is interior on SuperHigh. As in previous steps of this proof, for all $y \in [z_0, z_1]$, if $\varepsilon$ is sufficiently close to 0, then $\underline{\pi}(y, e_A(z_0) \pm \varepsilon) = \pi_3(y, e_A(z_0) \pm \varepsilon)$.

In an auxiliary step, we argue that for almost all $y \in [z_0, z_1]$ the function $\tilde{e}_A \mapsto \pi_3(y, \tilde{e}_A)$ is differentiable at $\tilde{e}_A = e_A(z_0)$. To that end, it suffices to show check differentiability of $\tilde{e}_A \mapsto \beta(y, \tilde{e}_A)$ at $\tilde{e}_A = e_A(z_0)$ for almost all $y \in [z_0, z_1]$. Recall, for all $y \in [z_0, z_1]$ and $\tilde{e}_A \in [0, 1]$ the definitions:

$$\beta(y, \tilde{e}_A) = \max_{x \in [\underline{x}, \bar{x}]} \frac{\lambda(x) - \lambda(y) - u_A(\tilde{e}_A)}{x + \tau},$$
$$\hat{Z}(\lambda(y) + u_A(\tilde{e}_A))) = \operatorname*{argmax}_{x \in [\underline{x}, \bar{x}]} \frac{\lambda(x) - (\lambda(y) + u_A(\tilde{e}_A))}{x + \tau},$$

where we already used that $e_P(y)$ is strictly positive, and hence that $\beta(y, \tilde{e}_A)$ is strictly positive for $\tilde{e}_A$ close to $e_A(z_0)$. For a point $y$ such that $\hat{Z}(\lambda(y)+u_A(e_A(z_0))$ is a singleton, the Envelope Theorem (Milgrom and Segal, 2002, Theorem 3) implies that $\beta(y, \tilde{e}_A)$ is differentiable in $\tilde{e}_A$ at $\tilde{e}_A = e_A(z_0)$. Thus we show that for almost all $y \in [z_0, z_1]$ the set $\hat{Z}(\lambda(y) + u_A(e_A(z_0))$ is a singleton. For all $y$, the set $\hat{Z}(\lambda(y) + u_A(e_A(z_0))$ is an interval since $\lambda$ is concave and lies below the affine function $x \mapsto \beta(y, e_A(z_0))(x + \tau) + \lambda(y) + u_A(e_A(z_0))$. Since $\lambda$ is strictly increasing on $[z_0, z_1]$, a routine argument verifies that if $y < y'$, then $\max \hat{Z}(\lambda(y) + u_A(e_A(z_0)) < \min \hat{Z}(\lambda(y') + u_A(e_A(z_0))$. Consequently, for all $y \in [z_0, z_1]$, if $\hat{Z}(\lambda(y) + u_A(e_A(z_0))$ is a non-degenerate interval, then its interior contains a rational number that is not contained in an interval of the collection $\{\hat{Z}(\lambda(y') + u_A(e_A(z_0)) \colon y' \in [z_0, z_1] \setminus \{y\}\}$. It follows that there are at most countably many $y \in [z_0, z_1]$ for which $\hat{Z}(\lambda(y) + u_A(e_A(z_0))$ is non-degenerate.

Find $y, y' \in [z_0, z_1]$ such that $y < y'$ and $\tilde{e}_A \mapsto \pi_3(z, \tilde{e}_A)$ and $\tilde{e}_A \mapsto \pi_3(z', \tilde{e}_A)$ are differentiable at $e_A(z_0)$. This is possible by the claim just proven and since, by assumption, the interval $[z_0, z_1]$ is non-degenerate. Since $e_A(z_0)$ maximizes $\pi_3(y, \cdot)$ (on a neighborhood of $e_A(z_0)$), the following first-order condition holds:

$$0 = c_P(\beta(y, e_A(z_0))) - c'_A(e_A(z_0)) + (1 - e_A(z_0)) \frac{c'_P(\beta(y, e_A(z_0)))u'_A(e_A(z_0))}{\hat{z}(\lambda(y) + u_A(e_A(z_0)) + \tau}.$$

An analogous first-order condition holds for $y'$ (obtained by replacing all instances of

$y$ with $y'$). Since $\lambda$ is strictly increasing on $[z_0, z_1]$, we have $\hat{z}(\lambda(y) + u_A(e_A(z_0)) + \tau \leq \hat{z}(\lambda(y') + u_A(e_A(z_0)) + \tau$ and $\beta(y, e_A(z_0)) > \beta(y', e_A(z_0))$. Consequently, the first-order condition cannot hold for both $y$ and $y'$; contradiction. □

It remains to show the claims regarding the efficient agent-effort, the agent's interim utility, and the principal's profit. The effort $e_A(y)$ is strictly below (resp. strictly above) the efficient agent-effort $e_A^{\text{eff}}(y)$ for $y \in Y_0 \cup \text{SuperLow}$ (resp. $y \in \text{SuperHigh}$) as one can show by inspecting the first-order conditions for maximizing $\underline{\pi}(y, \cdot)$. On Middle, the two efforts are equal, by inspecting the first-order conditions. On $Y_6$ $(= [\sup(\text{SuperHigh}), \bar{x}])$ the two efforts both equal 0. On the interior of Low, we have $e_A < e_A^{\text{eff}}$ since on this interval $e_A$ is strictly increasing (as proven earlier), $e_A^{\text{eff}}$ is strictly decreasing, and the two coincide at the top of the interval (i.e., at $\inf(\text{Middle})$). By a similar argument, $e_A > e_A^{\text{eff}}$ on the interior of High.

The principal's profit is increasing since the profit at each type $y$ is given by $\max_{\tilde{e}_A \in [0,1]} \underline{\pi}(y, \tilde{e}_A)$, where $\underline{\pi}(y, \tilde{e}_A)$ is increasing in $y$. Finally, Lemma B.2 shows that the agent's utility is given by $U_m(y) = \max(u_A(e_A(y)), y - \lambda(y))$ for all $y$. For $y \in Y_0$, using $e_P(y) = \alpha(y) = 1$, one may verify $e_A(y)$ is constantly equal to the unique minimizes $\operatorname{argmin}_{\tilde{e}_A \in [0,1]} \pi_1(y, 1)$, and is hence strictly positive; since also $\lambda(y) = y$, we have $U_m(y) = u_A(e_A(y)) > y - \lambda(y)$ and $U_m(y)$ is constant in $y$ for in $y \in Y_0$. For $y \in \text{SuperLow} \cup \ldots \cup \text{SuperHigh}$, the claims regarding $U_m$ follow from the definitions of the intervals and the properties of $e_A$ and $\lambda$ established in earlier steps. The point $\underline{y}$ $(= \inf(\text{SuperLow}))$ is either in $Y_0$ or SuperLow, and hence $U_m(\underline{y}) = u_A(e_A(\underline{y})) > \underline{y} - \lambda(\underline{y})$. On $y \in Y_6$, we know $\lambda(y) = \max \lambda$ and $e_A(y) = 0$, and thus $U_m(y) = y - \max \lambda$. Finally, these claims imply that $U_m$ is minimized at $y = \sup(\text{Low})$ and there given by $u_A(e_A(y)$; since $e_A(y) > 0$, we have $U_m(y) > 0$; in particular, $U_m$ is bounded away from 0 across all types. □

## B.5 Non-random audits: Proof of Theorem 3.2

Proposition B.2 asserts every tight mechanism $m$ satisfies $m \in T(\lambda_m)$ and $\lambda_m \in \Lambda$. Hence, Theorem 3.2 will follow from the following theorem that characterizes mechanisms with non-random audits in the image of $\Lambda$ under $T$.

**Theorem B.2.** *Let $m$ be a mechanism. The following are equivalent.*

*(1) Mechanism $m$ has non-random audits and there is $\lambda \in \Lambda$ such that $m \in T(\lambda)$.*

*(2) There is a face value $y_0 \in [\underline{x}, \bar{x}]$ such that $m$ is a debt-with-relief mechanism (with face value $y_0$) with relief $\bar{r}_A = c_A'(\bar{e}_A)$ and for all $x$ agent-effort $e_A(x)$ given by $e_A(x) = \bar{e}_A \mathbf{1}_{(x \in [\underline{x}, y_0))}$, where $\bar{e}_A = \operatorname{argmin}_{\tilde{e}_A \in [0,1]} \tilde{e}_A c_A'(\tilde{e}_A) + (1 - \tilde{e}_A) c_P(1) > 0$. Moreover, $\lambda_m(x) = \min(x, y_0)$ for all $x$.*

*(3) There is $y_0 \in [\underline{x}, \bar{x}]$ such that, defining $\lambda(y) = \min(y, y_0)$ for all $y$, it holds $m \in T(\lambda)$.*

*Proof of Theorem 3.2.* We show (1) implies (2), the other claims being similar. Recall $e_P(y) = \max(\alpha(y), \beta(y, e_A(y))$ for all $y \in [\underline{x}, \bar{x}]$ and $\beta < 1$. Since $\alpha \geq 0$ and $e_P$ maps to $\{0, 1\}$, we find $e_P = \alpha$. Since $\alpha$ is decreasing, there is $y_0 \in \mathbb{R}$ such that $\alpha(y) = 1$ for all $y \in [\underline{y}, y_0)$, and $\alpha(y) = 0$ for all $y \in (y_0, 1]$. Using this formula for $\alpha$, one may verify that $\lambda(y) = \min(y, y_0)$ holds for all $y$. Hence also $\alpha(y_0) = 0$. Thus, $e_P(y) = \mathbf{1}_{(y < y_0)}$.

We next characterize $e_A(y)$, for arbitrary $y$. Recall that $e_A(y)$ maximizes $\underline{\pi}(y, \cdot)$ (Proposition B.3). If $y \geq y_0$, using $e_P(y) = 0$, it is easy to check that $e_A(y) = 0$ uniquely maximizes $\underline{\pi}(y, \cdot)$. For $y < y_0$, using $e_P(y) = 1$ we obtain $\underline{\pi}(y, \tilde{e}_A) = y - u_A(\tilde{e}_A) - c_A(\tilde{e}_A) - (1 - \tilde{e}_A) c_P(1)$ for all $\tilde{e}_A \in [0, 1]$, and hence $e_A(y) = \bar{e}_A$.

We next consider the refund $r_A$ and the induced loss. Fix $y$. Proposition B.3 asserts $\Pi_m(y) = \underline{\pi}(y, e_A(y))$. Spelling out this equation shows:

$$e_P(y) r_P(y) + (1 - e_P(y)) r_\emptyset(y) = \min\left(y - u_A(\tilde{e}_A), \lambda(y)\right)$$

Using the formula for $e_P$, we thus find $r_P(y) = 0$ for $y \in [\underline{x}, y_0)$, and $r_\emptyset(y) = y - y_0$ for $y \in [y_0, \bar{x}]$. Thus, for all $y \in [\underline{x}, y_0)$ the refund $r_A(y)$ is given by $\bar{r}_A = c_A'(\bar{e}_A)$ in order to incentivize effort $\bar{e}_A$. Thus, $m$ is debt-with-relief with threshold $y_0$ and relief $c_A'(\bar{e}_A)$. Direct computation now shows $\lambda_m(y) = \lambda(y) = \min(y, y_0)$. □

# C How to tighten a mechanism?

The following lemma is essential for our proof of Theorem 3.3.

**Lemma C.1.** *Let $\lambda, \lambda^* \in \Lambda$, let $m \in T(\lambda)$, and let $m^* \in T(\lambda^*)$. If $(\Pi_m, \lambda) \leq (\Pi_{m^*}, \lambda^*)$, then $(\Pi_m, \lambda) = (\Pi_{m^*}, \lambda^*)$.*

For a moment, let us assume this lemma. Theorem 3.3 is claim (1) of the following:

**Theorem C.1.** *Let $m$ be a mechanism.*

*(1) If $m^*$ is obtained through steps (1) to (4) of the tightening algorithm applied to $m$, then $m^*$ is tight and tighter than $m$.*

*(2) Mechanism $m$ is tight if and only if there exists $\lambda \in \Lambda$ such that $m \in T(\lambda)$; in this case, $\lambda = \lambda_m$.*

*Proof of Theorem C.1.* Claim (1) is a corollary of claim (2) and Proposition B.1. We also already know from Proposition B.2 that if $m$ is tight, then $\lambda_m \in \Lambda$ and $m \in T(\lambda_m)$. Thus, it remains to show: *if $\lambda \in \Lambda$ such that $m \in T(\lambda)$, then $m$ is tight and $\lambda_m = \lambda$.*

To show that $m$ is tight, let $m'$ be a mechanism such that $(\Pi_m, \lambda_m) \leq (\Pi_{m'}, \lambda_{m'})$. We show $(\Pi_m, \lambda_m) = (\Pi_{m'}, \lambda_{m'})$. We know $\lambda \leq \lambda_m$ (Lemma B.1). By invoking Proposition B.1, find $\tilde{\lambda} \in \Lambda$ and $m^* \in T(\tilde{\lambda})$ such that $(\Pi_{m'}, \lambda_{m'}) \leq (\Pi_{m^*}, \tilde{\lambda})$ (i.e., apply the tightening operator to $m'$). Thus, $(\Pi_m, \lambda) \leq (\Pi_{m^*}, \tilde{\lambda})$. Crucially, Lemma C.1 now implies $(\Pi_m, \lambda) = (\Pi_{m^*}, \tilde{\lambda})$. Thus, also $(\Pi_m, \lambda_m) = (\Pi_{m'}, \lambda_{m'})$.

The identity $\lambda = \lambda_m$ follows by simply applying the arguments from the previous paragraph to $m' = m$. □

It remains to prove Lemma C.1. Key to the proof are the binding incentive constraints, defined next.

## C.1 Binding incentive constraints

Given $\lambda \in \Lambda$ and $m \in T(\lambda)$, recall the definition $\underline{y} = \max\{y \in [\underline{x}, \bar{x}] \colon \lambda(y) = y\}$. Let $\bar{y} = \min\{y \in [\underline{x}, \bar{x}] \colon \lambda(y) = \max \lambda\}$, so that $\underline{y} \leq \bar{y}$. Theorems B.1 and B.2 show that $e_P$ is constantly 1 below $\underline{y}$, interior on $(\underline{y}, \bar{y})$, and constantly 0 above $\bar{y}$. If $m$ has random audits, then $e_P(\underline{y}) > 0$; if $m$ has non-random audits, then $e_P(\underline{y}) = 0$. From Definition B.2, it follows that for all $x$ and $y$ such that $y \leq x$ it holds

$$\lambda(x) \leq e_P(y)x + \min\left((1 - e_P(y))y, \lambda(y) + u_A(e_A(y)) + e_P(y)\tau\right). \tag{26}$$

We define $y$'s binding IC types as the set of types $x$ satisfying (26) with equality.

**Definition C.1** (Binding ICs)**.** Let $\lambda \in \Lambda$ and $m \in T(\lambda)$. For all $y \in [\underline{y}, \bar{x}]$, define

$$\begin{aligned}
\phi(y) &= \min\left((1 - e_P(y))y, \lambda(y) + u_A(e_A(y)) + e_P(y)\tau\right), \\
\hat{X}(y) &= \{x \in [y, \bar{x}] \colon \lambda(x) = e_P(y)x + \phi(y)\}.
\end{aligned}$$

The set $\hat{X}(y)$ is non-empty for all $y \in [\underline{y}, \bar{x}]$ such that $e_P(y) > 0$ since the principal's effort $e_P(y)$ is set as low as possible subject to (26).[23] If $e_P(y) = 0$ (i.e., $y \geq \bar{y}$), then $\hat{X}(y)$ is simply the interval $[y, \bar{x}]$ of types above $y$ (since $\lambda$ is constant above $\bar{y}$). Notice that $y \leq \min \hat{X}(y)$ holds definitionally for all $y$.

In view of (26), concavity and continuity of $\lambda$ imply that $\hat{X}(y)$ is a non-empty compact interval; moreover, the correspondence $\hat{X}$ is upper hemicontinuous since $\lambda$, $e_A$ and $e_P$ are right-continuous at $\underline{y}$ and continuous on $(\underline{y}, \bar{x}]$ (Theorems B.1 and B.2)

The next lemma establishes crucial properties of $\hat{X}$.

**Lemma C.2.** *Let $\lambda \in \Lambda$, $m \in T(\lambda)$. Let $\lambda_m$ be the loss function induced by $m$.*

*(1) If $A$ is a non-empty closed subinterval of $[\underline{y}, \bar{y}]$, then the image $\hat{X}(A)$ is a non-empty interval, where $\hat{X}(A)$ means the union $\hat{X}(A) = \cup_{y \in A} \hat{X}(y)$.*

*(2) If $z, y \in [\underline{y}, \bar{y})$ and $z < y$, then, if $e_P(z) > e_P(y)$, then $\max \hat{X}(z) \leq \min \hat{X}(y)$; if $e_P(z) = e_P(y)$, then $\hat{X}(z) = \hat{X}(y)$*

*(3) All $y \in (\underline{y}, \bar{y})$ satisfy $y < \min \hat{X}(y)$.*

*(4) For all $x \in [\underline{y}, \bar{y}]$ there exists $y \in [\underline{y}, x]$ such that $x \in \hat{X}(y)$. Moreover, the image $\hat{X}([\underline{y}, x])$ equals the interval $[\underline{y}, \max \hat{X}(x)]$.*

*Proof of Lemma C.2.* Let $A$ be a non-empty closed subinterval of $[\underline{y}, \bar{y}]$. Since $\hat{X}$ is upper hemicontinuous and its values are non-empty compact intervals, the following fact (de Clippel (2008, Lemma 2)) implies that $\hat{X}(A)$ is a non-empty interval: *Let $[a_0, a_1]$ be an interval in $\mathbb{R}$, let $s \in \mathbb{R}$ and let $\Psi \colon [a_0, a_1] \twoheadrightarrow \mathbb{R}$ be a correspondence with non-empty convex values and a compact graph. If there are $s_0 \in \Psi(a_0)$ and $s_1 \in \Psi(a_1)$ such that $s_0 \leq s \leq s_1$, then there is $a \in [a_0, a_1]$ such that $s \in \Psi(a)$.*

Next, let $y, z \in [\underline{y}, \bar{y})$ and $z < y$. Recall that $e_P$ is decreasing, and so $e_P(z) \geq e_P(y)$. If $e_P(z) = e_P(y)$, then also $\phi(y) = \phi(z)$ (since, else, one of $\hat{X}(y)$ and $\hat{X}(z)$ would be empty) and thus $\hat{X}(z) = \hat{X}(y)$. Thus let $e_P(z) > e_P(y)$. Let $x_y \in \hat{X}(y)$ and $x_z \in \hat{X}(z)$. In view of (26), thus $\lambda(x_y) = e_P(y)x_y + \phi(y) \leq e_P(z)x_y + \phi(z)$ and $\lambda(x_z) = e_P(z)x_z + \phi(z) \leq e_P(y)x_z + \phi(y)$. Add the two inequalities to obtain $0 \leq (x_y - x_z)(e_P(z) - e_P(y))$. Thus $x_z \leq x_y$.

Next, let $\underline{y} < \bar{y}$ and let $y \in (\underline{y}, \bar{y})$. Thus $e_P(y) > 0$ and $\lambda(y) < y$ (Lemma B.5).

[23] To see this, distinguish two cases. First, let $y > \underline{y}$. Hence, $y > \lambda(y)$, and hence the respective suprema in the definitions of $\alpha(y)$ and $\beta(y, e_A(y))$ are attained at some point $x$ above $y$. Using $e_P(y) = \max(\alpha(y), \beta(y, e_A(y))) > 0$ and rearranging, one may verify $\lambda(x) = e_P(y)x + \phi(y)$, meaning $x \in \hat{X}(y)$. Second, let $y = \underline{y}$. Hence, $\lambda(y) = y$, and hence $\phi(y) = (1 - e_P(y))y$. Hence $\lambda(y) = y = e_P(y)y + \phi(y)$, meaning $y \in \hat{X}(y)$.

Thus $\lambda(y) \leq e_P(y)y + (1 - e_P(y))y$ and $\lambda(y) < e_P(y)y + \lambda(y) + u_A(e_A(y)) + e_P(y)\tau$. In particular, $y \notin \hat{X}(y)$.

Finally, let $x \in [\underline{y}, \bar{y}]$. Since $e_P$ is decreasing, claim (2) just proven imply that the image $\hat{X}([\underline{y}, x])$ is contained in $[\min \hat{X}(\underline{y}), \max \hat{X}(x)]$. Clearly, $\min \hat{X}(\underline{y})$ and $\max \hat{X}(x)$ are both in $\hat{X}([\underline{y}, x])$. Claim (1) thus implies that $\hat{X}([\underline{y}, x]) = [\min \hat{X}(\underline{y}), \max \hat{X}(x)]$. Using $\lambda(\underline{y}) = \underline{y}$, one may verify $\lambda(\underline{y}) = e_P(\underline{y})\underline{y} + \phi(\underline{y})$;[24] Thus, $\hat{X}([\underline{y}, x])$ equals the interval $[\underline{y}, \max \hat{X}(x)]$. Finally, it follows from claim (1) that there is $y \in [\underline{y}, x]$ such that $x \in \hat{X}(y)$. □

## C.2 Proof of Lemma C.1

Let $\lambda, \lambda^* \in \Lambda$, and $m \in T(\lambda)$, and $m^* \in T(\lambda^*)$ be such that $(\Pi_m, \lambda) \leq (\Pi_{m^*}, \lambda^*)$. We show $(\Pi_m, \lambda) = (\Pi_{m^*}, \lambda^*)$. It suffices to prove $\lambda = \lambda^*$ since then $m \in T(\lambda)$ and $m^* \in T(\lambda^*)$ imply $\Pi_m = \Pi_{m^*}$.

We next establish an important auxiliary claim under the hypothesis of Lemma C.1.

**Auxiliary claim.** *For all $y \in [\underline{x}, \bar{x}]$, if $\lambda(y) = \lambda^*(y)$, then $(e_A(y), e_P(y)) = (e_A^*(y), e_P^*(y))$.*

*Proof.* The (in)equalities $\lambda(y) = \lambda^*(y)$ and $\lambda \leq \lambda^*$ imply $M(y, \lambda^*) \subseteq M(y, \lambda)$; i.e., to sustain the pointwise higher $\lambda^*$ the principal has fewer choices than when sustaining $\lambda$. Hence, also $\Pi_{m^*} \leq \Pi_m$. Thus $\Pi_{m^*} = \Pi_m$ and $m(y) \in T(y, \lambda^*)$. Since also $m^*(y) \in T(y, \lambda^*)$, Corollary B.1 implies $(e_A(y), e_P(y)) = (e_A^*(y), e_P^*(y))$. □

The idea of the proof is now as follows. Given a type $y$ where we have established $\lambda(y) = \lambda^*(y)$, the auxiliary claim implies $e_P(y) = e_P^*(y)$. This suggests $\lambda(x) = \lambda^*(x)$ for all types $x$ who contemplated deviating to $y$; specifically, we confirm this for types $x$ in $y$'s binding IC correspondence under $(m, \lambda)$. By "repeating" these steps, we eventually deduce $\lambda = \lambda^*$

We distinguish two cases. First, consider the easy case where $m$ has non-random audits. Invoking Theorem B.2, there exists $y_0$ such that $e_P(y) = \mathbf{1}(y < y_0)$ and $\lambda(y) = \min(y, y_0)$ for all $y \in [\underline{x}, \bar{x}]$. Thus immediately $\lambda(y) = \lambda^*(y)$ for all $y \in [\underline{y}, y_0]$. Next, since $\lambda(y_0) = \lambda^*(y_0) = y_0$ and $e_P(y_0) = e_A(y_0) = 0$, the auxiliary claim implies $e_P^*(y_0) = e_A^*(y_0) = 0$. Thus $\lambda^*(y_0) = \sup \lambda^*$. Since $\lambda^*$ is increasing, we conclude that $\lambda(y) = \lambda^*(y) = y_0$ holds for all $y \in [y_0, \bar{x}]$.

[24] Indeed, have we have $\phi(\underline{y}) = \min((1 - e_P(\underline{y}))\underline{y}, \underline{y} + u_A(e_A(\underline{y})) + e_P(\underline{y})\tau) = (1 - e_P(\underline{y}))\underline{y}$, and thus $\lambda(\underline{y}) = \underline{y} = e_P(\underline{y})\underline{y} + \phi(\underline{y})$.

In what follows, let $m$ have random audits. We have to show $\lambda = \lambda^*$.

In the remainder of the proof, we have to take some notational care and be aware of the dependence of certain objects on $(\lambda, m)$ and $(\lambda^*, m^*)$. Specifically, we denote by $\hat{X}$ and $\phi$ the objects defined in Definition C.1 for $(m, \lambda)$, whereas $\hat{X}^*$ and $\phi^*$ denote the counterparts for $(m^*, \lambda^*)$. Likewise, the functions $\alpha, \dots, \underline{\pi}$ are all as in Appendix B.3.2 for $(m, \lambda)$, while $\alpha^*, \dots, \underline{\pi}^*$ are the counterparts for $(m^*, \lambda^*)$.

Define $\bar{z} = \max\{x \in [\underline{x}, \bar{x}] \colon \forall y \in [\underline{x}, x],\ \lambda(y) = \lambda^*(y)\}$, which is well-defined since $\lambda$ and $\lambda^*$ agree at $\underline{x}$ and are continuous. We show $\bar{z} = \bar{x}$, proving $\lambda = \lambda^*$.

Define $\underline{y} = \max\{y \colon \lambda(y) = y\}$, $\bar{y} = \min\{y \colon \lambda(y) = \max \lambda\}$. Since $m$ has random audits, we have $\underline{y} < \bar{y}$ (Theorem B.1).

To begin with, we note $\underline{y} \leq \bar{z}$. Indeed, all $y \in [\underline{x}, \underline{y}]$ satisfy $\lambda(y) = y$, and hence certainly $\lambda(y) = \lambda^*(y)$ since $\lambda \leq \lambda^* \leq \mathrm{id}$. Thus, $\underline{y} \leq \bar{z}$.

**Step 1.** *There is $\varepsilon > 0$ such that $(e_A, e_P)$ and $(e_A^*, e_P^*)$ agree on $[\underline{y}, \underline{y} + \varepsilon]$.*

*Proof.* First, we claim that the efforts $e_A$ and $e_A^*$ are bounded away from 0 on $[\underline{y}, \underline{y} + \varepsilon]$ for $\varepsilon > 0$ sufficiently small. For $e_A$, one easily verifies that $e_A$ is bounded away from 0 on a neighborhood of $\underline{y}$ using Theorem B.1 and the definition of $\underline{y}$. Turning to $e_A^*$, let $Y_0^*$ and SuperLow$^*$ be as in Theorem B.1 applied to $(m^*, \lambda^*)$. As already proven, $\lambda^*(\underline{y}) = \underline{y}$ holds. Thus, $\underline{y} \leq \max\{y \colon \lambda^*(y) = y\}$, and thus Theorem B.1 applied to $(m^*, \lambda^*)$ implies $\underline{y} \in Y_0^* \cup \mathrm{SuperLow}^*$. Using Theorem B.1, one now easily verifies $e_A^*$ is bounded away from 0 on a neighborhood of $\underline{y}$.

Thus, $e_A$ and $e_A^*$ are bounded away from 0 on $[\underline{y}, \underline{y} + \varepsilon]$ for $\varepsilon > 0$ sufficiently small. Since $\lambda(\underline{y}) = \lambda^*(\underline{y}) = \underline{y}$, for $\varepsilon$ sufficiently small and $y \in [\underline{y}, \underline{y} + \varepsilon]$, it holds

$$y - u_A(e_A(y)) < \min(\lambda(y), \lambda^*(y)) \quad \text{and} \quad y - u_A(e_A^*(y)) < \min(\lambda(y), \lambda^*(y)).$$

Fix such $\varepsilon > 0$. We show $e_A$ and $e_A^*$ agree on $[\underline{y}, \underline{y} + \varepsilon]$. Let $y \in [\underline{y}, \underline{y} + \varepsilon]$. Thus $e_P(y) = \alpha(y)$ and $e_P^*(y) = \alpha^*(y)$, and, for all $\tilde{e}_A \in \{e_A(y), e_A^*(y)\}$, we have $\underline{\pi}(y, \tilde{e}_A) = y - u_A(\tilde{e}_A) - c_A(\tilde{e}_A) - (1 - \tilde{e}_A)\alpha(y)$ and $\underline{\pi}^*(y, \tilde{e}_A) = y - u_A(\tilde{e}_A) - c_A(\tilde{e}_A) - (1 - \tilde{e}_A)\alpha^*(y)$. Now recall that $e_A(y)$ maximizes $\underline{\pi}(y, \cdot)$ (yielding profit $\Pi_m(y)$), while $e_A^*(y)$ maximizes $\underline{\pi}^*(y, \cdot)$ (yielding profit $\Pi_{m^*}(y)$). Since $\lambda \leq \lambda^*$, we deduce $\alpha(y) = \alpha^*(y)$; that is, $e_P(y) = e_P^*(y)$. Using also that the map $\tilde{e}_A \mapsto y - u_A(\tilde{e}_A) - c_A(\tilde{e}_A) - (1 - \tilde{e}_A)\alpha(y)$ is strictly quasiconcave (Assumption 3), we also deduce $e_A(y) = e_A^*(y)$. □

**Step 2.** *There is $\varepsilon > 0$ such that $\underline{y} + \varepsilon \leq \bar{z}$.*

*Proof.* Let $\varepsilon > 0$ be as in the previous step. To show $\underline{y} + \varepsilon \leq \bar{z}$, we show $\lambda$ and $\lambda^*$ agree on $[\underline{y}, \underline{y} + \varepsilon]$. Fix $x \in [\underline{y}, \underline{y} + \varepsilon]$. Lemma C.2 implies that there is $y \in [\underline{y}, x]$ such that $x \in \hat{X}(y)$. Hence also $y \in [\underline{y}, \underline{y} + \varepsilon]$, and hence $(e_A(y), e_P(y)) = (e_A^*(y), e_P^*(y))$ by the choice of $\varepsilon$. Therefore, $\lambda(x) = e_P(y)x + \phi(y) = e_P^*(y)x + \phi^*(y) \geq \lambda^*(x)$. Since $\lambda(x) \leq \lambda^*$, also $\lambda(x) = \lambda^*(x)$. □

**Step 3.** *It holds $\bar{y} \leq \bar{z}$.*

*Proof.* Since there is $\varepsilon > 0$ such that $\lambda_m$ and $\lambda^*$ agree on $[\underline{x}, \underline{y} + \varepsilon]$, it suffices to show the following: If $x \in (\underline{y}, \bar{y})$ is such that $\lambda$ and $\lambda^*$ agree on $[\underline{x}, x]$, then there exists $\delta > 0$ such that $\lambda$ and $\lambda^*$ agree on $[\underline{x}, x + \delta]$. Fix $x \in (\underline{y}, \bar{y})$ such that $\lambda$ and $\lambda^*$ agree on $[\underline{x}, x]$. Lemma C.2 implies that the image $\hat{X}([\underline{y}, x])$ equals the interval $[\underline{y}, \max \hat{X}(x)]$. Since $x \in (\underline{y}, \bar{y})$, we have $e_P(x) \in (0, 1)$ and hence Lemma C.2 implies $x < \min \hat{X}(x)$. Now put $\delta = \max(\hat{X}(x)) - x$. Thus, $\delta > 0$ and $\hat{X}([\underline{y}, x]) = [\underline{y}, x + \delta]$. We prove that $\lambda$ and $\lambda^*$ agree on $(x, x + \delta]$. Thus, let $x' \in (x, x + \delta]$. Hence, there is $y \in [\underline{x}, x']$ such that $x' \in \hat{X}(y)$. Since $y \leq x$, we have $\lambda(y) = \lambda^*(y)$ (by the assumption on $x$). The auxiliary claim implies $(e_A(y), e_P(y)) = (e_A^*(y), e_P^*(y))$. Thus, $\lambda(x') = e_P(y)x' + \phi(y) = e_P^*(y)x' + \phi^*(y)$. Since $x' \geq y$, also $e_P^*(y)x' + \phi^*(y) \geq \lambda^*(x')$. In particular, $\lambda(x') \geq \lambda^*(x')$. Since $\lambda \leq \lambda^*$, we conclude $\lambda(x') = \lambda^*(x')$. □

**Step 4.** *It holds $\bar{x} = \bar{z}$.*

*Proof.* Since $\bar{y} \leq \bar{z}$, it suffices to show that $\lambda$ and $\lambda^*$ agree on $[\bar{y}, \bar{x}]$. Theorem B.1 establishes $\lambda(\bar{y}) = \max \lambda$ and $e_P(\bar{y}) = e_A(\bar{y}) = 0$. Since $\bar{y} \leq \bar{z}$, we know $\lambda(\bar{y}) = \lambda^*(\bar{y})$, and hence the auxiliary claim implies $e_P^*(\bar{y}) = e_A^*(\bar{y}) = 0$. Thus also $\lambda^*(\bar{y}) = \max \lambda^*$ (by inspecting the definitions of $\alpha^*$ and $\beta^*$). Summarizing, $\max \lambda = \lambda(\bar{y}) = \lambda^*(\bar{y}) = \max \lambda^*$. Since $\lambda$ and $\lambda^*$ are increasing, we conclude that $\lambda$ and $\lambda^*$ agree on $[\bar{y}, \bar{x}]$. □

Since $\bar{x} = \bar{z}$, we conclude $\lambda = \lambda^*$. □

# D Optimal mechanisms

## D.1 Existence and essential tightness

**Definition D.1** (Essential tightness)**.** A mechanism $m$ is *essentially tight* if there is a tight mechanism $m^*$ such that $F$-almost all types $y$ satisfy

$$(\Pi_m(y), U_m(y), e_A(y), e_P(y)) = (\Pi_{m^*}(y), U_{m^*}(y), e_A^*(y), e_P^*(y)).$$

**Lemma D.1.** *A tight optimal mechanism exists. Every optimal mechanism is essentially tight.*

*Proof of Lemma D.1.* We first show that a tight optimal mechanism exists. Given $\lambda \in \Lambda$ and $y \in [\underline{x}, \bar{x}]$, define $\Pi^*(y, \lambda) = \max_{\tilde{m}(y) \in M(y,\lambda)} \Pi(y, \tilde{m}(y))$, i.e. the profit from applying $T$ at $y$ under $\lambda$. For every mechanism there is a tight mechanism with a pointwise higher profit (Lemma 3.1). In view of Theorem C.1, to prove that a tight optimal mechanism exists, it thus suffices to show that $\lambda \mapsto \int \Pi^*(y, \lambda) \, \mathrm{d}F(y)$ admits a maximizer across $\lambda \in \Lambda$. Since all functions in $\Lambda$ are Lipschitz continuous with constant 1 (Appendix B.3.1) and map to $[\underline{x}, \bar{x}]$, the Arzelà-Ascoli Theorem implies that $\Lambda$ is compact in the supremum norm. For each fixed $y$, the correspondence $\lambda \mapsto E(y, \lambda)$ is upper hemicontiuous and has non-empty compact values (by inspection), and thus $\lambda \mapsto \Pi^*(y, \lambda)$ is upper semicontinuous (Aliprantis and Border, 2006, Lemma 17.30). Fatou's Lemma now implies that $\lambda \mapsto \int \Pi^*(y, \lambda) \, \mathrm{d}F(y)$ is upper semicontinuous; indeed, if $(\lambda_n)_n$ is a sequence in $\Lambda$ converging to a point $\lambda$, then $\limsup_{n\to\infty} \int \Pi^*(y, \lambda_n) \, \mathrm{d}F(y) \leq \int \limsup_{n\to\infty} \Pi^*(y, \lambda_n) \, \mathrm{d}F(y) \leq \int \Pi^*(y, \lambda) \, \mathrm{d}F(y)$, where the first inequality is by Fatou's Lemma and the second inequality is by upper semicontinuity for each fixed $y$. Thus, $\lambda \mapsto \int \Pi^*(y, \lambda) \, \mathrm{d}F(y)$ admits a maximizer across $\lambda \in \Lambda$.

Now let $m$ be optimal. We show $m$ is essentially tight. Let $\lambda_m$ denote $m$'s induced loss function. Using Lemma B.3, find $\lambda \in \Lambda$ such that $(e_A, e_P) \in E(\lambda)$ and $\Pi_m(y) \leq \hat{\Pi}(y, \lambda(y), e_A(y), e_P(y))$ for all $y \in [\underline{x}, \bar{x}]$. Find $m^* \in T(\lambda)$. Theorem C.1 implies $m^*$ is tight. We show that profit and the efforts under $m$ and $m^*$ agree at $F$-almost all types. Lemma B.2 implies $(e_A^*, e_P^*) \in \hat{T}(\lambda)$ and $\hat{\Pi}(y, \lambda(y), e_A(y), e_P(y)) \leq \hat{\Pi}(y, \lambda(y), e_A^*(y), e_P^*(y)) = \Pi_{m^*}(y)$ for all $y$. Since $m$ is optimal, $F$-almost all $y$ satisfy

$$\Pi_m(y) = \hat{\Pi}(y, \lambda(y), e_A(y), e_P(y)) = \hat{\Pi}(y, \lambda(y), e_A^*(y), e_P^*(y)) = \Pi_{m^*}(y).$$

For $y$ such that $\hat{\Pi}(y, \lambda(y), e_A(y), e_P(y)) = \hat{\Pi}(y, \lambda(y), e_A^*(y), e_P^*(y))$, the inclusions $(e_A^*, e_P^*) \in \hat{T}(\lambda)$ and $(e_A, e_P) \in E(\lambda)$ imply $(e_A, e_P) \in \hat{T}(\lambda)$. But Corollary B.1 asserts that $\hat{T}$ is a singleton. Thus, $(e_A, e_P) = (e_A^*, e_P^*)$ for $F$-almost all types. □

## D.2 Proof of Theorem 4.1

**Step 1.** *If $m$ is optimal and tight, then $e_P$ is strictly positive except at $\bar{x}$.*

*Proof of Step 1.* Let $m$ be tight and optimal. Denote $\lambda = \lambda_m$, so that $m \in T(\lambda)$.

Denote $\bar{y} = \sup\{y \in [\underline{x}, \bar{x}] \colon e_P(y) > 0\}$. We show $\bar{y} = \bar{x}$.

Theorems B.1 and B.2 imply $e_P(\bar{x}) = 0$ and that $\lambda$ is constantly $\lambda(\bar{y})$ on $[\bar{y}, \bar{x}]$. Since $m \in T(\lambda)$, Lemma B.2 implies $\Pi(x, m(x)) = \hat{\Pi}(x, \lambda(x), e_A(x), e_P(x))$ and $(e_A(x), e_P(x)) \in E(x, \lambda)$ for all $x$.

We perturb the mechanism as follows. Fix $\eta \in (0, 1)$. Let $y_\eta = \inf\{y \in [\underline{x}, \bar{x}] \colon e_P(y) < \eta\}$. Since $e_P(\bar{x}) = 0$, this infimum is well-defined. Since $e_P$ is right-continuous and decreases (Theorems B.1 and B.2), we have $e_P(y_\eta) \leq \eta$ and $y_\eta \leq \bar{y}$. For all $x \in [\underline{y}, \bar{x}]$, let $\tilde{\lambda}(x) = \bar{\lambda} + \eta(x - \bar{y})$; for all other $x$, let $\tilde{\lambda}(x) = \lambda(x)$. Note $\tilde{\lambda} \leq \mathrm{id}$ since $\lambda \leq \mathrm{id}$ and $\eta < 1$. Further, $\tilde{\lambda} \geq \lambda$ since $\lambda$ is constant on $[\bar{y}, \bar{x}]$.

We next show that, for all $x$ and $y$ such that $y \leq x$, it holds

$$\tilde{\lambda}(x) \leq \tilde{e}_P(y)x + \min\left((1 - \tilde{e}_P(y))y, \tilde{\lambda}(y) + u_A(e_A(y)) + \tilde{e}_P(y)\tau\right). \tag{27}$$

For $x$ below $\bar{y}$, this inequality is immediate since $\tilde{e}_P \geq e_P$ and $\tilde{\lambda}(x) = \lambda(x)$. Thus, let $x \geq \bar{y}$. Note, $\tilde{e}_P(y) = \max(e_P(y), \eta)$ holds for all $y$ since $e_P$ decreases. Hence,

$$\begin{aligned}
&\tilde{\lambda}(x) \\
&= \lambda(\bar{y}) + \eta(x - \bar{y}) \\
&\leq \lambda(\bar{y}) + \tilde{e}_P(y)(x - \bar{y}) \\
&\leq e_P(y)\bar{y} + \min\left((1 - e_P(y))y, \lambda(y) + u_A(e_A(y)) + e_P(y)\tau\right) + \tilde{e}_P(y)(x - \bar{y}) \qquad (28) \\
&\leq \tilde{e}_P(y)x + \min\left((1 - \tilde{e}_P(y))y, \tilde{\lambda}(y) + u_A(e_A(y)) + \tilde{e}_P(y)\tau\right) + \tilde{e}_P(y)(x - \bar{y}) \\
&= \tilde{e}_P(y)x + \min\left((1 - \tilde{e}_P(y))y, \tilde{\lambda}(y) + u_A(e_A(y)) + \tilde{e}_P(y)\tau\right),
\end{aligned}$$

where the inequality (28) follows from $(e_A, e_P) \in E(\lambda)$.

In view of (27), Lemma B.2 implies that there is a mechanism whose profit is at least $\hat{\Pi}(x, \tilde{\lambda}(x), e_A(x), \tilde{e}_P(x))$ for every $x$. Before bounding the expected profit from the perturbation, we note that $x - u_A(e_A(x)) \geq \tilde{\lambda}(x) = \lambda(\bar{y}) + \eta(x - \bar{y})$ holds all $x \in [\bar{y}, \bar{x}]$. Indeed, Theorems B.1 and B.2 imply $e_A(x) = 0$ for all such $x$, and we already noted $\tilde{\lambda}(x) \leq x$ for all $x$.

Collecting our work, we have the following lower bound (the first inequality holds since $m$ is optimal, the second inequality uses the bounds just derived, the third

inequality uses convexity of $c_P$, and the final inequality is by inspection):

$$\begin{aligned}
0 &\geq \int_{[\underline{x},\bar{x}]} \left(\hat{\Pi}(x,\tilde{\lambda}(x),e_A(x),\tilde{e}_P(x)) - \hat{\Pi}(x,\lambda(x),e_A(x),e_P(x))\right) \mathrm{d}F(x) \\
&\geq \int_{[\bar{y},\bar{x}]} \left(\min\left(x-u_A(e_A(x)),\lambda(\bar{y})+\eta(x-\bar{y})\right) - \min\left(x-u_A(e_A(x)),\lambda(x)\right)\right) \mathrm{d}F(x) \\
&\quad - \int_{[y_\eta,\bar{x}]} \left(1-e_A(x)\right)\left(c_P(\eta)-c_P(e_P(x))\right) \mathrm{d}F(x) \\
&\geq \int_{[\bar{y},\bar{x}]} \eta(x-\bar{y})\,\mathrm{d}F(x) - \int_{[y_\eta,\bar{x}]} (1-e_A(x))\eta c_P'(\eta)\,\mathrm{d}F(x) \\
&\geq \int_{[\bar{y},\bar{x}]} \eta(x-\bar{y})\,\mathrm{d}F(x) - \eta c_P'(\eta).
\end{aligned}$$

Now divide by $\eta > 0$ and pass to the limit $\eta \to 0$ to find $0 \geq \int_{[\bar{y},\bar{x}]}(x-\bar{y})\,\mathrm{d}F(x) - c_P'(0)$. By assumption, $c_P'(0) = 0$. Since $\bar{x} = \max(\operatorname{supp} F)$, we conclude $\bar{y} = \bar{x}$. □

**Step 2.** *If $m$ is optimal and tight, then $m$ has random audits.*

*Proof.* This step uses that $F$ is continuous at $\bar{x}$. Let $m$ have non-random audits. We show $m$ is not optimal. According to Theorem 3.2, there is a type $y_0$ such that $m$ is a debt-with-relief mechanism with threshold $y_0$ and relief $\bar{r}_A$ as specified by Theorem 3.2. In view of the previous step, to show that $m$ is not optimal it suffices to show $m$ is not optimal if $y_0 = \bar{x}$. Thus, let $y_0 = \bar{x}$. Define $k = \min_{\tilde{e}_A} \tilde{e}_A c_A'(\tilde{e}_A) + (1-\tilde{e}_A)c_P(1)$. For all $x \in [\underline{x},\bar{x}]$ the profit is given by $\min(x,\bar{x}) - k\mathbf{1}(x < \bar{x})$. Let $\varepsilon > 0$. Consider the debt-with-relief mechanism $m_\varepsilon$ with face value $\bar{x}-\varepsilon$ (and the same relief as $m$); profit of this mechanism is equals $\min(x,\bar{x}-\varepsilon) - k\mathbf{1}(x < \bar{x}-\varepsilon)$ for all $x$. Thus we have (the final line invokes continuity of $F$ at $\bar{x}$):

$$\begin{aligned}
&\int_{[\underline{x},\bar{x}]} (\Pi_m(x) - \Pi_{m_\varepsilon}(x))\,\mathrm{d}F(x) \\
=&\int_{[\underline{x},\bar{x}]} \left(\min(x,\bar{x}) - \min(x,\bar{x}-\varepsilon) - k\left(\mathbf{1}_{(x<\bar{x})} - \mathbf{1}_{(x<\bar{x}-\varepsilon)}\right)\right) \mathrm{d}F(x) \\
&\leq \varepsilon\left(1-F(\bar{x}-\varepsilon)\right) - k\int_{[\bar{x}-\varepsilon,\bar{x})} 1\,\mathrm{d}F(x) \\
=&(\varepsilon-k)\left(1-F(\bar{x}-\varepsilon)\right).
\end{aligned}$$

Since $\bar{x} = \max(\operatorname{supp}(F))$ and $k > 0$, for $\varepsilon > 0$ sufficiently close to 0, the upper bound $(\varepsilon-k)(1-F(\bar{x}-\varepsilon))$ is strictly negative. Thus, $m$ is not optimal. □

**Step 3.** *If $m$ is optimal and tight, then principal-effort $e_P$ is bounded away from 1, and the agent's utility $U_m$ is bounded away from 0.*

*Proof.* By the previous step, $m$ has random audits. In particular, Theorem B.1 applies. Let $\lambda$ denote the induced loss function of $m$. Let $Y_0, \text{SuperLow}, \dots, \text{SuperHigh}, \underline{y}$ be as in Theorem B.1, so that $Y_0 = [\underline{x}, \underline{y})$, and $\underline{y} = \inf \text{SuperLow}$.

Theorem B.1 already asserts that the agent's utility $U_m$ is bounded away from 0. Thus, it remains to show that $e_P$ is bounded away from 1. For later reference, recall that $e_P(y) = 1$ holds if and only if $y \in Y_0$, and $e_P$ is right-continuous at $\max Y_0$. Further, all types $y \in Y_0 \cup \text{SuperLow}$ have a strict incentive to advance the full surplus, i.e. $U_m(y) > y - \lambda_m(y)$; consequently, $r_P(y) = r_\emptyset(y) = 0$ for all such $y$ (else, re-optimize the refunds to obtain a contradiction to the tightness of $m$).

For $\varepsilon > 0$, perturb the mechanism $m$ by decreasing $e_P(y)$ to $e_{P,\varepsilon}(y) = \min(1 - \varepsilon, e_P(y))$ for all $y \in [\underline{x}, \bar{x}]$, and leaving all other parts of the mechanism unchanged. Denote the perturbed mechanism by $m_\varepsilon$. We claim $m_\varepsilon$ is IC for $\varepsilon$ sufficiently small. Specifically, we choose $\varepsilon \in (0, \infty)$ to satisfy two properties:

- Since $U_m$ is bounded away from 0, if $\varepsilon$ is sufficiently close to 0 then all $x \in [\underline{x}, \bar{x}]$ satisfy $x - U_m(x) \leq (1 - \varepsilon)x$.
- Since $e_P$ is right-continuous and decreasing, and equals 1 on $[\underline{x}, \underline{y})$, there is $\delta_\varepsilon \geq 0$ such that $e_{P,\varepsilon}(y) = e_P(y)$ if and only if $y \geq \underline{y} + \delta_\varepsilon$. As $\varepsilon \to 0$, also $\delta_\varepsilon \to 0$ since $e_P$ decreases, and strictly decreases on SuperLow. For $\varepsilon$ sufficiently small, we thus have $\underline{y} + \delta_\varepsilon \in \text{SuperLow}$ since $\underline{y} = \inf \text{SuperLow}$ and SuperLow is an interval with a non-empty interior. We choose $\varepsilon$ so that $\underline{y} + \delta_\varepsilon \in \text{SuperLow}$.

We verify IC for this choice of $\varepsilon$. Since $e_P$ and $e_{P,\varepsilon}$ differ only below $\underline{y} + \delta_\varepsilon$, since $\underline{y} + \delta_\varepsilon \in \text{SuperLow}$, and since $r_P(y) = r_\emptyset(y) = 0$ for all $y \in Y_0 \cup \text{SuperLow}$, it holds $U_m = U_{m_\varepsilon}$, i.e. every type (in $[\underline{x}, \bar{x}]$) has same on-path utility under $m_\varepsilon$ as under $m$. The incentives to deviate to a type outside $[\underline{x}, \underline{y} + \delta_\varepsilon]$ are clearly unaffected. Thus take $[\underline{x}, \underline{y} + \delta_\varepsilon]$. Since $e_{P,\varepsilon}(y) = 1 - \varepsilon$ and $r_\emptyset(y) = 0$, IC demands $x - U_m(x) \leq (e_{P,\varepsilon}(y) - \varepsilon)x = (1 - \varepsilon)x$, which holds by the choice of $\varepsilon$.

Since $m$ is optimal and $m_\varepsilon$ is IC, the interval $[\underline{x}, \underline{y} + \delta_\varepsilon]$ has $F$-measure 0. Since $\underline{x} = \min(\text{supp}(F))$ and since $[\underline{x}, \underline{y} + \delta_\varepsilon]$ is the set where $e_P$ is at least $1 - \varepsilon$, we conclude $e_P$ is bounded away from 1. □

## D.3 Trade-offs across types: Theorem 4.2

### D.3.1 Preparations for the characterization

This section records auxiliary lemmata for the proof of Theorem 4.2. We first derive necessary conditions for optimality for mechanisms whose loss functions are particularly well-behaved. Let $\Lambda^*$ be the set of increasing strictly concave functions $\lambda\colon [\underline{x}, \bar{x}] \to \mathbb{R}$ such that $\lambda(\underline{x}) = \underline{x}$ and $\lambda(x) < x$ for all $x \in (\underline{x}, \bar{x}]$.

From Appendix C.1, given $\lambda \in \Lambda$ and $m \in T(\lambda)$, recall the definition of the binding IC correspondence $\hat{X}$: for all $y$,

$$\hat{X}(y) = \{x \in [y, \bar{x}]\colon \lambda(x) = e_P(y)x + \min\left((1 - e_P(y))y, \lambda(y) + u_A(e_A(y)) + e_P(y)\tau\right)\}.$$

**Lemma D.2.** *Let $\lambda \in \Lambda^*$, $m \in T(\lambda)$, and let $\hat{X}$ be $m$'s binding IC correspondence. Then, on $(\underline{x}, \bar{x})$ the correspondence $\hat{X}$ is singleton-valued, and its unique selection $\hat{x}$ is increasing and continuous.*

*Proof of Lemma D.2.* Let $y \in (\underline{x}, \bar{x})$. The principal's effort $e_P$ is interior on $(\underline{x}, \bar{x})$ (use Theorems B.1 and B.2). Hence, Lemma C.2 that $\hat{X}(y)$ is non-empty. Since $\lambda$ is strictly concave, it is immediate that $\hat{X}(y)$ is a singleton. Let $\hat{x}$ denote the unique selection on $(\underline{x}, \bar{x})$. Since $\hat{X}$ is upper hemicontinuous, it is immediate that $\hat{x}$ is continuous. Since $e_P$ is decreasing and $\hat{X}$ is singleton-valued, Lemma C.2 implies $\hat{x}$ is increasing; indeed, let $z < y$, so that $e_P(y) \leq e_P(z)$; if $e_P(y) = e_P(z)$, then Lemma C.2 implies $\hat{X}(y) = \hat{X}(z)$, whence $\hat{x}(y) = \hat{x}(z)$; if $e_P(y) < e_P(z)$, then Lemma C.2 implies $\max \hat{X}(z) \leq \min \hat{X}(y)$, whence $\hat{x}(z) \leq \hat{x}(y)$. □

We next derive necessary first-order conditions for optimality for mechanisms with loss functions in $\Lambda^*$. Using that loss functions in $\Lambda^*$ induce unique binding ICs (as just shown, Lemma D.2), an Envelope Theorem (Theorem 3 of Milgrom and Segal (2002)) lets us differentiate the principal's profit with respect to perturbations of the loss function, yielding necessary first-order conditions for optimality.

Given a type $x$ and $\lambda \in \Lambda$, recall that $\Pi^*(x, \lambda) = \max_{m(x) \in M(x,\lambda)} \Pi(x, m(x))$ denotes the profit obtained by applying $T$.

**Lemma D.3.** *Let $\lambda \in \Lambda^*$ and $m \in T(\lambda)$. Let $\hat{x}$ denote the unique selection from $m$'s binding IC correspondence, and let $I$ and $D$ be as in (15) for $m$ and $\hat{x}$. Let $\eta\colon [\underline{x}, \bar{x}] \to \mathbb{R}$ be a continuous function that is constantly 0 except possibly on a closed*

*subset of $A$. Denote*

$$v = \int ((1 + D(y))\eta(y) - I(y)\eta(\hat{x}(y)))\,\mathrm{d}F(y). \tag{29}$$

*If $v \neq 0$, then all $\varepsilon \neq 0$ sufficiently close to 0 satisfy $\lambda + \varepsilon\eta \in \Lambda_0$ and*

$$\int (\Pi^*(y, \lambda + \varepsilon\eta) - \Pi^*(y, \lambda))\,\mathrm{d}F(y) \geq |\varepsilon| \cdot |v|.$$

*Proof of Lemma D.3.* Recall from Lemma B.2 that

$$\Pi^*(x, \lambda) = \max_{(\tilde{e}_A, \tilde{e}_P) \in E(x,\lambda)} \hat{\Pi}(x, \lambda(x), \tilde{e}_A, \tilde{e}_P)$$

holds for all $x$. Thus, it suffices to find $(e_{A,\varepsilon}, e_{P,\varepsilon})$ such that, for all $\varepsilon > 0$ sufficiently small, it holds $(e_{A,\varepsilon}, e_{P,\varepsilon}) \in E(\lambda + \varepsilon\eta)$ and $\lambda + \varepsilon\eta \in \Lambda_0$, and

$$\int \hat{\Pi}(\lambda + \varepsilon\eta, e_{A,\varepsilon}, e_{P,\varepsilon})\,\mathrm{d}F - \int \Pi^*(\lambda)\,\mathrm{d}F \geq \varepsilon v.$$

Let $\varepsilon \geq 0$. Denote $\zeta(x) = -\eta(x)/u_A(e_A(x))$ for all $x \in [\underline{x}, \bar{x})$, and $\zeta(\bar{x}) = 0$. We next define a pair $(e_{A,\varepsilon}, e_{P,\varepsilon})$. Let $\mathrm{SuperLow}, \ldots, \mathrm{SuperHigh}$ be as in the conclusion of Theorem B.1 for $(m, \lambda)$. Using that $\lambda$ is strictly increasing and that $\lambda(y) < y$ holds for all $y \in (\underline{x}, \bar{x})$, one may verify via Theorem B.1 that $e_P(y)$ and $e_A(y)$ are interior for all $y \in (\underline{x}, \bar{x})$; given interior $e_P(y)$, Lemma C.2 implies $y < \hat{x}(y)$.

We define $e_{A,\varepsilon}$ as follows:

(1) if $y \in \mathrm{Low}$, let $e_{A,\varepsilon}(y)$ solve $y = \lambda(y) + \varepsilon\eta(y) + u_A(e_{A,\varepsilon}(y))$;
(2) if $y \in \mathrm{High}$, let $e_{A,\varepsilon}(y)$ solve $(1 - \alpha_{\lambda+\varepsilon\eta}(y))y = \lambda(y) + \varepsilon\eta(y) + u_A(e_{A,\varepsilon}(y)) + \alpha_{\lambda+\varepsilon\eta}(y)\tau$;
(3) if $y \in \mathrm{SuperLow} \cup \mathrm{Middle} \cup \mathrm{SuperHigh}$, let $e_{A,\varepsilon}(y) = e_A(y) + \varepsilon\zeta(y)$.

We define $e_{P,\varepsilon}$ for all $y$ by $e_{P,\varepsilon}(y) = \max(\alpha_\varepsilon(y), \beta_\varepsilon(y, e_{A,\varepsilon}(y)))$, where $\alpha_\varepsilon(\bar{x}) = \beta_\varepsilon(\bar{x}, \cdot) = 0$, and for all $y \in [\underline{x}, \bar{x})$,

$$\alpha_\varepsilon(y) = \max\left(0, \sup_{x \in (y, \bar{x}]} \frac{\lambda(x) + \varepsilon\eta(x) - y}{x - y}\right);$$

$$\beta_\varepsilon(y, e_{A,\varepsilon}(y))) = \max\left(0, \sup_{x \in [y, \bar{x}]} \frac{\lambda(x) - \lambda(y) + \varepsilon(\eta(x) - \eta(y)) - u_A(e_{A,\varepsilon}(y))}{x + \tau}\right).$$

It is easy to verify that $(e_{A,\varepsilon}, e_{P,\varepsilon}) \in E(\lambda + \varepsilon\eta)$ holds for $\varepsilon$ sufficiently close to 0. For $\varepsilon = 0$, the pair $(e_{A,\varepsilon}, e_{P,\varepsilon})$ agrees with the efforts $(e_A, e_P)$ from $m$.

To proceed with the proof of Lemma D.3, we calculate the following derivatives.

**Lemma D.4.** *If $y$ is in the interior of one of the interval* SuperLow, ..., SuperHigh, *then*

$$\text{if } y \in \text{SuperLow} \cup \ldots \cup \text{High}, \qquad \left.\frac{\partial e_{P,\varepsilon}(y))}{\partial \varepsilon}\right|_{\varepsilon=0} = \frac{\eta(\hat{x}(y))}{\hat{x}(y) - y}; \tag{30a}$$

$$\text{if } y \in \text{SuperHigh}, \qquad \left.\frac{\partial e_{P,\varepsilon}(y)}{\partial \varepsilon}\right|_{\varepsilon=0} = \frac{\eta(\hat{x}(y)) - \eta(y) - u'_A(e_A(y))\zeta(y)}{\hat{x}(y) + \tau}; \tag{30b}$$

$$\text{if } y \in \text{Low}, \qquad u'_A(e_A(y)) \left.\frac{\partial e_{A,\varepsilon}(y)}{\partial \varepsilon}\right|_{\varepsilon=0} = -\eta(y); \tag{30c}$$

$$\text{if } y \in \text{High}, \qquad u'_A(e_A(y)) \left.\frac{\partial e_{A,\varepsilon}(y)}{\partial \varepsilon}\right|_{\varepsilon=0} = -\eta(y) - \frac{(y+\tau)}{\hat{x}(y) - y}\eta(\hat{x}(y)). \tag{30d}$$

*If $A$ is a closed subset of $(\underline{x}, \bar{x})$, then all derivatives in (30) are bounded across $A$.*

*Proof of Lemma D.4.* Recall from Theorem B.1 that $e_P(y) = \alpha_0(y) > \beta_0(y, e_A(y))$ holds for all $y$ in the interior of SuperLow $\cup$ Low $\cup$ Middle, while $e_P(y) = \beta_0(y, e_A(y)) > \alpha_0(y)$ holds for all $y$ in the interior of SuperHigh. Hence, fixing such $y$, if $\varepsilon$ is sufficiently small, then the same inequalities hold for $e_{P,\varepsilon}, \alpha_\varepsilon, \beta_\varepsilon$. For $y \in$ High, we have $e_{P,\varepsilon}(y) = \alpha_\varepsilon(y) = \beta_\varepsilon(y, e_{A,\varepsilon}(y))$ for all $\varepsilon \geq 0$, by virtue of the choice of $e_{A,\varepsilon}(y)$.

As in the proof of Lemma B.5, for all $y \in (\underline{x}, \bar{x})$ there exists $\delta > 0$ such that

$$\alpha_\varepsilon(y) = \max_{x \in [y+\delta, \bar{x}]} \frac{\lambda(x) + \varepsilon\eta(x) - y}{x - y}. \tag{31}$$

holds for all $\varepsilon$ sufficiently close to 0. Since $\lambda$ is strictly concave, the type $\hat{x}(y)$ is the unique maximizer of (31) at $\varepsilon = 0$. By Theorem 3 of Milgrom and Segal (2002), therefore, we get $\left.\frac{\partial \alpha_\varepsilon(y))}{\partial \varepsilon}\right|_{\varepsilon=0} = -1/(\hat{x}(y) - y)$. Thus also $\left.\frac{\partial e_{P,\varepsilon}(y))}{\partial \varepsilon}\right|_{\varepsilon=0} = -1/(\hat{x}(y) - y)$ for $y$ in SuperLow $\cup$ Low $\cup$ Middle $\cup$ High. Thus (30a) holds.

The equation (30b) follows by a similar application of Theorem 3 of Milgrom and Segal (2002), using again that $\hat{x}(y)$ is the unique maximizer in the limit. The equations (30c) and (30d) follow from implicit differentiation and formulae for the derivative of $e_{P,\varepsilon}$ with respect to $\varepsilon$.

Finally, for the claim regarding boundedness, recall $\hat{x}(y) - y > 0$ for all $y \in (\underline{x}, \bar{x})$. Since $\hat{x}$ is continuous, it holds $\inf_{y \in A}(\hat{x}(y) - y) > 0$. From here, it is easy to see that

all derivatives in (30) are bounded across $A$. $\square$

We now proceed with the proof of Lemma D.3. Let $\Delta(\varepsilon) = \int(\hat{\Pi}(\lambda+\varepsilon\eta, e_{A,\varepsilon}, e_{P,\varepsilon}) - \Pi^*(\lambda))\,\mathrm{d}F$. We calculate:

$$\begin{aligned}\Delta(\varepsilon) = &\int_{[\underline{x},\bar{x}]} (\min(y - u_A(e_{A,\varepsilon}(y)), \lambda(y) + \varepsilon\eta(y)) - \min(y - u_A(e_A(y)), \lambda(y)))\,\mathrm{d}F(y)\\ &+ \int_{[\underline{x},\bar{x}]} -(c_A(e_{A,\varepsilon}(y)) - c_A(e_A(y)))\,\mathrm{d}F(y)\\ &+ \int_{[\underline{x},\bar{x}]} -((1-e_{A,\varepsilon}(y))c_P(e_{P,\varepsilon}(y)) - (1-e_A(y))c_P(e_P(y)))\,\mathrm{d}F(y).\end{aligned}$$

Recall $y - u_A(e_A(y)) < \lambda(y)$ holds for all $y$ in the interior of SuperLow, while $y - u_A(e_A(y)) > \lambda(y)$ holds for all $y$ in the interior of Middle $\cup$ High $\cup$ SuperHigh. Hence, for each $y$, if $\varepsilon$ is sufficiently close to 0, then the same inequalities hold for $e_{A,\varepsilon}$ and $\lambda + \varepsilon\eta$. For $y \in$ Low, we have $y - u_A(e_{A,\varepsilon}(y)) = \lambda(y) + \varepsilon\eta$ for all $\varepsilon$. Divide the formula for the difference $\Delta(\varepsilon)$ by $\varepsilon \neq 0$ and take the limit. Lebesgues' Dominated Convergence Theorem implies that $\Delta$ is differentiable at $\varepsilon = 0$, and the derivative $\Delta'(0)$ is:[25]

$$\begin{aligned}\Delta'(0) = &\int_{y\in\text{SuperLow}} -u_A'(e_A(y))\zeta(y)\,\mathrm{d}F(y) + \int_{y\in\text{Low}\cup\ldots\cup\text{SuperHigh}} \eta(y)\,\mathrm{d}F(y)\\ &+ \int_{y\in[\underline{x},\bar{x}]} -(c_A'(e_A(y)) - c_P(e_P(y)))\left.\frac{\partial e_{A,\varepsilon}(y)}{\partial\varepsilon}\right|_{\varepsilon=0}\mathrm{d}F(y)\\ &+ \int_{y\in[\underline{x},\bar{x}]} -(1-e_A(y))c_P'(e_P(y))\left.\frac{\partial e_{P,\varepsilon}(y)}{\partial\varepsilon}\right|_{\varepsilon=0}\mathrm{d}F(y);\end{aligned} \tag{32}$$

We now plug in the derivatives from Lemma D.4, observing that $F$ is assumed to be absolutely continuous. Then, using the definitions of $I$ and $D$, we can write $\Delta'(0)$ as

$$\begin{aligned}\Delta'(0) = &\int_{\text{SuperLow}} -u_A(e_A'(y))\zeta(y)\,\mathrm{d}F(y) + \int_{\text{Low}\cup\ldots\cup\text{SuperHigh}} \eta(y)\,\mathrm{d}F(y)\\ &+ \int_{\text{SuperLow}\cup\text{Middle}\cup\text{SuperHigh}} -D(y)u_A(e_A(y))\zeta(y)\,\mathrm{d}F(y)\end{aligned}$$

[25] To apply Dominated Convergence, we are using that the integrands in (32) are bounded (across all types). Boundedness follows since: $\eta$ is continuous; $\eta$ is 0 except possibly on a closed subset $A$ of $(\underline{x},\bar{x})$; the derivatives $\frac{\partial e_{P,\varepsilon}(y)}{\partial\varepsilon}$ and $\frac{\partial e_{A,\varepsilon}(y)}{\partial\varepsilon}$ are bounded across $A$ (Lemma D.4); and since $\zeta(y) = 1/u_A(e_A(y))$ is bounded across $y \in A$ as $e_A$ is strictly positive and continuous on $(\underline{x},\bar{x})$.

$$
\begin{aligned}
&+ \int_{\text{Low}} D(y)\eta(y)\,\mathrm{d}F(y) \\
&+ \int_{\text{High}} D(y)\left(\eta(y) + \frac{y+\tau}{\hat{x}(y)-y}\eta(\hat{x}(y))\right)\mathrm{d}F(y) \\
&+ \int_{\text{SuperLow}\cup\text{Middle}\cup\text{SuperHigh}} -I(y)\eta(\hat{x}(y))\,\mathrm{d}F(y) \\
&+ \int_{\text{High}} -\left(I(y) - D(y)\frac{y+\tau}{\hat{x}(y)-y}\right)\eta(\hat{x}(y))\,\mathrm{d}F(y) \\
&+ \int_{\text{SuperHigh}} -I(y)\left(\eta(\hat{x}(y)) - \eta(y) - u_A(e_A(y))\zeta(y)\right)\mathrm{d}F(y).
\end{aligned}
$$

Plugging in for $\zeta$, the expression for $\Delta'(0)$ simplifies to $v$ as defined in (29), i.e. $\Delta'(0) = v = \int ((1+D(y))\eta(y) - I(y)\eta(\hat{x}(y)))\,\mathrm{d}F(y)$. Since $\Delta'(0)$ is the derivative of $\int \hat{\Pi}(\lambda + \varepsilon\eta, e_{A,\varepsilon}, e_{P,\varepsilon})\,\mathrm{d}F - \int \Pi^*(\lambda)\,\mathrm{d}F$ with respect to $\varepsilon$ at 0, we are done. □

**Lemma D.5.** *Let $\lambda \in \Lambda^*$ and $m \in T(\lambda)$. Let $\hat{x}$ be the unique selection from the binding IC correspondence for $(m,\lambda)$. Let* SuperLow*,* Middle*, and* SuperHigh *be as in Theorem B.1. Let $I$ and $D$ be as in (15) for $(m,\hat{x})$. Then*

*(1) all $y \in$ SuperLow satisfy $1 + D(y) = 0$.*

*(2) all $y \in$ Middle satisfy $D(y) = 0$.*

*(3) all $y \in$ SuperHigh satisfy $D(y) = I(y)$.*

*Proof of Lemma D.5.* Let $y \in$ SuperHigh. From Theorem B.1, we know $\pi_3(y,\tilde{e}_A) = \underline{\pi}_3(y,\tilde{e}_A)$ holds for all $\tilde{e}_A$ close to $e_A(y)$. We also recall $e_A(y) \in \operatorname{argmax}_{\tilde{e}_A} \underline{\pi}_3(y,\tilde{e}_A)$. Using that $\hat{x}(y)$ is the unique point in $\hat{X}(y)$, the Envelope Theorem (Milgrom and Segal, 2002, Theorem 3) shows that $\underline{\pi}_3(y,\tilde{e}_A)$ is differentiable with respect to $\tilde{e}_A$ at $e_A(y)$ with derivative $I(y) - D(y)$. Thus $I(y) = D(y)$. For $y \in$ SuperLow, we instead consider the derivative of $\pi_1(y,\tilde{e}_A)$ with respect to $\tilde{e}_A$ at $e_A(y)$; for $y \in$ Middle, consider $\pi_2(y,\cdot)$. □

### D.3.2 Proof of Theorem 4.2

Abbreviate $\lambda = \lambda_m$. We rely on Lemma D.3 and an approximation argument. Find a sequence $(\lambda_n)_{n\in\mathbb{N}}$ in $\Lambda^*$ converging uniformly to $\lambda$.[26]

[26]For example, recalling that $\lambda$ is concave and Lipschitz-continuous with constant 1, find a derivative $\lambda'$ of $\lambda$ that is decreasing, continuous from the right, and satisfies $\lambda' \le 1$. Now, for all $n$ and $x \in [\underline{x}, \bar{x}]$, let $\lambda_n'(x) = -x/n + n\int_{[x,x+1/n]} \lambda' \,\mathrm{d}\,\mathrm{Leb}$ and $\lambda_n(x) = \underline{x} + \int_{[\underline{x},x]} \lambda_n' \,\mathrm{d}\,\mathrm{Leb}$, where Leb denotes Lebesgue measure. Then $(\lambda_n)_{n\in\mathbb{N}}$ is a sequence in $\Lambda^*$ converging uniformly to $\lambda$.

We next pick the candidate increasing selection from $\hat{X}$, as follows. For all $n$, find $m_n \in T(\lambda_n)$. Denote the associated efforts by $(e_{A,n}, e_{P,n})$. According to Lemma D.2, the binding IC correspondence for $(m_n, \lambda_n)$ is singleton-valued and increasing as a function. Let $\hat{x}_n$ denote the increasing selection from the correspondence. By Helly's Selection Theorem and by possibly passing to a subsequence, the sequence $(\hat{x}_n)_{n\in\mathbb{N}}$ converges pointwise. For all $y \in (\underline{x}, \bar{x})$, we put $\hat{x}(y) = \lim_{n\to\infty} \hat{x}_n(y)$. We confirm below that $\hat{x}$ is a selection from the binding IC correspondence of $m$.

Let $I$ and $D$ be as in (15) for $m$ and using the selection $\hat{x}$ just constructed. For each $n$, let $I_n$ and $D_n$ be as in (15) for $m_n$ and the selection $\hat{x}_n$.

For the approximation argument, we use the following technical lemma, proven at the very end of the argument.

**Lemma D.6.** *All of the following hold.*

*(1) If $\eta$ is a continuous function that is constantly 0 except possibly on a closed subset of $(\underline{x}, \bar{x})$, then there exists $\varepsilon > 0$ such that both $\int \tilde{\lambda} \mapsto \Pi^*(y, \tilde{\lambda})\,\mathrm{d}F(y)$ and $\tilde{\lambda} \mapsto \int \Pi^*(y, \tilde{\lambda} + \varepsilon\eta)\,\mathrm{d}F(y)$ are continuous at $\lambda$ on $\Lambda$.*

*(2) $\hat{x}$ is a selection from the binding IC correspondence of $m$.*

*(3) For all $y \in (\underline{x}, \bar{x})$, it holds $(I_n(y), D_n(y)) \to (I(y), D(y))$ as $n \to \infty$.*

*(4) For every closed subset $A$ of $(\underline{x}, \bar{x})$, both $|I|$ and $|D|$ are bounded above across $A$.*

We complete the proof in several steps.

**Step 1.** *Equations (16b) to (16d) all hold.*

*Proof.* We show (16d), the equations (16b) and (16c) being similar. Given $y \in$ SuperHigh, we have to show $I(y) = D(y)$. For all $n$, let SuperHigh$_n$ be as in Theorem B.1 for $(\lambda_n, m_n)$. Lemma D.5 implies that all $z \in$ SuperHigh$_n$ satisfy $I_n(z) = D_n(z)$. Since $I_n \to I$ and $D_n \to D$ pointwise, it suffices to show $y \in$ SuperHigh$_n$ for sufficiently large $n$. Inspecting the definitions of SuperHigh and SuperHigh$_n$ in (25), for large enough $n$ the inclusion $y \in$ SuperHigh$_n$ follows easily since $(e_{A,n}(y), e_{P,n}(y), \lambda_n(y))$ converges to $(e_A(y), e_P(y), \lambda(y))$ as $n \to \infty$. □

Say a closed subinterval $[a, b]$ of $(\underline{x}, \bar{x})$ is *invertible* if $\mathrm{Leb}\{y \in (\underline{x}, \bar{x}) \setminus [a, b] \colon \hat{x}(y) \in [\hat{x}(a), \hat{x}(b)]\} = 0$, where Leb denotes Lebesgue measure.

**Step 2.** *If $[a, b]$ is a closed invertible subinterval of $(\underline{x}, \bar{x})$, then*

$$\int_{[a,b]} I\,\mathrm{d}F = \int_{[\hat{x}(a),\hat{x}(b)]} (1 + D)\,\mathrm{d}F.$$

*Proof.* Towards a contradiction, let $\int_{[a,b]} I \,\mathrm{d}F - \int_{[\hat{x}(a),\hat{x}(b)]}(1+D)\,\mathrm{d}F \neq 0$. Since $[a,b]$ is invertible, this difference equals $\int I(y)\mathbf{1}_{\hat{x}(y)\in[\hat{x}(a),\hat{x}(b)])}\,\mathrm{d}F - \int(1+D(y))\mathbf{1}_{y\in[\hat{x}(a),\hat{x}(b)]}\,\mathrm{d}F$. Hence, there is a continuous function $\eta$ such that $\int I(y)\eta(\hat{x}(y)) - (1+D(y))\eta(y)\,\mathrm{d}F \neq 0$; e.g., take a sequence $(\eta_k)_{k\in\mathbb{N}}$ of continuous functions that converges pointwise to the indicator function for the closed interval $[\hat{x}(a),\hat{x}(b)]$; for $k$ sufficiently large, Lebesgue's Dominated Convergence Theorem implies $\int I(y)\eta_k(\hat{x}(y))\,\mathrm{d}F \neq \int(1+D(y))\eta_k(y)\,\mathrm{d}F$. Since $[a,b]\subset(\underline{x},\bar{x})$, we may choose $\eta$ to be constantly 0 except on a closed subset of $(\underline{x},\bar{x})$.

We next claim, $|\int I_n(y)\eta(\hat{x}_n(y))\,\mathrm{d}F - \int(1+D_n(y))\eta(y)\,\mathrm{d}F \neq 0$ for $n$ sufficiently large. Indeed, recall that $I_n$, $D_n$ and $\hat{x}_n$, respectively, converge pointwise to $I$, $D$, and $\hat{x}$, respectively, and that $I$ and $D$ are bounded on every closed subset of $(\underline{x},\bar{x})$. Thus, the claim follows from an application of Lebesgue's Dominated Convergence Theorem.

Consequently, there is $v>0$ such that $|\int I_n(y)\eta(\hat{x}_n(y))\,\mathrm{d}F - \int(1+D_n(y))\eta(y)\,\mathrm{d}F \geq v > 0$ for all but finitely many $n$. For such $n$, Lemma D.3 implies that all sufficiently small $\varepsilon>0$ satisfy $\int(\Pi^*(y,\lambda_n+\varepsilon\eta)-\Pi^*(y,\lambda_n))\,\mathrm{d}F \geq v\varepsilon$. We choose $\varepsilon>0$ sufficiently small so that, additionally, both $\int\tilde{\lambda}\mapsto\Pi^*(y,\tilde{\lambda})\,\mathrm{d}F(y)$ and $\tilde{\lambda}\mapsto\int\Pi^*(y,\tilde{\lambda}+\varepsilon\eta)\,\mathrm{d}F(y)$ are continuous at $\lambda$ (Lemma D.6). Since $\lambda_n \xrightarrow{n\to\infty} \lambda$, we conclude $\int(\Pi^*(y,\lambda+\varepsilon\eta) - \Pi^*(y,\lambda))\,\mathrm{d}F(y) \geq v\varepsilon$. In particular, $m$ is not optimal; contradiction. □

**Step 3.** *On $(\underline{x},\bar{x})$, the selection is $\hat{x}$ is strictly increasing and continuous. Moreover, every closed subinterval $[a,b]$ of $(\underline{x},\bar{x})$ satisfies $\int_{[a,b]} I\,\mathrm{d}F = \int_{[\hat{x}(a),\hat{x}(b)]}(1+D)\,\mathrm{d}F$.*

*Proof.* Recall that $\hat{x}$ is (weakly) increasing. We show $\hat{x}$ is strictly increasing. Let $x$ be a point in the image of $(\underline{x},\bar{x})$ under $\hat{x}$. Let $\hat{x}^{-1}(x)$ be the pre-image of $x$. Denote $z=\inf\hat{x}^{-1}(x)$ and $y=\sup\hat{x}^{-1}(x)$. Since $\hat{x}$ is increasing, $\hat{x}$ constantly equals $x$ on $(y,z)$. We show $z=y$, proving that $\hat{x}$ is strictly increasing. By construction, $[z,y]$ is invertible. Thus, $\int_{[z,y]} I\,\mathrm{d}F = \int_{[\hat{x}(z),\hat{x}(y)]}(1+D)\,\mathrm{d}F$. But $[\hat{x}(z),\hat{x}(y)]=\{x\}$, and $F$ is absolutely continuous. Thus $\int_{[z,y]} I\,\mathrm{d}F = 0$. By inspection, $I$ is non-zero if $e_P$ is non-zero. In particular, $I$ is non-zero except at $\bar{x}$. Thus, $\int_{[z,y]} I\,\mathrm{d}F = 0$ requires that $F$ assign measure 0 to $[z,y]$. Since $F$ has full support, we conclude $z=y$.

Next, since $\hat{x}$ is strictly increasing, every closed subinterval $[a,b]$ of $(\underline{x},\bar{x})$ of $(\underline{x},\bar{x})$ is invertible, yielding $\int_{[a,b]} I\,\mathrm{d}F = \int_{[\hat{x}(a),\hat{x}(b)]}(1+D)\,\mathrm{d}F$.

It remains to show that $\hat{x}$ is continuous. We use two auxiliary steps.

First, we show $\hat{x}(y)\in(\inf(\mathrm{Low}),\bar{x}]$ holds for all $y\in(\underline{x},\bar{x})$. Towards a contradiction, suppose $\hat{x}(y)\in[\underline{x},\inf(\mathrm{Low}))$ for some $y\in(\underline{x},\bar{x})$. Since $\hat{x}$ is strictly increasing, the

interval $[\hat{x}(\underline{x}), \hat{x}(y)]$ is contained in $[\underline{x}, \inf(\text{Low})]$. Recalling that $1 + D$ is constantly 0 on Low, we find $\int_{\hat{x}([\underline{x},y])}(1 + D)\,\mathrm{d}F = 0$. Thus, also $\int_{[\underline{x},y]} I\,\mathrm{d}F = 0$. Since $e_P$ is strictly positive on $(\underline{x}, \bar{x})$, also $I$ is strictly positive on $(\underline{x}, \bar{x})$, and hence $\underline{x} = y$; contradiction.

Second, we show that for every subinterval $[a, b]$ of $(\underline{x}, \bar{x})$, the function $1 + D$ is strictly positive on $[\hat{x}(a), \hat{x}(b)]$. Let $x \in [\hat{x}(a), \hat{x}(b)]$. By the previous paragraph and since $\hat{x}$ is increasing, we have $x > \inf(\text{Low})$. Inspecting the definition of $D$, it holds $1 + D(x) > 0$ if and only if $u'_A(e_A(x)) + c'_A(e_A(x)) - c_P(e_P(x))$. Using that $u'_A + c'_A$ is strictly increasing (Assumption 3), the inequality $1 + D(x) > 0$ holds if and only $e_A(x)$ is strictly larger than the effort that maximizes $\pi_1(x, \cdot)$, which one may confirm using Theorem B.1 and the inequality $x > \inf(\text{Low})$.

We now show that $\hat{x}$ is continuous at every $y \in (\underline{x}, \bar{x})$. Let $\varepsilon > 0$ be sufficiently small such that $y \pm \varepsilon \in (\underline{x}, \bar{x})$. Thus, $\int_{[y-\varepsilon,y+\varepsilon]} I\,\mathrm{d}F = \int_{[\hat{x}(y-\varepsilon),\hat{x}(y+\varepsilon)]}(1 + D)\,\mathrm{d}F$. For $\varepsilon \to 0$, the integral on the left vanishes. As just shown, $1 + D$ is strictly positive on $[\hat{x}(y - \varepsilon), \hat{x}(y + \varepsilon)]$. Since $F$ has full support, the difference $\hat{x}(y - \varepsilon) - \hat{x}(y + \varepsilon)$ must also vanish as $\varepsilon \to 0$. Since $\hat{x}$ is increasing, we infer that $\hat{x}$ is continuous at $y$. □

It remains to prove Lemma D.6.

*Proof of Lemma D.6.* We first show claim (1). Let $\eta$ be a continuous function that is constantly 0 except possibly on a closed subset of $(\underline{x}, \bar{x})$. Since $m$ is optimal, the principal's effort is bounded away from 1. Since $\lambda$ is $m$'s induced loss function, one may verify using Theorem B.1 that $\lambda(x) < x$ holds for all $x \in (\underline{x}, \bar{x}]$. Since $\eta$ is constantly 0 on a neighborhood of $\underline{x}$, for all $\varepsilon$ sufficiently close to 0 also $\lambda(x) + \varepsilon\eta(x) < x$ for all $x \in (\underline{x}, \bar{x}]$. Fix such a number $\varepsilon$. Let $\lambda^* \in \{\lambda, \lambda + \varepsilon\eta\}$. We show $\tilde{\lambda} \mapsto \Pi^*(y, \tilde{\lambda} + \varepsilon)$ is continuous at $\lambda^*$, which implies the claim. Let $y \in [\underline{x}, \bar{x}]$. Recall that $\Pi^*(y, \lambda^*)$ equals the maximum of $\hat{\Pi}(y, \lambda^*(y), \tilde{e}_A, \tilde{e}_P)$ across $(\tilde{e}_A, \tilde{e}_P)$ subject to the constraint $\lambda^*(x) \leq \tilde{e}_P x + \min((1 - \tilde{e}_P)y, \lambda^*(y) + u_A(\tilde{e}_A) + \tilde{e}_P\tau)$ for all $x \in [y, \bar{x}]$. Let $E(y, \lambda^*)$ denote the set of pairs $(\tilde{e}_A, \tilde{e}_P)$ satisfying this constraint. This constraint always holds if $x = y$ since $\lambda^*(y) \leq y$. Since $\lambda^*(x) < x$ for all $x \in (\underline{x}, \bar{x}]$, setting $\tilde{e}_P = 1$ implies that the constraint holds strictly for all $x \in (y, \bar{x}]$. With this observation, the constraint correspondence $\tilde{\lambda} \mapsto E(y, \tilde{\lambda})$ is lower hemicontinuous at $\lambda^*$. A routine argument shows the constraint correspondence is upper hemicontinuous. Thus, by Berge's Maximum Theorem, $\tilde{\lambda} \mapsto \Pi^*(y, \tilde{\lambda})$ is continuous at $\lambda^*$. Since $y$ was arbitrary, Lebesgue's Dominated Convergence Theorem implies that $\tilde{\lambda} \mapsto \int \Pi^*(y, \tilde{\lambda} + \varepsilon)\,\mathrm{d}F(y)$ is continuous at $\lambda^*$.

In an auxiliary step, we show that for all $y \in (\underline{x}, \bar{x})$ the sequence of efforts $(e_{A,n}(y), e_{P,n}(y))_{n\in\mathbb{N}}$ converges to $(e_A(y), e_P(y))$. Fix $y$. Take an arbitrary subsequence of efforts, and pass to a subsequence along which both efforts are convergent. Denote the limit by $(\tilde{e}_A(y), \tilde{e}_P(y))$. As in the previous paragraph, the correspondence $E(y, \cdot)$ is continuous at $\lambda$. Thus, Berge's Maximum Theorem implies that $\hat{T}(y, \cdot)$ is upper hemicontinuous at $\lambda$. In particular, since $(\lambda_n)_{n\in\mathbb{N}}$ converges to $\lambda$ and since $(e_{A,n}(y), e_{P,n}(y)) \in \hat{T}(y, \lambda_n)$ holds for all $n$, the limit $(\tilde{e}_A(y), \tilde{e}_P(y))$ is in $\hat{T}(y, \lambda)$. Since $\hat{T}(y, \lambda)$ is a singleton (Corollary B.1), we conclude $(\tilde{e}_A(y), \tilde{e}_P(y)) = (e_A(y), e_P(y))$. Thus, every subsequence of efforts admits a further subsequence converging to $(e_A(y), e_P(y))$. Thus, the entire sequence converges to the same limit, as desired.

Using that $(e_{A,n}, e_{P,n})_{n\in\mathbb{N}}$ converges to $(e_A, e_P)$ on $(\underline{x}, \bar{x})$, that $(\lambda_n)_{n\in\mathbb{N}}$ converges to $\lambda$, it is easy to see that $\hat{x}$ is a selection from the binding IC correspondence, and that $(I_n)_{n\in\mathbb{N}}$ and $(D_n)_{n\in\mathbb{N}}$, resp., converge pointwise to $I$ and $D$, resp., on $(\underline{x}, \bar{x})$.

Finally, we show that $|I|$ and $|D|$ are bounded above on every closed subset $A$ of $(\underline{x}, \bar{x})$. Indeed, Lemma C.2 implies $y < \hat{x}(y)$ for all $y \in (\underline{x}, \bar{x})$. Since the binding IC correspondence (of $m$) is upper hemicontinuous, in fact $\inf_{y\in A}(\hat{x}(y) - y) > 0$. We also know $\inf_{y\in A} e_A(y) > 0$ since $e_A$ is continuous and strictly positive on $(\underline{x}, \bar{x})$. From here, it is easy to see that $|I|$ and $|D|$ are bounded above on $A$. □

With Lemma D.6 established, the proof is complete. □

# E Miscellaneous

## E.1 Example of non-strictly decreasing principal effort

Figure 5 shows the efforts of a tight mechanism with random audits. The principal's effort is decreasing (as per Theorem 3.1), but is not strictly decreasing on SuperHigh. The environment is as in the example from Figure 1: the type space is $[0, 1]$, the costs are $c_A(e) = e^4$ and $c_P(e) = 2e^2$ for all $e \in [0, 1]$, and the principal's private funds are $\tau = 0.2$. We obtained the efforts from Figure 5 by applying step (4) of the tightening algorithm to $\tilde{\lambda}$ given by $\tilde{\lambda}(x) = \min\left(0.2 + 0.1x, \sqrt{1+x} - 1\right)$ for all $x \in [0, 1]$. In this mechanism, the principal's effort is constant on an interval around type $x = 0.3$. In our numerical solution, the principal's effort $e_P(x)$ is within $10^{-6}$ of 0.1 at all types $x \in [0.25, 0.35]$ on a grid of types with spacing $10^{-3}$. It is no accident that the constant value of 0.1 also equals the slope of $\tilde{\lambda}$ when it is affine. The fact that $e_P$ is constant

on an interval of types relies on the fact that all types in this interval have non-unique binding IC types.

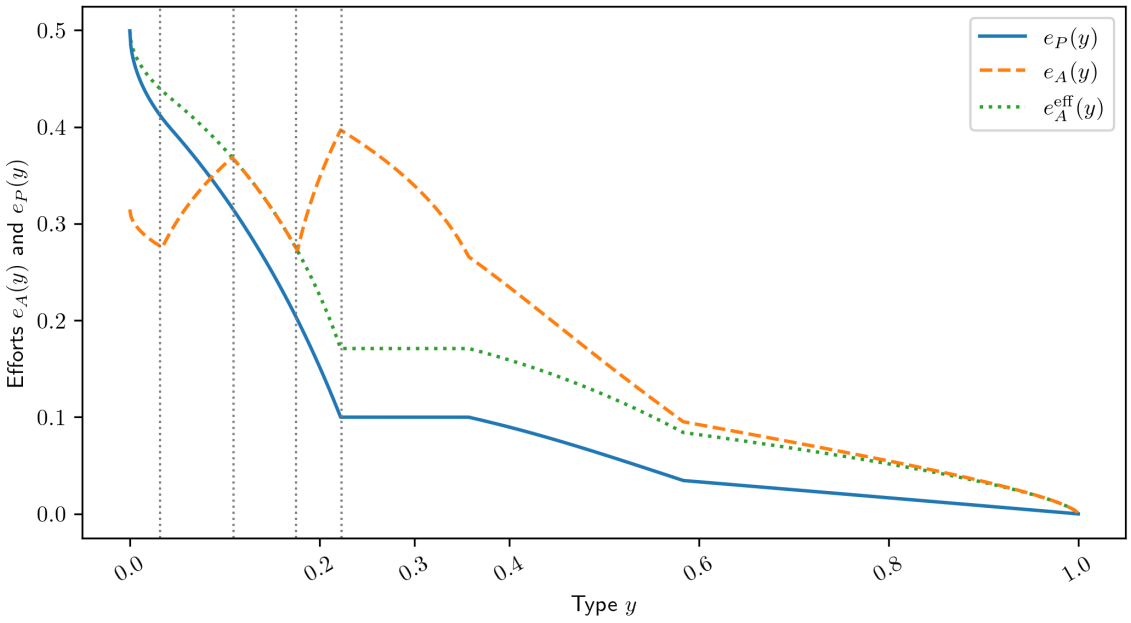


**Figure 5.** Effort levels in a tight mechanism with random audits where the principal's effort is not strictly decreasing.

## E.2 Undominated mechanisms

**Definition E.1.** A mechanism $m$ is *undominated* if for all mechanisms $m^*$ such that $\Pi_m \leq \Pi_{m^*}$ it holds $\Pi_m = \Pi_{m^*}$.

By applying Lemma 3.1, it is immediate that for every undominated mechanism there is a tight mechanism with a type-by-type equal profit. We next show that undominated mechanisms are in fact tight themselves.

**Undominated mechanisms are tight.** Let $m$ be undominated. Let $\lambda_m$ denote $m$'s induced loss function. Find $m^* \in T(\lambda_m)$. Lemma 3.1 implies $(\Pi_m, \lambda_m) \leq (\Pi_{m^*}, \lambda_{m^*})$. Since $m$ is undominated, also $\Pi_m = \Pi_{m^*}$. Thus, also $m \in T(\lambda_m)$. In view of Theorem C.1, it follows that $m$ is tight if we can show $\lambda_m \in \Lambda$, i.e. $\lambda$ is increasing, concave, and $\lambda_m(\underline{x}) = \underline{x}$ and $\lambda_m \leq \mathrm{id}$. We can can show $\lambda_m \in \Lambda$ by repeating the steps from Appendix B.1 that show that tight mechanisms have an induced loss function in $\Lambda$ (Proposition B.2); with some work one can show that these steps strictly increase

the profit for at least one type if $\lambda_m$ does not coincide with $\lambda^+$ or $\tilde{\lambda}$ as constructed in the tightening algorithm. Since $m$ is undominated, also $\lambda_m = \tilde{\lambda}$, and thus $\lambda_m \in \Lambda$.

**There is a tight mechanism that is not undominated.** Intuitively, a tight mechanism $m$ may entail the principal to audit the lowest type $\underline{x}$ with certainty in order to sustain a high postulated induced loss function. However, no type of the agent may lose the full surplus by being truthful, and thus auditing $\underline{x}$ with certainty is actually excessive for providing incentives. Thus, by slightly decreasing $e_P(\underline{x})$, the principal strictly raises profits without disturbing incentives. Thus, $m$ is not undominated. Note, however, that decreasing $e_P(\underline{x})$ may also decrease the induced loss function, and so the decrease does not contradict the tightness of $m$.

For an example, let $\lambda$ be an arbitrary function in $\Lambda$ such that for some type $\underline{y} \in (\underline{x}, \bar{x}]$, it holds $\lambda(x) = x$ for all types $x \in [\underline{x}, \underline{y}]$, such that $\lambda$ is strictly increasing on $[\underline{y}, \bar{x}]$, and such that $\lambda(\bar{x}) < \bar{x}$. Find $m \in T(\lambda)$. Thus, $m$ is tight (Theorem C.1). Using Theorems B.1 and B.2, one can show $e_P(y) = 1$ for all $y \in [\underline{x}, \underline{y})$, and that all types in $[\underline{x}, \bar{x}]$ have a strictly positive utility from advancing the full surplus. By slightly decreasing $e_P(\underline{x})$ and changing nothing else, the principal obtains another mechanism whose profits are type-by-type higher, strictly so at $\underline{x}$. (The proof of Theorem 4.1 has the same argument.) Thus, $m$ is not undominated.

# F References


Ahmadzadeh, Amirreza (2024): "Costly state verification with Limited Commitment," .

Ahmadzadeh, Amirreza and Stephan Waizmann (2024): "Mechanism Design with Costly Inspection," .

Aliprantis, Charalambos D. and Kim C. Border (2006): *Infinite Dimensional Analysis : A Hitchhikers Guide*, Springer Berlin Heidelberg.

Allingham, Michael G and Agnar Sandmo (1972): "Income tax evasion: A theoretical analysis," *Journal of public economics*, 1 (3-4), 323–338.

Andreoni, James, Brian Erard, and Jonathan Feinstein (1998): "Tax compliance," *Journal of economic literature*, 36 (2), 818–860.

Asseyer, Andreas and Ran Weksler (2024): "Certification Design with Common Values," *Econometrica*, 92 (3), 651–686.

Ball, Ian and Deniz Kattwinkel (2025): "Probabilistic Verification in Mechanism Design," .

Ball, Ian and Jan Knoepfle (2024): "Should the Timing of Inspections be Predictable?" .

Ball, Ian and Teemu Pekkarinen (2024): "Optimal Auction Design with Contingent Payments and Costly Verification," .

Ben-Porath, Elchanan, Eddie Dekel, and Barton L Lipman (2014): "Optimal allocation with costly verification," *American Economic Review*, 104 (12), 3779–3813.

Ben-Porath, Elchanan, Eddie Dekel, and Barton L. Lipman (2019): "Mechanisms with evidence: Commitment and robustness," *Econometrica*, 87 (2), 529–566.

——— (2023): "Sequential mechanisms for evidence acquisition," .

Beshkar, Mostafa and Eric W Bond (2017): "Cap and escape in trade agreements," *American Economic Journal: Microeconomics*, 9 (4), 171–202.

Bester, Helmut, Matthias Lang, and Jianpei Li (2021): "Signaling versus auditing," *The RAND Journal of Economics*, 52 (4), 859–883.

Border, Kim C and Joel Sobel (1987): "Samurai accountant: A theory of auditing and plunder," *Review of Economic Studies*, 54 (4), 525–540.

Boyd, John H and Bruce D Smith (1994): "How good are standard debt contracts? Stochastic versus nonstochastic monitoring in a costly state verification environment," *Journal of Business*, 539–561.

Brzustowski, Thomas and Albin Erlanson (2024): "Optimal Allowance with Limited Auditing Capacity," .

Bull, Jesse (2008a): "Costly evidence production and the limits of verifiability," *The BE Journal of Theoretical Economics*, 8 (1).

——— (2008b): "Mechanism design with moderate evidence cost," *The BE Journal of Theoretical Economics*, 8 (1).

Castro-Pires, Henrique, Hector Chade, and Jeroen Swinkels (2024): "Disentangling moral hazard and adverse selection," *American economic review*, 114 (1), 1–37.

Celik, Gorkem (2006): "Mechanism design with weaker incentive compatibility constraints," *Games and Economic Behavior*, 56 (1), 37–44.

Chander, Parkash and Louis L Wilde (1998): "A general characterization of optimal income tax enforcement," *Review of Economic Studies*, 65 (1), 165–183.

Chen, Yi-Chun, Gaoji Hu, and Xiangqian Yang (2022): "Information Design in Allocation with Costly Verification," .

de Clippel, Geoffroy (2008): "An axiomatization of the inner core using appropriate reduced games," *Journal of Mathematical Economics*, 44 (3-4), 316–323.

Dye, Ronald A (1985): "Disclosure of nonproprietary information," *Journal of accounting research*, 123–145.

Epitropou, Markos and Rakesh Vohra (2019): "Optimal on-line allocation rules with verification," in *Algorithmic Game Theory: 12th International Symposium, SAGT 2019, Athens, Greece, September 30–October 3, 2019, Proceedings 12*, Springer, 3–17.

Erlanson, Albin and Andreas Kleiner (2019): "A note on optimal allocation with costly verification," *Journal of Mathematical Economics*, 84, 56–62.

——— (2020): "Costly verification in collective decisions," *Theoretical Economics*, 15 (3), 923–954.

——— (2024): "Optimal allocations with capacity constrained verification," .

Gale, Douglas and Martin Hellwig (1985): "Incentive-compatible debt contracts: The one-period problem," *The Review of Economic Studies*, 52 (4), 647–663.

Halac, Marina and Pierre Yared (2020): "Commitment versus flexibility with costly verification," *Journal of Political Economy*, 128 (12), 4523–4573.

Holmström, Bengt (1979): "Moral hazard and observability," *The Bell journal of economics*, 74–91.

Hu, Gaoji (2024): "Screening by (In) accurate Inspection," *Available at SSRN 4797356*.

Jiang, Shaofei (2024): "Persuasion via sequentially acquired evidence," .

Jovanovic, Boyan (1982): "Truthful disclosure of information," *The Bell Journal of Economics*, 36–44.

Kaplow, Louis (2011a): "On the optimal burden of proof," *Journal of Political Economy*, 119 (6), 1104–1140.

——— (2011b): "Optimal proof burdens, deterrence, and the chilling of desirable behavior," *American Economic Review*, 101 (3), 277–280.

Kartik, Navin and Olivier Tercieux (2012): "Implementation with evidence," *Theoretical Economics*, 7 (2), 323–355.

Kattwinkel, Deniz and Jan Knoepfle (2023): "Costless information and costly verification: A case for transparency," *Journal of Political Economy*, 131 (2), 504–548.

Khalfan, Nawaaz (2023): "Optimal Allocation with Noisy Inspection," .

Khalfan, Nawaaz and Rakesh Vohra (2024): "Sequential Information Acquisition and Optimal Search," .

Li, Yunan (2020): "Mechanism design with costly verification and limited punishments," *Journal of Economic Theory*, 186, 105000.

——— (2021): "Mechanism design with financially constrained agents and costly verification," *Theoretical Economics*, 16 (3), 1139–1194.

Li, Zihao and Jonathan Libgober (2023): "The Dynamics of Verification when Searching for Quality," .

Lichtig, Avi and Helene Mass (2024): "Optimal Testing in Disclosure Games," Tech. rep., University of Bonn and University of Mannheim, Germany.

Luttmer, Erzo FP and Monica Singhal (2014): "Tax morale," *Journal of economic perspectives*, 28 (4), 149–168.

Madarasz, Kristof and Marek Pycia (2023): "Information Choice: Cost over Content," .

Malenko, Andrey (2019): "Optimal dynamic capital budgeting," *Review of Economic Studies*, 86 (4), 1747–1778.

Milgrom, Paul and Ilya Segal (2002): "Envelope theorems for arbitrary choice sets," *Econometrica*, 70 (2), 583–601.

Monnet, Cyril and Erwan Quintin (2005): "Optimal contracts in a dynamic costly state verification model," *Economic Theory*, 26, 867–885.

Mookherjee, Dilip and Ivan Png (1989): "Optimal auditing, insurance, and redistribution," *The Quarterly Journal of Economics*, 104 (2), 399–415.

Myerson, Roger B (1982): "Optimal coordination mechanisms in generalized principal–agent problems," *Journal of mathematical economics*, 10 (1), 67–81.

Palonen, Petteri and Teemu Pekkarinen (2022): "Mechanism Design with Auditing and Avoidance: Tax Evasion Story," Available at SSRN 4230382.

Patel, Rohit and Can Urgun (2022): "Costly verification and money burning," .

Perez-Richet, Eduardo and Vasiliki Skreta (2024): "Score-based mechanisms," .

Pham, Hien (2024): "Adverse Selection with Costly Verification," *Available at SSRN 4997973*.

Popov, Latchezar (2016): "Stochastic costly state verification and dynamic contracts," *Journal of Economic Dynamics and Control*, 64, 1–22.

Pram, Kym (2023): "Learning and Evidence in Insurance Markets," *International Economic Review*, 64 (4), 1685–1714.

Preusser, Justus (2022): "Costly evidence and the value of commitment," .

Ravikumar, B and Yuzhe Zhang (2012): "Optimal auditing and insurance in a dynamic model of tax compliance," *Theoretical Economics*, 7 (2), 241–282.

Siegel, Ron and Bruno Strulovici (2023): "Judicial mechanism design," *American Economic Journal: Microeconomics*, 15 (3), 243–270.

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

Stahl, Konrad and Roland Strausz (2017): "Certification and market transparency," *Review of Economic Studies*, 84 (4), 1842–1868.

Strausz, Roland and Daniel Krähmer (2024): "Unidirectional Incentive Compatibility," .

Tirole, Jean (2010): *The theory of corporate finance*, Princeton university press.

Townsend, Robert M (1979): "Optimal contracts and competitive markets with costly state verification," *Journal of Economic theory*, 21 (2), 265–293.

Verrecchia, Robert E (1983): "Discretionary disclosure," *Journal of accounting and economics*, 5, 179–194.

Wang, Cheng (2005): "Dynamic costly state verification," *Economic Theory*, 25, 887–916.

Whitmeyer, Mark and Kun Zhang (2022): "Costly Evidence and Discretionary Disclosure," .