Herweg, Fabian; Müller, Daniel

Working Paper

The Optimality of Simple Contracts: Moral Hazard and Loss Aversion

Bonn Econ Discussion Papers, No. 17/2008

Provided in Cooperation with:
Bonn Graduate School of Economics (BGSE), University of Bonn

Suggested Citation: Herweg, Fabian; Müller, Daniel (2008) : The Optimality of Simple Contracts: Moral Hazard and Loss Aversion, Bonn Econ Discussion Papers, No. 17/2008, University of Bonn, Bonn Graduate School of Economics (BGSE), Bonn

This Version is available at:
https://hdl.handle.net/10419/27177

Bonn Econ Discussion Papers

Discussion Paper 17/2008

The Optimality of Simple Contracts: Moral Hazard and Loss Aversion

by

Fabian Herweg, Daniel Müller, Philipp Weinschenk

September 2008

Bonn Graduate School of Economics
Department of Economics
University of Bonn
Adenauerallee 24 - 42
D-53113 Bonn

The Bonn Graduate School of Economics is
sponsored by the

Deutsche Post World Net
MAIL EXPRESS LOGISTICS FINANCE

The Optimality of Simple Contracts: Moral Hazard and Loss Aversion*

Fabian Herweg, Daniel Müller,
and Philipp Weinschenk†

October 8, 2008

This paper extends the standard principal-agent model with moral hazard to allow for agents having reference-dependent preferences according to Kőszegi and Rabin (2006, 2007). The main finding is that loss aversion leads to fairly simple contracts. In particular, when shifting the focus from standard risk aversion to loss aversion, the optimal contract is a simple bonus contract, i.e. when the agent's performance exceeds a certain threshold he receives a fixed bonus payment. Moreover, if the agent is sufficiently loss averse, it is shown that the first-order approach is not necessarily valid. If this is the case the principal may be unable to fine-tune incentives. Strategic ignorance of information by the principal, however, allows to overcome these problems and may even reduce the cost of implementation.

JEL classification: D8; M1; M5

Keywords: Agency Model; Moral Hazard; Reference-Dependent Preferences; Loss Aversion

1 Introduction

"The recent literature provides very strong evidence that contractual forms have large effects on behavior. As the notion that "incentive matters" is one of the central tenets of economists of every persuasion, this should be comforting to the community. On the other hand, it raises an old puzzle: if contractual form matters so much, why do we observe such a prevalence of fairly simple contracts?"

- Bernard Salanié

*In preparing this paper we have greatly benefited from comments made by Patrick Bolton, Jörg Budde, Paul Heidhues, Martin Hellwig, Botond Kőszegi, Patrick Schmitz, and Urs Schweizer. We also thank seminar participants at University of Bonn, as well as participants at the IMEBE at Alicante (2008), EEA/ESEM at Milan (2008), and at the annual congress of the Verein für Socialpolitik at Graz (2008). The usual disclaimer applies.

†Fabian Herweg, University of Bonn; Daniel Müller, University of Bonn; Philipp Weinschenk, University of Bonn and Max Planck Institute for Research on Collective Goods Bonn, Corresponding author. E-mail address: fherweg@uni-bonn.de (F. Herweg).

The question asked by Salanié (2003), why observed contracts often display far less complexity than predicted by economic theory, neither is new nor is the answer fully understood. While Prendergast (1999) already referred to the discrepancy between theoretically predicted and actually observed contractual form about a decade ago, over time this question was raised again and again, for example recently by Lazear and Oyer (2007). The most simple incentive contract one can think of is a bonus contract, with a bonus being a payment made for achieving some level of performance. And indeed, according to Joseph and Kalwani (1998), bonuses are a form of incentive pay widely used by a large variety of organizations, in particular within sales organizations. As Oyer (1998), however, points out, facing an annual sales quota provides incentives for salespeople to manipulate prices and timing of business to maximize their own income rather than their firms' profits. This observation raises "the interesting question of why these nonlinear contracts are so prevalent. [...] It appears that there must be some benefit of these contracts that outweighs these apparent costs" (Lazear and Oyer (2007)). Simple contracts are not only common in labor contexts but also in insurance markets. A prevalent form of insurance contracts is a straight-deductible contract widely used, for example, in automobile insurance.[1] As Dionne and Gagné (2001) point out, however, "deductible contracts can introduce perverse effects when falsification behavior is potentially present". With fraudulent claims being a major problem in the car insurance market,[2] which is – at least partially – due to straight deductible contracts, the prevalence of this particular contractual form seems puzzeling.[3]

To give one possible explanation for the widespread use of the contractual arrangements just described, we consider a principal-agent model with moral hazard, framed as an employer-employee relationship, which is completely standard but for one twist: the agent is assumed to have reference-dependent preferences according to Kőszegi and Rabin (2006, 2007), and in consequence is loss averse. In expectations the agent suffers from deviations from his reference point. By offering a simple contract which specifies only few different wage payments, the principal can reduce the scope for the agent to experience a loss, thereby lowering the payment necessary to compensate the agent for ex ante expected losses. In the extreme case of a purely loss averse agent, this logic leads to a literal bonus contract being optimal. Put differently, no matter how rich the performance measure, the principal offers only two different wages, a high wage for "good" performance, and a low wage for "bad" performance.

[1] For evidence on deductibles in the automobile insurance see Puelz and Snow (1994) or Chiappori et al. (2006).

[2] Caron and Dionne (1997) estimated the cost of fraud in the Québec automobile insurance market in 1994 at $100 million, just under 10% of total claims. For an estimation of the costs of fraudulent claims in the United States, see Foppert (1994).

[3] As was shown by Rothschild and Stiglitz (1976), the use of deductibles can theoretically be explained if the insurance market is subject to adverse selection. Besides adverse selection, however, moral hazard plays an important role in automobile insurance. Deductibles were found to be optimal under moral hazard by Holmström (1979) if the insured person's action influences only the probability of an accident but not its severity. As pointed out by Winter (2000), however, "[d]riving a car more slowly and carefully reduces *both* the probability of an accident and the likely costs of an accident should it occur." Thus, existing theories cannot explain the prevalence of deductibles in these markets.

We present our model of a principal-agency which is subject to moral hazard in Section 2. The principal, who is both risk and loss neutral, does not observe the agent's effort directly. Instead, he observes a measure of performance which is correlated – though imperfectly – with the agent's effort decision. Our model departs from the classical principal-agent relationship by assuming that the agent has reference-dependent preferences in the sense of Kőszegi and Rabin (2006, 2007). This recent concept of reference-dependent preferences posits that a decision maker – next to intrinsic consumption utility from an outcome – also derives gain-loss utility from comparing the actual outcome with his rational expectations about outcomes. More precisely, the sensation of gains and losses is derived by comparing a given outcome to all possible outcomes. To illustrate this point consider an employee who receives a wage of \$5000 for good performance, a wage of \$4400 for mediocre performance, and a wage of \$4000 for bad performance. If the employee's performance is bad he experiences the sensation of a loss of \$400 and of a loss of \$1000, with the weights on the two losses equal to the probability with which he expected to perform fairly or well, respectively. If the employee's performance is mediocre, this generates mixed feelings, a loss of \$600 and a gain of \$400.[4] The key feature of the Kőszegi-Rabin model is that expectations matter in determining the reference point.[5] This assumption is based mainly on findings in the psychological literature. For instance, Mellers et al. (1999) and Breiter et al. (2001) document that both the actual outcome and unattained possible outcomes affect subjects' satisfaction with their payoff. Just very recently two remarkable contributions to the economic literature also provided evidence that expectations play an important role in the determination of the reference point. In a real-effort experiment, Abeler et al. (2008) find strong evidence for individuals taking their expectations as a reference point, rather than the status quo, as was most often assumed in the wake of Kahneman and Tversky's original formulation of prospect theory (1979). Post et al. (2008), on the other hand, analyze decision making in a large-stake game show and come to the conclusion, that observed behavior "is consistent with the idea that the reference point is based on expectations." The Kőszegi-Rabin concept is successfully applied by Heidhues and Kőszegi (2008) to provide a theoretical explanation for an old puzzle from the industrial organization literature known as focal pricing: by introducing consumer loss aversion into a standard model of price competition with differentiated products, they give an answer to the question why non-identical competitors charge identical (focal) prices for differentiated products.

As a benchmark, in Section 3 we first consider the case of a purely risk averse agent. This visit to Holmström (1979)'s world yields a familiar result: Under the optimal contract signals that are more indicative of higher effort are rewarded strictly higher, thereby giving rise to a strictly increasing wage profile. We then turn to the analysis of a purely loss averse agent, who does not exhibit risk aversion in the usual sense. After providing

[4]For at least suggestive evidence on mixed feelings, see Larsen et al. (2004).

[5]The feature that the reference point is determined by the decision maker's forward-looking expectations is shared with the disappointment-aversion models of Bell (1985), Loomes and Sugden (1986), and Gul (1991).

sufficient conditions for the first-order approach to be valid, we establish our main result: when the agent is loss averse, the principal considers it optimal to offer a bonus contract which comprises of only two different wage payments. No matter how rich the set of possible realizations of the performance measure, the optimal contract entails a minimum of wage differentiation in the sense that the set of all possible signals is partitioned into only two subsets: signals contained in the one subset are rewarded with a strictly higher wage than signals in the complementary subset. We already briefly touched the intuition underlying this finding. With the agent's action being unobservable, the necessity to create incentives makes it impossible for the principal to bear the complete risk. With losses looming larger than equally sized gains, this ex ante imposes an expected net loss on the agent. This overall expected net loss equals the sum over the ex ante expected wage differences weighted with the product of the corresponding probabilities. To illustrate, let us return to the example introduced above. Suppose the agent expects to perform well, moderately, or poorly with probability p_G, p_M and p_B, respectively. Then, ex ante, the agent expects a wage difference – or net loss – of \$600 with probability $p_M p_G$, a net loss of \$400 with probability $p_B p_M$, and a net loss of \$1000 with probability $p_B p_G$. The agent demands to be compensated for his overall expected net loss, which the principal therefore seeks to minimize. Consider, for a sake of argumentation, a principal who has to improve incentives. There are two ways to do so. First, the principal can introduce a new wage spread, i.e., pay slightly different wages for two signals that were rewarded equally in the original wage scheme, while keeping the differences between all other neighboring wages constant. Secondly, the principal can increase an existing wage spread, holding constant all other spreads between neighboring wages. Both procedures increase the overall expected net loss by increasing the size of some of the expected losses without reducing others. Introducing a new wage spread, however, additionally increases the overall expected net loss by increasing the ex-ante expected probability of experiencing a loss. Therefore, in order to improve incentives it is advantageous to increase a particular existing wage spread without adding to the contractual complexity in the sense of increasing the number of different wages. In this sense, reference-dependent preferences according to Kőszegi and Rabin introduce a notion of endogenous complexity cost based on psychological foundations.

Thereafter, we establish several properties displayed by the optimal contract. Let a signal that is more likely to be observed the higher the agent's effort be referred to as a *good* signal. We find that the subset of signals that are rewarded with the high wage contains either only good signals, though possibly not all good signals, or all good signals and possibly a few bad signals as well.[6] Moreover, it is shown that, at least under a certain condition, it is optimal for the principal to order the signals according to their relative

[6]The theoretical prediction that inferior performance may also well be rewarded with a bonus is in line with both Joseph and Kalwani (1998)'s suggestion that organizations tend to view the payment of a bonus as a reward for good or even acceptable performance rather than an award for exceptional performance, and Churchill et al. (1993)'s prescription that bonuses should be based on objectives that can be achieved with reasonable rather than Herculean efforts.

informativeness (likelihood ratio), i.e., the agent receives the high wage for all signals that are more indicative for high effort than a cutoff signal. Though wage payments are only weakly increasing in the likelihood ratio, this finding resembles Holmström (1979)'s result for a risk averse agent, where the incentive scheme is strictly increasing in likelihood ratios. Last, we establish an interesting comparative static property: we show that an increase in the agent's degree of loss aversion may allow the principal to use a lower-powered incentive scheme in order to implement a desired level of effort. The reason is that a higher degree of loss aversion may be associated with a stronger incentive for the agent to choose a high effort in order to reduce the probability of incurring a loss. This finding immediately relates to a train of thought found in Kőszegi and Rabin (2006), who reason that under loss aversion the agent's motivation goes beyond pure monetary incentives. Section 3 concludes with a discussion of the general case where the agent is both risk averse and loss averse. It is shown that our results are robust towards a small degree of risk aversion. Moreover, we give a heuristic reasoning why a reduction in the complexity of the contract is also to be expected to be optimal for a non-negligible degree of risk aversion, and we back this argument up with a numerical example, which confirms our conjecture.[7]

Returning to the case of a purely loss averse agent, in Section 4 we relax the assumptions that guaranteed validity of the first-order approach. Moreover, to keep the analysis without first-order approach tractable, we only consider binary measures of performances. If the agent's degree of loss aversion is sufficiently high and if the performance measure is – in an intuitive sense – sufficiently informative, then only extreme actions – work as hard as possible or do not work at all – are incentive compatible. Put differently, the principal may face severe problems in fine-tuning the agent's incentives. These implementation problems, however, can be remedied if the principal can commit herself to turning a blind eye from time to time, that is, by stochastically ignoring the low realization of the performance measure. Besides alleviating implementation problems, turning a blind eye may also lower the cost of implementing a certain action. An interesting implication of these findings is that the sufficiency part of Blackwell's celebrated theorem does not hold in our model when the agent has reference-dependent preferences.

After briefly summarizing our main findings, in Section 5 we conclude by discussing robustness of our results with respect to imposed functional assumptions and the equilibrium concept applied to solve for the behavior of the loss averse agent.

Related Literature Before presenting our model, we would like to relate our paper to the small but steadily growing literature that analyzes the implications of loss aversion on incentive design.[8] With reference-dependent preferences being at the heart of loss aversion

[7] This finding also relates to the observation that, within a firm, pay for individuals often seems to be less variable than productivity, as recently surveyed by Lazear and Shaw (2007). Our model suggests an alternative explanation for this pay compression outside the realms of inequity aversion, tournament theory, and influence activities.

[8] Beside loss aversion there are other behavioral biases that are incorporated into models of incentive design. For instance, O'Donoghue and Rabin (1999) analyze optimal incentive schemes for time inconsistent agents, and Englmaier and Wambach (2006) characterize the optimal contract for the

on the one hand, but with no unifying approach provided how to determine a decision maker's reference point on the other hand, it is little surprising that all contributions differ in this particular aspect. While Dittmann et al. (2007) posit that the reference income is exogenously given by the previous year's fixed wage, Iantchev (2005), who considers a market environment with multiple principals competing for the services of multiple agents, applies the concept of Rayo and Becker (2007). Here, an agent's reference point is endogenously determined by the equilibrium conditions in the market. When focusing on a particular principal-agent pair, however, both the principal and the agent take the reference point as exogenously given. An exogenous reference point does not always seem plausible. Starting out from the premise that the reference point is forward looking and depends on the distributions of outcomes, as suggested by ample evidence, De Meza and Webb (2007) consider both exogenous as well as endogenous formulations of the reference point. Concluding that the disappointment concept of Gul (1991), which equates the reference point with the certainty equivalent of the income distribution, does yield some questionable implications,[9] De Meza and Webb propose that the reference income is the median income. Giving a brief heuristic reasoning why this may be a reasonable first pass, they argue that making the reference point equal to the median income captures the idea that the agent incurs a loss at all incomes for which it is odds-on that a higher income would be drawn. Taking median income as reference income, however, suffers from the obvious drawback that it is discontinuous in the underlying probability distribution.[10] By weighting each gain or loss with the probability that it actually occurs, the concept of reference-dependent preferences introduced by Kőszegi and Rabin (2006) avoids this kind of discontinuity. With the reference point being determined by the decision maker's expectations about outcomes, this most recent approach pursues the road most consistently that expectations matter in the determination of the reference point.

All of the aforementioned contributions explore questions of both empirical importance as well as theoretical interest: Dittmann et al. (2007) find that a loss-aversion model dominates an equivalent risk-aversion model in explaining observed CEO compensation contracts if the reference point is equal to the previous year's fixed wage. Iantchev (2005) finds evidence for his theoretically predicted results in panel data from Safelite Glass Corporation. Last, by explaining why bonuses are paid for good performance rather than penalties for poor performance, De Meza and Webb (2007) provide a theoretical underpinning for the frequent usage of option-like incentive schemes in CEO compensation. The contractual form predicted by these papers, however, is rather complex: while the

case of an inequity averse agent in the sense of Fehr and Schmidt (1999). For a review of behavioral economics of organizations see Camerer and Malmendier (2007).

[9] De Meza and Webb consider two otherwise identical agents who differ only in their degree of loss aversion. They point out that with the certainty equivalent as reference point, there are situations where the less loss-averse agent experiences a loss, but the more loss-averse agent does not.

[10] For example, suppose that with a probability of .51 a manager earns \$1m and with a probability of .49 he earns \$2m. With median income as reference point the manager will never suffer a loss because his reference income is \$1m. A small shift in probabilities, however, makes the median income equal to \$2m. Now, the agent suffers a loss in almost 50% of all cases.

optimal contract typically displays a range where pay is independent of performance, for performance above this range payment varies with performance in a fairly complex way, depending crucially on the underlying distribution of signals. Theoretical predictions differ in whether or not the optimal contract includes punishment for very poor performance or where in the wage schedule the optimal contract features discontinuities. Thus, none of these papers provides a rationale for the prevalence of fairly simple contracts, bonus contracts in particular.[11]

To the best of our knowledge, Daido and Itoh (2007) is the only paper that also applies the concept of reference dependence à la Kőszegi and Rabin to a principal-agent setting. The focus of Daido and Itoh, however, greatly differs from ours. Assuming that the performance measure comprises of only two signals, two types of self-fulfilling prophecy are explained, the Galatea and the Pygmalion effects.[12] While sufficient to capture these two effects, the assumption of a binary measure of performance does not allow one to inquire into the form that contracts take under moral hazard.

2 The Model

There are two parties, a principal and an agent.[13] The principal offers a one-period employment contract to the agent. If the agent accepts the contract, then he chooses an effort level $a \in \mathcal{A} \equiv [0,1]$. The agent's action a equals the probability that the principal receives a benefit $B > 0$. The principal's expected net benefit is

$$\pi = aB - E[W] \,,$$

where W is the compensation payment the principal pays to the agent.[14] The principal is assumed to be risk and loss neutral, thus she maximizes π. We wish to inquire into the form that contracts take under moral hazard and loss aversion. Therefore, we focus on the cost minimization problem to implement a certain action $\hat{a} \in (0,1)$.[15]

The action choice $a \in \mathcal{A}$ is private information of the agent and not observable for the principal. Furthermore, it is assumed that the realization of B is not directly observable. A possible interpretation is that B corresponds to a complex good whose quality cannot be determined by a court, thus a contract cannot depend on the realization of B. Instead of observing the agent's action a or whether the benefit B was realized or not, the principal

[11] De Meza and Webb (2007) find conditions under which a simple bonus contract is optimal. For this to be the case, however, they assume that the reference point is exogenously given and that all wage payments are in the loss region, where the agent is assumed to be risk-loving.

[12] Roughly speaking, the former effect refers to empirical findings that an agent's self-expectation about his performance is an important determinant of his actual performance, whereas the latter effect refers to the phenomenon that a principal's expectation about the agent's performance has an impact on the agent's actual performance.

[13] The model is similar to the one used by MacLeod (2003) to analyze subjective performance measures. He does not discuss loss averse agents.

[14] The particular functional form of the principal's profit function is not crucial for our analysis. We assume this specific structure merely for illustrative purposes.

[15] The second-best action maximizes the principal's expected benefit, aB, minus the minimum cost of implementing action a. The overall optimal contract exhibits the same characteristics as the contract that minimizes the cost of implementing an arbitrary action $\hat{a}$.

observes a contractible measure of performance, with $s \in \mathcal{S} \equiv \{1, \ldots, S\}$ being the realization of the performance measure or the signal. Let $S \geq 2$. The probability of observing signal s conditional on B being realized is denoted by γ_s^H. Accordingly, γ_s^L is the probability of observing signal s conditional on B not being realized. With this notation, the unconditional probability of observing signal s for a given action a is $\gamma_s(a) \equiv a\gamma_s^H + (1-a)\gamma_s^L$. For technical convenience, we make the following assumption.

Assumption (A1): *For all $s, \tau \in \mathcal{S}$ with $s \neq \tau$,*

(i) $\gamma_s^H/\gamma_s^L \neq 1$ *(informative signals),*

(ii) $0 < \gamma_s^H/\gamma_s^L < \infty$ *(full support),*

(iii) $\gamma_s^H/\gamma_s^L \neq \gamma_\tau^H/\gamma_\tau^L$ *(different signals).*

The assumption $\gamma_s^H/\gamma_s^L \neq 1$ for any s is a technical assumption that holds generically. It guarantees that any signal s is either a good or a bad signal, in the sense that the overall probability of observing that signal unambiguously increases or decreases in a. By assuming that $0 < \gamma_s^H/\gamma_s^L < \infty$ for all s, the standard full support assumption is satisfied, since for $a \in \mathcal{A}$, all signals occur with positive probability. Last, the assumption $\gamma_s^H/\gamma_s^L \neq \gamma_\tau^H/\gamma_\tau^L$ for all $s \neq \tau$ ensures that the signals can unambiguously be ranked according to the relative impact of an increase in effort on the probability of observing a particular signal.[16]

The contract which the principal offers to the agent consists of a payment for each realization of the performance measure, $\{w_s\}_{s=1}^S \in \mathbb{R}^S$.[17]

The agent is assumed to have reference-dependent preferences in the sense of Kőszegi and Rabin (2006): Overall utility from consuming $\boldsymbol{x} = (x_1, \ldots, x_K) \in \mathbb{R}^K$ – when having reference level $\boldsymbol{r} = (r_1, \ldots, r_K) \in \mathbb{R}^K$ for each dimension of consumption – is given by

$$v(\boldsymbol{x}|\boldsymbol{r}) \equiv \sum_{k=1}^{K} m_k(x_k) + \sum_{k=1}^{K} \mu(m_k(x_k) - m_k(r_k)).$$

Put verbally, overall utility is assumed to have two components: consumption utility and gain-loss utility. Consumption utility, also called intrinsic utility, from consuming in dimension k is denoted by $m_k(x_k)$. How a person feels about gaining or losing in a dimension is assumed to depend in a universal way on the changes in consumption utility associated with such gains and losses. The universal gain-loss function $\mu(\cdot)$ satisfies the assumptions imposed by Tversky and Kahneman (1991) on their "value function".[18] In our model, the agent's consumption space comprises of two dimensions, money income ($x_1 = W$) and effort ($x_2 = a$). The agent's intrinsic utility for money is assumed to be a

[16] Formally, for all $a \in [0,1]$, $(\gamma_s^H - \gamma_s^L)/\gamma_s(a) > (\gamma_\tau^H - \gamma_\tau^L)/\gamma_\tau(a) \iff \gamma_s^H/\gamma_s^L > \gamma_\tau^H/\gamma_\tau^L$.

[17] The restriction $w_s \in \mathbb{R}$ for all $s \in S$ is standard in the principal-agent literature and also in accordance with observed practice. In a later section, however, we comment on this assumption.

[18] Roughly speaking, $\mu(z)$ is strictly increasing, continuous for all z, twice differentiable for all $z \neq 0$ with $\mu(0) = 0$, convex over the range of losses, and concave over the range of gains. For a more formal statement of these properties, see Bowman et al. (1999).

strictly increasing, (weakly) concave, and unbounded function. Formally, $m_1(W) = u(W)$ with $u'(\cdot) > \varepsilon > 0$, $u''(\cdot) \leq 0$. The intrinsic disutility from exerting effort $a \in [0, 1]$ is a strictly increasing, strictly convex function of effort, $m_2(a) = -c(a)$ with $c'(\cdot) > 0$, $c''(\cdot) > 0$, $c'(0) = 0$, and $\lim_{a \to 1} c(a) = \infty$. We assume that the gain-loss function is piece-wise linear,

$$\mu(x) = \begin{cases} x \ , & \text{for } x \geq 0 \\ \lambda x, & \text{for } x < 0 \end{cases} .$$

The parameter λ characterizes the weight put on losses relative to gains.[19] The weight on gains is normalized to one. When $\lambda > 1$, the agent is loss averse in the sense that losses loom larger than equally-sized gains.[20] Last, the agent has an outside employment opportunity (or reservation utility) yielding expected utility $\bar{u}$.

Following Kőszegi and Rabin (2006, 2007), the agent's reference point is determined by his rational expectations about outcomes. A given outcome is then evaluated by comparing it to all possible outcomes, where each comparison is weighted with the probability with which the alternative outcome occurs ex-ante. With the actual outcome being itself uncertain, the agent's ex-ante expected utility is obtained by averaging over all these comparisons.[21] We apply the concept of choice-acclimating personal equilibrium (CPE) as defined in Kőszegi and Rabin (2007), which assumes that a person correctly predicts her choice set, the environment she faces, in particular the set of possible outcomes and how the distribution of these outcomes depends on her decisions, and her own reaction to this environment. The eponymous feature of CPE is that the agent's reference point is affected by his choice of action. As pointed out by Kőszegi and Rabin, CPE refers to the analysis of risk preferences regarding outcomes that are resolved long after all decisions are made. This environment seems well-suited for many principal-agent relationships. For often the outcome or the return of a project becomes observable, and thus performance-based wage compensation of the agent feasible, long after the agent finished working on that project. Under CPE, the expectations relative to which a decision's outcome is evaluated are formed at the moment when the decision is made, and therefore incorporate

[19] Alternatively, one could assume that $\mu(x) = \eta x$ for gains and $\mu(x) = \eta\lambda x$ for losses, where $\eta \geq 0$ can be interpreted as the weight attached to gain-loss utility relative to intrinsic utility. Our implicit normalization $\eta = 1$ is without loss of generality due to the applied concept of choice-acclimating personal equilibrium (CPE). Carrying η through the whole analysis would only replace $(\lambda - 1)$ by $\eta(\lambda - 1)$ in all formulas.

[20] The assumption of a piece-wise linear gain-loss function is not uncommon in the literature on incentive design with loss averse agents, see De Meza and Webb (2007), Daido and Itoh (2007). In their work on asset pricing, Barberis et al. (2001) also apply this particular functional form, reasoning that "curvature is most relevant when choosing between prospects that involve only gains or between prospects that involve only losses. For gambles that can lead to both gains and losses, [...] loss aversion at the kink is far more important than the degree of curvature away from the kink."

[21] Suppose the actual outcome $\boldsymbol{x}$ and the vector of reference levels $\boldsymbol{r}$ are distributed according to distribution functions F and G, respectively. As introduced above, overall utility from two arbitrary vectors $\boldsymbol{x}$ and $\boldsymbol{r}$ is given by $v(\boldsymbol{x}|\boldsymbol{r})$. With the reference point being distributed according to probability measure G, the utility from a certain outcome is the average of how this outcome feels compared to all other possible outcomes, $U(\boldsymbol{x}|G) = \int v(\boldsymbol{x}|\boldsymbol{r}) \, dG(\boldsymbol{r})$. Last, with $\boldsymbol{x}$ being drawn according to probability measure F, utility is given by $E[U(F|G)] = \iint v(\boldsymbol{x}|\boldsymbol{r}) \, dG(\boldsymbol{r}) \, dF(\boldsymbol{x})$. Due to the applied equilibrium concept, choice acclimating personal equilibrium, we will have $F = G$.

the implications of the decision. More precisely, suppose the agent chooses action a and that signal s is observed. The agent receives wage w_s and incurs effort cost $c(a)$. While the agent expected signal s to come up with probability $\gamma_s(a)$, with probability $\gamma_\tau(a)$ he expected signal $\tau \neq s$ to be observed. If $w_\tau > w_s$, the agent experiences a loss of $\lambda(u(w_s) - u(w_\tau))$, whereas if $w_\tau < w_s$, the agent experiences a gain of $u(w_s) - u(w_\tau)$. If $w_s = w_\tau$, there is no sensation of gaining or losing involved. The agent's utility from this particular outcome is given by

$$u(w_s) + \sum_{\{\tau|w_\tau<w_s\}} \gamma_\tau(a)(u(w_s) - u(w_\tau)) + \sum_{\{\tau|w_\tau\geq w_s\}} \gamma_\tau(a)\lambda(u(w_s) - u(w_\tau)) - c(a).$$

Averaging over all possible outcomes yields the agent's expected utility from choosing action a:

$$E[U(a)] = \sum_{s=1}^{S} \gamma_s(a)\Bigg\{u(w_s) \quad + \sum_{\{\tau|w_\tau<w_s\}} \gamma_\tau(a)(u(w_s) - u(w_\tau)) + \sum_{\{\tau|w_\tau\geq w_s\}} \gamma_\tau(a)\lambda(u(w_s) - u(w_\tau))\Bigg\} - c(a).$$

Note that since the agent's expected and actual effort choice coincide, there is neither a gain nor a loss in the effort dimension.

We conclude this section by briefly summarizing the underlying timing of the described principal-agent relationship.

1) The principal makes a take-it-or-leave-it offer $\{w_s\}_{s=1}^S$ to the agent.

2) The agent either accepts or rejects the contract. If the agent rejects the contract the game ends and each party receives her/his reservation payoff. If the agent accepts the contract the game moves to the next stage.

3) The agent chooses his action and forms rational expectations about the monetary outcomes. The agent's rational expectations about the realization of the performance measure determine his reference point.

4) Both parties observe the realization of the performance measure and payments are made according to the contract.

3 The Analysis

Let the inverse function of the agent's intrinsic utility of money be $h(\cdot)$, i.e., $h(\cdot) := u^{-1}$. Put differently, the monetary cost for the principal to offer the agent utility u_s is $h(u_s) = w_s$. Due to the assumptions imposed on $u(\cdot)$, $h(\cdot)$ is a strictly increasing and weakly convex function. Following Grossman and Hart (1983) we regard $\boldsymbol{u} = \{u_1, \ldots, u_S\}$ as the principal's control variables in her cost minimization problem to implement action $\hat{a} \in (0, 1)$. The principal offers the agent a contract that specifies for each signal a

monetary payment or, equivalently, an intrinsic utility level. With this notation the agent's expected utility from exerting effort a is given by

$$E[U(a)] = \sum_{s\in\mathcal{S}} \gamma_s(a)u_s - (\lambda-1)\sum_{s\in\mathcal{S}}\sum_{\{\tau|u_\tau>u_s\}} \gamma_\tau(a)\gamma_s(a)(u_\tau-u_s) - c(a). \tag{1}$$

From the above formulation of the agent's utility it becomes clear that λ captures not only the weight put on losses relative to gains, but $(\lambda-1)$ also characterizes the weight put on gain-loss utility relative to intrinsic utility. Thus, for $\lambda \leq 2$, the weight attached to gain-loss utility is below the weight attached to intrinsic utility. Note that for $\lambda = 1$ the agent's expected utility equals expected net intrinsic utility. Thus, for $\lambda = 1$ we are in the standard case without loss aversion. For a given contract $\boldsymbol{u}$, the agent's marginal utility of effort amounts to

$$\begin{aligned} E[U'(a)] = &\sum_{s\in\mathcal{S}}(\gamma_s^H-\gamma_s^L)u_s \\ &-(\lambda-1)\sum_{s\in\mathcal{S}}\sum_{\{\tau|u_\tau>u_s\}}[\gamma_\tau(a)(\gamma_s^H-\gamma_s^L)+\gamma_s(a)(\gamma_\tau^H-\gamma_\tau^L)](u_\tau-u_s)-c'(a). \end{aligned} \tag{2}$$

Suppose the principal wants to implement action $\hat{a} \in (0,1)$. The optimal contract minimizes the expected wage payment to the agent subject to the usual incentive compatibility and individual rationality constraints:

$$\begin{aligned} &\min_{u_1,\ldots,u_S} \sum_{s\in\mathcal{S}} \gamma_s(\hat{a})h(u_s) \\ \text{subject to}\quad &E[U(\hat{a})] \geq \bar{u}\ , &\text{(IR)}\\ &\hat{a} \in \arg\max_{a\in\mathcal{A}} E[U(a)]\ . &\text{(IC)} \end{aligned}$$

As a first benchmark consider the case where the agent's action choice is observable and contractible, i.e., the incentive constraint (IC) is absent. In order to implement action $\hat{a}$ in this first-best situation, the principal pays the agent $u^{FB} = \bar{u} + c(\hat{a})$ irrespective of the realization of the performance measure if the agent chooses the desired action, thereby compensating him for his outside option and his effort cost.

At this point we simplify the analysis by imposing two assumptions. These assumptions are sufficient to guarantee that the principal's cost minimization problem exhibits the following two important properties. First, there are incentive-compatible wage contracts, i.e., contracts under which it is optimal for the agent to choose the desired action $\hat{a}$. Existence of such contracts is not generally satisfied with the agent being loss averse. Second, the first-order approach is valid, i.e., the incentive constraint to implement action $\hat{a}$ can equivalently be represented as follows: $E[U'(\hat{a})] = 0$. The first assumption that we introduce requires that the weight attached to gain-loss utility does not exceed the weight put on intrinsic utility.

Assumption (A2): *No dominance of gain-loss utility,* $\lambda \leq 2$.

As carefully laid out in Kőszegi and Rabin (2007), CPE implies a strong notion of risk aversion, in the sense that a decision maker may choose stochastically dominated options when $\lambda > 2$, i.e. when his weight attached to the impact of loss aversion exceeds the weight attached to consumption utility.[22] The reason is that, with losses looming larger than gains of equal size, the person ex-ante expects to experience a net loss. In consequence, if reducing the scope of possibly incuring a loss is the decision maker's primary concern, that person would rather give up the slim hope of experiencing a gain at all in order to avoid the disappointment in case of not experiencing this gain. In our model, if the agent is sufficiently loss averse, the principal may be unable to implement any action $\hat{a} \in (0,1)$. The reason is that the agent minimizes the ex-ante expected net loss by choosing one of the two extreme actions. The values of λ for which this behavior is optimal for the agent crucially depend on the precise structure of the performance measure. Assumption $(A2)$ is sufficient, but by far not necessary, to ensure that there is a contract such that $\hat{a} \in (0,1)$ is incentive compatible. In Section 4, we relax Assumption $(A2)$ and discuss in detail the implications of $\lambda > 2$ on the contractual arrangement. Though calibrationally not inconsistent, the tendency to choose stochastically dominated options seems counterintuitive.[23] Next to ensuring existence of an incentive compatible contract, $(A2)$ rules out that our findings are driven by such counterintuitive behavior of the agent.

To keep the analysis tractable we impose the following assumption which ensures – given $(A2)$ holds – that the first-order approach is valid.[24]

Assumption (A3): *Convex marginal cost function,* $\forall\, a \in [0,1]:\ c'''(a) \geq 0$.

We want to emphazise that – given $(A2)$ – Assumption $(A3)$ is a sufficient but not necessary condition for the first-order approach to be applicable. For the first-order approach to be valid it would also suffice to have λ sufficiently small, or the slope of the marginal cost function sufficiently steep. Our results require the validity of the first-order approach, not that Assumption $(A3)$ holds. In Section 4 we shed some more light on what happens when the first-order approach is not valid.

Lemma 1: *Given* $(A1)$-$(A3)$*, the constraint set of the principal's minimization problem is non-empty for all* $\hat{a} \in (0,1)$.

Proof: See Appendix.

[22] Suppose a loss-averse person has to choose between two lotteries: lottery 1 pays x for sure; lottery 2 pays $x+y$ with probability p, where $y > 0$, and x otherwise. Then, for each $\lambda > 2$, the decision maker prefers the dominated lottery 1 if $p < (\lambda - 2)/(\lambda - 1)$. For further details on this point, see Kőszegi and Rabin (2007).

[23] The "uncertainty effect" identified by Gneezy et al. (2006) refers to people valuing a risky prospect less than its worst possible outcome. While this may be interpreted as experimental evidence for people having preferences for stochastically dominated options, this finding crucially relies on the lottery currency not being stated in purely monetary terms. Therefore, we believe that in the context of wage contracts most people do not choose dominated options.

[24] The validity of the first-order approach under assumptions (A1)-(A3) is rigorously proven in the appendix. The reader should be aware, however, that the proof requires some notation introduced later on. We therefore recommend to defer reading the proof until having read the preliminary considerations up to Section 3.1.

The above lemma states that there are wage contracts such that the agent is willing to accept the contract and then chooses the desired action. Moreover, we will show that a second-best optimal contract exists. This, however, is shown separately for the three cases that are analyzed in this section.

Sometimes it will be convenient to state the constraints in terms of increases in intrinsic utilities instead of absolute utilities. Note that whatever contract $\{\hat{u}_s\}_{s\in\mathcal{S}}$ the principal offers, we can always relabel the signals such that this contract is equivalent to a contract $\{u_s\}_{s=1}^S$ with $u_{s-1} \leq u_s$ for all $s \in \{2,\ldots,S\}$. This, in turn, allows us to write the contract as $u_s = u_1 + \sum_{\tau=2}^{s} b_\tau$, where $b_\tau = u_\tau - u_{\tau-1} \geq 0$ is the increase in intrinsic utility for money when signal τ instead of signal $\tau - 1$ is observed. Let $\boldsymbol{b} = (b_2,\ldots,b_S)$. Using this notation allows us to rewrite the individual rationality constraint as follows:

$$u_1 + \sum_{s=2}^{S} b_s \left[\sum_{\tau=s}^{S} \gamma_\tau(\hat{a}) - \rho_s(\hat{\boldsymbol{\gamma}},\lambda,\hat{a})\right] \geq \bar{u} + c(\hat{a}) \;, \tag{IR$'$}$$

where

$$\rho_s(\hat{\boldsymbol{\gamma}},\lambda,\hat{a}) := (\lambda-1)\left[\sum_{\tau=s}^{S}\gamma_\tau(\hat{a})\right]\left[\sum_{t=1}^{s-1}\gamma_t(\hat{a})\right]. \tag{3}$$

Let $\boldsymbol{\rho}(\hat{\boldsymbol{\gamma}},\lambda,\hat{a}) = (\rho_2(\hat{\boldsymbol{\gamma}},\lambda,\hat{a}),\ldots,\rho_S(\hat{\boldsymbol{\gamma}},\lambda,\hat{a}))$. The first part of the agent's utility, $u_1 + \sum_{s=2}^{S} b_s(\sum_{\tau=s}^{S}\gamma_\tau(\hat{a}))$, is the expected intrinsic utility for money. Due to loss aversion, however, the agent's utility has a second negative component, the term $\boldsymbol{b}'\boldsymbol{\rho}(\hat{\boldsymbol{\gamma}},\lambda,\hat{a})$. Where does this term come from? With bonus b_s being paid to the agent whenever a signal higher or equal to s is observed, the agent expects to receive b_s with probability $\sum_{\tau=s}^{S}\gamma_\tau(\hat{a})$. With probability $\sum_{t=1}^{s-1}\gamma_t(\hat{a})$, however, a signal below s will be observed, and the agent will not be paid bonus b_s. Thus, with "probability" $[\sum_{\tau=s}^{S}\gamma_\tau(\hat{a})][\sum_{t=1}^{s-1}\gamma_t(\hat{a})]$ the agent experiences a loss of λb_s. Analogous reasoning implies that the agent will experience a gain of b_s with the same probability. With losses looming larger than gains of equal size, in expectation the agent suffers from deviations from his reference point. This ex-ante expected net loss is captured by the term, $\boldsymbol{b}'\boldsymbol{\rho}(\hat{\boldsymbol{\gamma}},\lambda,\hat{a})$, which we will refer to as the agent's "loss premium".[25] A crucial point is that the loss premium increases in the complexity of the contract. When there is no wage differentiation at all, i.e., $\boldsymbol{b} = \boldsymbol{0}$, then the loss premium vanishes. If, in contrast, the contract specifies many different wage payments, then the agent ex-ante considers a deviation from his reference point very likely. Put differently, for each additional wage payment an extra negative term enters the agent's loss premium and therefore reduces his expected utility.[26]

[25] Our notion of the agent's loss premium is highly related to the average self-distance of a lottery defined by Kőszegi and Rabin (2007). Let $D(\boldsymbol{u})$ be the average self-distance of incentive scheme $\boldsymbol{u}$, then $[(\lambda-1)/2]D(\boldsymbol{u}) = \boldsymbol{b}'\boldsymbol{\rho}(\hat{\boldsymbol{\gamma}},\lambda,\hat{a})$.

[26] While the exact change of the loss premium from adding more and more wage payments is hard to grasp, this point can heuristically be illustrated by considering the upper bound of the loss premium. Suppose the principal sets $n \leq S$ different wages. It is readily verified that the loss premium is bounded from above by $(\lambda-1)[(u_S-u_1)/2]\times[(n-1)/n]$, and that this upper bound increases as n increases. Note, however, that even for $n\to\infty$ the upper bound of the loss premmium is finite.

Given the first-order approach is valid, the incentive constraint can be rewritten as

$$\sum_{s=2}^{S} b_s \beta_s(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a}) = c'(\hat{a}), \tag{IC$'$}$$

where we defined

$$\beta_s(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a}) := \left(\sum_{\tau=s}^{S}(\gamma_\tau^H - \gamma_\tau^L)\right)\left[1 - (\lambda - 1)\left(\sum_{t=1}^{s-1}\gamma_t(\hat{a})\right)\right] - (\lambda - 1)\left[\sum_{\tau=s}^{S}\gamma_\tau(\hat{a})\right]\left(\sum_{t=1}^{s-1}(\gamma_t^H - \gamma_t^L)\right).$$

Here, $\beta_s(\cdot)$ is the marginal effect on incentives of an increase in the wage payments for signals above $s-1$. Without loss aversion, i.e., $\lambda = 1$, this expression equals the marginal probability of observing at least signal s. If the agent is loss averse, however, then the absolute probability of observing at least signal s also plays a role in determining this marginal effect. The reason is that the loss premium is a quadratic function of the probability of observing at least signal s. Let $\boldsymbol{\beta}(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a}) = (\beta_2(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a}), \ldots, \beta_S(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a}))$.

As in the standard case, incentives are created solely by increases in intrinsic utilities, $\boldsymbol{b}$. In consequence, (IR$'$) is binding in the optimum. If this was not the case, i.e., if $\boldsymbol{b}$ satisfies (IC$'$) but (IR$'$) holds with strict inequality, then the principal can lower payment u_1 up to the point where the (IR$'$) is satisfied with equality. Thus, reducing u_1 while holding $\boldsymbol{b}$ constant lowers the principal's expected wage payment while preserving incentives.

It is obvious that (IC$'$) can only be satisfied if there exists at least one $\beta_s > 0$. If, for example, signals are ordered according to their likelihood ratios, then $\beta_s(\cdot) > 0$ for all $s = 2, \ldots, S$. More precisely, for a given ordering of signals, under $(A2)$ the following equivalence follows immediately from the fact that $\sum_{t=1}^{s-1}(\gamma_t^H - \gamma_t^L) = -\sum_{\tau=s}^{S}(\gamma_\tau^H - \gamma_\tau^L)$:

$$\beta_s(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a}) > 0 \iff \sum_{\tau=s}^{S}(\gamma_\tau^H - \gamma_\tau^L) > 0\,. \tag{4}$$

3.1 Two Polar Cases: Pure Risk Aversion vs. Pure Loss Aversion

In this part of the paper we analyze the two polar cases: The standard case where the agent is only risk averse but not loss averse, on the one hand, and the case of a loss averse agent with a risk-neutral intrinsic utility function, on the other hand.

Pure Risk Aversion

First consider an agent who is risk-averse in the usual sense, i.e., $h''(\cdot) > 0$, but does not exhibit loss aversion. Though not immediately obvious, the latter requirement corresponds to the case where $\lambda = 1$. To see this, remember that the agent compares each outcome with each possible other outcome. Thus the comparison of any two wages enters the agent's expected utility exactly twice, once as a loss and once as an equally-sized gain.

For $\lambda = 1$, the agent puts equal weights on gains and losses, so all these comparisons just cancel out, and we are left with

$$E[U(a)] = \sum_{s=1}^{S} \gamma_s(a) u_s - c(a).$$

With the agent not being loss averse, the first-order approach obviously is valid even without Assumption $(A3)$.

Proposition 1: *Given* $(A1)$, $h''(\cdot) > 0$, *and* $\lambda = 1$. *Then there exists a second-best optimal contract to implement* $\hat{a} \in (0,1)$. *The second-best contract has the property that* $u_s \neq u_\tau \ \forall s, \tau \in \mathcal{S}$ *and* $s \neq \tau$. *Moreover,* $u_s > u_\tau$ *if and only if* $\gamma_s^H/\gamma_s^L > \gamma_\tau^H/\gamma_\tau^L$.

Proof: See Appendix.

The result in Proposition 1 is not new, since it basically restates the well-known finding by Holmström (1979): With the relative impact of a marginal increase in effort on observing a signal being increasing in the likelihood ratio γ_s^H/γ_s^L, when the agent is risk averse, signals that are more indicative of higher effort are rewarded strictly higher. Things, however, look completely different when the agent is not risk averse but loss averse.

Pure Loss Aversion

Having considered the polar case of pure risk aversion, we now turn to the other extreme, a purely loss averse agent. Formally, intrinsic utility of money income is a linear function, $h''(\cdot) = 0$, and the agent is loss averse, $\lambda > 1$. As we have already reasoned, whatever contract the principal offers, relabeling the signals always allows us to represent this contract as an (at least weakly) increasing intrinsic utility profile. Therefore we can decompose the principal's problem into two steps: first, for a given ordering of signals, choose a nondecreasing profile of intrinsic utility levels that implements the desired action $\hat{a}$ at minimum cost; second, choose the signal ordering with the lowest cost of implementation. As we know from the discussion at the end of the previous section, a necessary condition for an upward-sloping incentive scheme to achieve incentive compatibility is that for the underlying signal ordering at least one $\beta_s(\cdot) > 0$. In what follows we restrict attention to the set of signal orderings that are incentive feasible in the afore-mentioned sense. Nonemptiness of this set follows immediately from Lemma 1.

Consider the first step of the principal's problem, i.e., taking the ordering of signals as given, find the nondecreasing payment scheme with the lowest cost of implementation. In what follows, we write the agent's intrinsic utility in terms of additional payments, $u_s = u_1 + \sum_{\tau=2}^{S} b_\tau$. With $h(\cdot)$ being linear, the principal's objective function is $C(u_1, \boldsymbol{b}) = u_1 + \sum_{s=2}^{S} b_s(\sum_{\tau=2}^{S} \gamma_\tau(\hat{a}))$. Remember that in the optimum, (IR') holds with equality. Inserting (IR') into the principal's objective allows us to write the cost minimization problem for a given order of signals in the following simple way:

Program ML:

$$\min_{\boldsymbol{b}\in\mathbb{R}_{+}^{S-1}} \ \boldsymbol{b}'\boldsymbol{\rho}(\hat{\boldsymbol{\gamma}},\lambda,\hat{a})$$
$$\text{subject to} \quad \boldsymbol{b}'\boldsymbol{\beta}(\hat{\boldsymbol{\gamma}},\lambda,\hat{a}) = c'(\hat{a}) \tag{IC$'$}$$

The minimization problem (ML) has a simple intuition. The principal seeks to minimize the agent's expected net loss subject to the incentive compatibility constraint. Similar to the case of pure risk aversion, where the principal would like to cut back the agent's risk premium, here she is interested in minimizing the agent's loss premium. Due to the incentive constraint, however, this loss premium has to be strictly positive.

It is important to realize that the principal's cost minimization problem for a given order of signals is a rather simple linear programming problem: minimize a linear objective function subject to one linear equality constraint. Since we restricted attention to orderings of signals with $\beta_s(\cdot) > 0$ for at least one signal s, a solution to (ML) exists. Due to the linear nature of problem (ML), (generically) this solution sets exactly one $b_s > 0$ and all other $b_s = 0$. Put differently, the problem is to find that b_s which creates incentives at the lowest cost.

So far we have seen that, for a given ordering of signals, the principal considers it optimal to offer the agent a bonus contract: pay a low wage for signals below some threshold, and a high wage for signals above this threshold. What remains to do for the principal, in a second step, is to find the signal ordering that leads to the lowest cost of implementation. With the number of different orders of signals being finite, this problem clearly has a solution.

Proposition 2: *Given (A1)-(A3), $h''(\cdot) = 0$ and $\lambda > 1$. Then there exists a second-best optimal contract to implement action $\hat{a} \in (0,1)$. The second-best optimal incentive scheme $\{u_s^*\}_{s=1}^S$ entails a minimum of (wage) differentiation in the sense that $u_s^* = u_H^*$ for $s \in \mathcal{B}^* \subset \mathcal{S}$ and $u_s^* = u_L^*$ for $s \in \mathcal{S} \setminus \mathcal{B}^*$, where $u_H^* > u_L^*$.*

Proof: See Appendix.

According to Proposition 2, the principal considers it optimal to offer the agent a bonus contract: the contract specifies a high wage $u_s = u_H^*$ for $s \in \mathcal{B}^*$ and a low wage $u_s = u_L^*$ for $s \notin \mathcal{B}^*$, where $u_L^* < u_H^*$.[27] This endeavor to reduce the complexity of the contract is plausible, since a high degree of wage differentiation increases the agent's loss premium: with the employment contract she offers to the agent, the principal determines the dimensionality of the agent's reference point. The higher the dimensionality of the reference point is, the more likely it is that the agent incurs a loss in a particular dimension.

[27] As is well-known, without loss aversion, a broad range of contracts – including simple bonus schemes – is optimal when both the agent and the principal are risk neutral. If, in addition, the agent is protected by limited liability, Park (1995) and Demougin and Fluet (1998) show that the optimal contract is a bonus scheme. These findings, however, immediately collapse when the agent is somewhat risk averse. Our findings, on the other hand, are robust towards introducing a slightly concave intrinsic utility function, as we will illustrate in Section 3.2.

Therefore, with the concept of reference-dependent preferences developed by Kőszegi and Rabin (2006), it truly pains a person to be exposed to numerous potential outcomes. This disutility of the agent from facing several possible (monetary) outcomes which he demands for to be compensated, makes it costly for the principal to offer complex contracts. In consequence, the optimal contract entails only a minimum of wage differentiation. To provide a more intuitive explanation for this finding, consider a principal who – starting out from a given wage scheme – has to improve incentives. There are basically two ways to do so. On the one hand, the principal can introduce a new wage spread, i.e., pay slightly different wages for two signals that were rewarded equally in the original wage scheme, while keeping the differences between all other neighboring wages constant. On the other hand, the principal can increase an existing wage spread, holding constant all other spreads between neighboring wages. Both procedures increase the loss premium by increasing the size of some of the the expected losses without reducing others. Introducing a new wage spread, however, additionally increases the loss premium by increasing the ex ante expected probability of experiencing a loss. Therefore, in order to improve incentives for a loss averse agent, it is advantageous to increase a particular existing wage spread without adding to the contractual complexity in the sense of increasing the number of different wages. Under the standard notion of a risk averse agent, however, one should not expect to encounter this tendency to reduce the complexitiy of contracts. The reason is that increasing incentives by introducing a small new wage spread is basically costless for the principal because locally the agent is risk neutral. Therefore, under risk aversion different outcomes are rewarded differently.

Up to now, however, we have not specified which signals are generally included in the set $\mathcal{B}^*$. In light of the above observation, the principal's problem boils down to choosing a binary partition of the set of signals, $\mathcal{B} \subset \mathcal{S}$, which characterizes for which signals the agent receives the high wage and for which signals he receives the low wage. The wages u_L and u_H are then uniquely determined by the corresponding individual rationality and incentive compatibility constraints. The problem of choosing the optimal partition of signals, $\mathcal{B}^*$, which minimizes the principal's expected cost of implementing action $\hat{a}$ is an integer programming problem. As is typical for this class of problems, and as is nicely illustrated by the well-known "0-1 Knapsack Problem", it is not possible to provide a general characterization of the solution. The "0-1 Knapsack Problem" refers to a hiker who has to select from a group of items, all of which may be suitable for his trip, a subset that has greatest value while not exceeding the capacity of his knapsack.[28] In order to highlight that it is not possible to provide a general answer which items should be taken along, suppose that the hiker is close to exhausting his knapsack's capacity. Without further specifications, one cannot tell whether the hiker should take one last relatively large item of high value, which possibly forces him to leave space unused, or rather several small items that neatly fill the knapsack, but each of which is of only little

[28] Suppose there are n items, each item j has a value $v_j > 0$ and a weight $w_j > 0$. Let the capacity of the knapsack be $c > 0$. The 0-1 Knapsack Problem may be formulated as the following maximization problem: max $\sum_{j=1}^{n} v_j x_j$ subject to $\sum_{j=1}^{n} w_j x_j \leq c$ and $x_j \in \{0, 1\}$ for $j = 1, \ldots, n$.

value.

Next to these standard intricacies of integer programming, there is an additional difficulty in our model: the principal's objective behaves non-monotonic when including an additional signal into the "bonus set" $\mathcal{B}$. This is due to different – possibly conflicting – targets that the principal pursues when deciding how to partition the set $\mathcal{S}$. From Program ML it follows that, for a given "bonus set" $\mathcal{B}$, the minimum cost of implementing action $\hat{a}$ is given by

$$C_{\mathcal{B}} = \bar{u} + c(\hat{a}) + \frac{c'(\hat{a})(\lambda-1)P_{\mathcal{B}}(1-P_{\mathcal{B}})}{[\sum_{s\in\mathcal{B}}\gamma_s^H - \gamma_s^L][1-(\lambda-1)(1-2P_{\mathcal{B}})]}\,, \tag{5}$$

where $P_{\mathcal{B}} := \sum_{s\in\mathcal{B}}\gamma_s(\hat{a})$. The above costs can be rewritten such that the principal's problem amounts to

$$\max_{\mathcal{B}\subset\mathcal{S}} \left[\sum_{s\in\mathcal{B}}(\gamma_s^H - \gamma_s^L)\right]\left\{\frac{1}{(\lambda-1)P_{\mathcal{B}}(1-P_{\mathcal{B}})} - \frac{1}{P_{\mathcal{B}}} + \frac{1}{1-P_{\mathcal{B}}}\right\}\,. \tag{6}$$

This objective function illustrates the tradeoff that the principal faces when deciding how to partition the signal space. The first term, $\sum_{s\in\mathcal{B}}(\gamma_s^H - \gamma_s^L)$, is the aggregate marginal impact of effort on the probability of the bonus $b := u_H - u_L$ being paid out. In order to create incentives for the agent, the principal would like to make this term as large as possible. This can be achieved by including only good signals in $\mathcal{B}$. The second term, on the other hand, is maximized by making the probability of paying the agent the high wage either as large as possible or as small as possible, depending on the exact signal structure and the action which is to be implemented. Intuitively, by making the event of paying the high wage very likely or unlikely, the principal minimizes the scope for the agent to experience a loss that he demands to be compensated for. Depending on the signal structure, these two goals may conflict with each other, which makes a complete characterization of the optimal contract very intricate. Nevertheless, it can be shown that the optimal contract displays the following very plausible property.

Proposition 3: *Let $\mathcal{S}^+ \equiv \{s \in \mathcal{S} | \gamma_s^H - \gamma_s^L > 0\}$. The optimal partition of the signals for which the high wage is paid, $\mathcal{B}^*$, has the following property: Either $\mathcal{B}^* \subseteq \mathcal{S}^+$ or $\mathcal{S}^+ \subseteq \mathcal{B}^*$.*

Proof: See Appendix.

Put verbally, the optimal partition of the signal set takes one of the two possible forms: the high wage is paid out to the agent (i) either only for good signals though possibly not for all good signals, or (ii) for all good signals and possibly a few bad signals as well. Loosely speaking, if the principal considers it optimal to pay the high wage very rarely, she will reward only good signals with the extra payment b. If, on the other hand, she wants the agent to receive the high wage with high probability, then she will reward at least all good signals.

Without further assumptions, due to the discrete nature of the problem it is hard to characterize the signals that are included in $\mathcal{B}^*$. Back to the "0-1 Knapsack Problem",

here it is well-established for the continuous version of the problem that the solution can easily be found by ordering the items according to their value-to-weight ratio.[29] Even though our problem is clearly more complex, we can obtain a similar result. Define $\kappa := \max_{\{s,t\}\subseteq\mathcal{S}} |\gamma_s(\hat{a}) - \gamma_t(\hat{a})|$. Assuming that κ is sufficiently small makes the principal's problem of choosing $\mathcal{B}^*$ similar to a continuous problem. With this assumption, we can show that it is optimal to order the signals according to their likelihood ratios.

Proposition 4: *Suppose κ is sufficiently small, then there exists a constant K such that $\mathcal{B}^* = \{s \in \mathcal{S} \mid \gamma_s^H/\gamma_s^L \geq K\}$.*

Proof: See Appendix.

Before moving on to the discussion of the more general case where the agent is both risk averse and loss averse, we would like to pause to point out an interesting comparative static result.

Proposition 5: *An increase in the agent's degree of loss aversion (i) decreases the necessary wage spread to implement action $\hat{a}$ if and only if $P_{\mathcal{B}^*} > 1/2$, given that the change in λ does not lead to a change of $\mathcal{B}^*$; (ii) strictly increases the minimum cost of implementing action $\hat{a}$.*

Proof: See Appendix.

Part (i) of Proposition 5 relates to the reasoning by Kőszegi and Rabin (2006) that if the agent is loss averse and expectations are the driving force in the determination of the reference point, then "in principal-agent models, performance-contingent pay may not only directly motivate the agent to work harder in pursuit of higher income, but also indirectly motivate [him] by changing [his] expected income and effort." As can be seen from (1), the agent's expected utility under the second-best contract comprises of two components, the first of which is expected net intrinsic utility from choosing effort level $\hat{a}$, $u_L + b^* \sum_{s\in\mathcal{B}^*} \gamma_s(\hat{a}) - c(\hat{a})$. Due to loss aversion, however, there is a second component: With expected losses looming larger than equally sized gains, in expectation the agent suffers from deviations from his reference point. While the strength of this effect is determined by the degree of the agent's loss aversion, λ, his action choice – together with the signal parameters – determines the probability that such a deviation from the reference point actually occurs. We refer to this probability, which is given by $P_{\mathcal{B}^*}(1-P_{\mathcal{B}^*})$, as loss probability. Therefore, when choosing his action, the agent has to balance off two possibly conflicting targets, maximizing expected net intrinsic utility and minimizing the loss probability. The loss probability, which is a strictly concave function of the agent's effort, is locally decreasing at $\hat{a}$ if and only if $P_{\mathcal{B}^*} > 1/2$. In this case, an increase in λ, which makes reducing the loss probability more important, may lead to the agent choosing a higher effort level, which in turn allows the principal to use lower-powered incentives.

[29] In the continuous "0-1 Knapsack Problem" the constraints on the variables $x_j \in \{0, 1\}$ are relaxed to $x_j \in [0, 1]$. The continuous problem was elegantly solved by Dantzig (1957).

This probably is the effect Kőszegi and Rabin had in mind when reasoning that under loss aversion the agent's motivation goes beyond pure monetary incentives. The principal, however, is not able to benefit from the fact that an increase in the agent's degree of loss aversion may facilitate the creation of incentives. Even though an increase in λ may allow for implementation of $\hat{a}$ by means of a lower-powered incentive scheme, according to part (ii) of Proposition 5, the overall cost of implementation strictly increases in the agent's degree of loss aversion.

3.2 The General Case: Loss Aversion and Risk Aversion

We now turn to the intermediate case where the agent is both risk averse and loss averse. The agent's intrinsic utility for money is a strictly increasing and strictly concave function, i.e., $u'(\cdot) > 0$ and $u''(\cdot) < 0$, which implies that $h(\cdot)$ is strictly increasing and strictly convex. Moreover, the agent is loss averse, i.e., $\lambda > 1$. From Lemma 1 we know that the constraint set of the principal's problem even in this general case is nonempty. By relabeling signals, each contract can be interpreted as a contract that offers the agent a (weakly) increasing intrinsic utility profile. This allows us to assess whether the agent perceives receiving u_s instead of u_t as a gain or a loss. As in the case of pure loss aversion, we analyze the optimal contract for a given feasible ordering of signals.

The principal's problem for a given arrangement of the signals is given by:

Program MG:

$$\min_{u_1,\ldots,u_S} \sum_{s=1}^{S} \gamma_s(\hat{a}) h(u_s)$$

subject to

$$\sum_{s=1}^{S} \gamma_s(\hat{a}) u_s - (\lambda - 1) \sum_{s=1}^{S-1} \sum_{t=s+1}^{S} \gamma_s(\hat{a}) \gamma_t(\hat{a}) [u_t - u_s] - c(\hat{a}) = \bar{u} \tag{IR$_G$}$$

$$\sum_{s=1}^{S} (\gamma_s^H - \gamma_s^L) u_s - (\lambda - 1) \sum_{s=1}^{S-1} \sum_{t=s+1}^{S} \left[\gamma_s(\hat{a})(\gamma_t^H - \gamma_t^L) + \gamma_t(\hat{a})(\gamma_s^H - \gamma_s^L) \right] [u_t - u_s] = c'(\hat{a}) \tag{IC$_G$}$$

$$u_S \geq u_{S-1} \geq \ldots \geq u_1 \tag{OC$_G$}$$

Note that the objective function is strictly convex and the constraints are all linear in $\boldsymbol{u} = \{u_1, \ldots, u_S\}$. Therefore, the Kuhn-Tucker theorem yields necessary and sufficient conditions for optimality. Put differently, if there exists a solution to the problem (MG) the solution is characterized by the partial derivatives of the Lagrangian associated with (MG) set equal to zero.

Lemma 2: *Given (A1)-(A3) and $h''(\cdot) > 0$, there exists a second-best optimal incentive scheme for implementing action $\hat{a} \in (0,1)$, denoted $\boldsymbol{u}^* = \{u_1^*, \ldots, u_S^*\}$.*

Proof: See Appendix

In order to interpret the first-order conditions of the Lagrangian to problem (MG) it is necessary to know whether the Lagrangian multipliers are positive or negative.

Lemma 3: *The Lagrangian multipliers of program (MG) associated with the incentive compatibility constraint and the individual rationality constraint are both strictly positive, i.e.,* $\mu_{IC} > 0$ *and* $\mu_{IR} > 0$*, respectively.*

Proof: See Appendix

Having established the sign of these Lagrangian multipliers, we now give a heuristic reasoning why pooling of information may well be optimal in this more general case where the agent is both risk averse and loss averse. For the sake of argumentation, suppose there is no pooling of information in the sense that it is optimal to set distinct wages for distinct signals, then all the order constraints are slack. Formally, if $u_s \neq u_{s'}$ for all $s, s' \in \mathcal{S}$ and $s \neq s'$, then $\mu_{O,s} = 0$. In this case, i.e., when none of the ordering constraints is binding, then the first-order condition of optimality with respect to u_s, $\partial \mathcal{L}(\boldsymbol{u})/\partial u_s = 0$, can be written as follows:

$$h'(u_s) = \underbrace{\left(\mu_{IR} + \mu_{IC}\frac{\gamma_s^H - \gamma_s^L}{\gamma_s(\hat{a})}\right)}_{=:H_s} \underbrace{\left[1 - (\lambda - 1)\left(2\sum_{t=1}^{s-1}\gamma_t(\hat{a}) + \gamma_s(\hat{a}) - 1\right)\right]}_{=:\Gamma_s} - \underbrace{\mu_{IC}(\lambda - 1)\left[2\sum_{t=1}^{s-1}(\gamma_t^H - \gamma_t^L) + (\gamma_s^H - \gamma_s^L)\right]}_{=:\Lambda_s}. \quad (7)$$

For $\lambda = 1$ we have $h'(u_s) = H_s$, the standard "Holmström-formula".[30] Note that $\Gamma_s > 0$ for $\lambda \leq 2$. More importantly, irrespective of the signal ordering, we have $\Gamma_s > \Gamma_{s+1}$. The third term, Λ_s, can be either positive or negative. If the compound signal of all signals below s is a bad signal, then $\Lambda_s < 0$.

Since the incentive scheme is nondecreasing, when the order constraints are not binding it has to hold that $h'(u_s) \geq h'(u_{s-1})$. Thus, if $\mu_{OC,s-1} = \mu_{OC,s} = \mu_{OC,s+1} = 0$ the following inequality is satisfied:

$$H_s \times \Gamma_s - \Lambda_s \geq H_{s-1} \times \Gamma_{s-1} - \Lambda_{s-1}. \quad (8)$$

Even when $H_s > H_{s-1}$, as it is the case when signals are ordered according to their likelihood ratio, it is not clear that inequality (8) is satisfied. In particular, when s and $s-1$ are good signals it seems to be likely that inequality (8) is violated, because then $\Lambda_s > \Lambda_{s-1}$ and $\Gamma_s < \Gamma_{s-1}$. In summary, it may well be that for a given incentive-feasible ordering of signals, and thus overall as well, the order constraints are binding, i.e., from the principal's point of view it may be optimal to offer a contract which is less complex than the signal space allows for. We illustrate this conjecture in the following with an example.

[30] See Holmström (1979).

Example: Suppose $h(u) = u^r$, with $r \geq 0$ being a measure for the agent's risk aversion. More precisely, the Arrow-Pratt measure for relative risk aversion of the agent's intrinsic utility function is $R = 1 - \frac{1}{r}$ and therefore constant. First, we show for the case of this intrinsic utility function of the CRRA type that the optimal contract is still a bonus contract when the agent is not only loss averse, but also slightly risk averse.

Proposition 6: *Given (A1)-(A3), $h(u) = u^r$ with $r > 1$, and $\lambda > 1$. Generically, for r sufficiently small the optimal incentive scheme $\{u_s^*\}_{s=1}^S$ is a bonus scheme, i.e., $u_s^* = u_H^*$ for $s \in \mathcal{B}^* \subset \mathcal{S}$ and $u_s^* = u_L^*$ for $s \in \mathcal{S} \backslash \mathcal{B}^*$ where $u_L^* < u_H^*$.*

Proof: See Appendix

Next, we demonstrate that pooling of signals may well be optimal even for a non-negligible degree of risk aversion. Suppose the agent's effort cost are $c(a) = (1/2)a^2$ and the effort level to be implemented is $\hat{a} = \frac{1}{2}$. Moreover, we assume that the reservation utility $\bar{u} = 10$, which guarantees that all utility levels are positive.[31] To keep the example as simple as possible, it is assumed that the agent's performance can take only three values, i.e., the agent's performance is either excellent (E), satisfactory (S) or inadequate (I). We consider two specifications of the performance measure. In the first specification the satisfactory signal is a good signal, whereas in the second specification it is a bad signal. Formally, in the first specification the conditional probabilities take the following values:

$$\begin{aligned} \gamma_E^H &= 5/10 & \gamma_E^L &= 1/10 \\ \gamma_S^H &= 4/10 & \gamma_S^L &= 3/10 \\ \gamma_I^H &= 1/10 & \gamma_I^L &= 6/10\ . \end{aligned}$$

The structure of the optimal contract for this specification and various values of r and λ is presented in Table 1.

r \ λ	1.0	1.1	1.3	1.5
1.5	$u_1 < u_2 < u_3$	$u_1 < u_2 = u_3$	$u_1 < u_2 = u_3$	$u_1 < u_2 = u_3$
2	$u_1 < u_2 < u_3$	$u_1 < u_2 < u_3$	$u_1 < u_2 = u_3$	$u_1 < u_2 = u_3$
3	$u_1 < u_2 < u_3$	$u_1 < u_2 < u_3$	$u_1 < u_2 = u_3$	$u_1 < u_2 = u_3$

Table 1: Structure of the optimal contract with two "good" signals.

Table 1 suggests that the optimal contract typically involves pooling of the two good signals, in particular when the agent's intrinsic utility is not too concave, i.e., if the agent is not too risk averse. Table 1 nicely illustrates the trade-off the principal faces when the agent is both, risk and loss averse: If the agent becomes more risk averse pooling is less

[31] Increasing $\bar{u}$ makes the agent less risk averse and thus is similar to a reduction in r.

likely to be optimal. If, on the other hand, he becomes more loss averse, pooling is more likely to be optimal.[32]

In the second specification we assume that there are two bad signals. The conditional probabilities are as follows:

$$\begin{aligned} \gamma_E^H &= 6/10 & \gamma_E^L &= 1/10 \\ \gamma_S^H &= 2/10 & \gamma_S^L &= 4/10 \\ \gamma_I^H &= 2/10 & \gamma_I^L &= 5/10\ . \end{aligned}$$

The results for this case are presented in Table 2.

r \ λ	1.0	1.1	1.3	1.5
1.5	$u_1 < u_2 < u_3$	$u_1 = u_2 < u_3$	$u_1 = u_2 < u_3$	$u_1 = u_2 < u_3$
2	$u_1 < u_2 < u_3$	$u_1 = u_2 < u_3$	$u_1 = u_2 < u_3$	$u_1 = u_2 < u_3$
3	$u_1 < u_2 < u_3$	$u_1 = u_2 < u_3$	$u_1 = u_2 < u_3$	$u_1 = u_2 < u_3$

Table 2: Structure of the optimal contract with two "bad" signals.

In this specification, a binary statistic that pools the two bad signals seems to be optimal almost always. The reason behind this observation is that the two bad signals are very similar. In consequence, paying the same wage for satisfactory as well as inadequate performance increases the risk premium only slightly. On the other hand, by pooling satisfactory and inadequate performance it becomes less likely for the agent ex-ante to experience a loss, i.e., the loss premium is reduced. Therefore, it is optimal for the principal to use a bonus scheme even when the agent's degree of loss aversion is small.

4 Implementation Problems, Turning a Blind Eye, and Stochastic Contracts

In this section we do not impose assumptions that guarantee the validity of the first-order approach. In particular, in order to explore the implications of a higher degree of loss aversion, we relax $(A2)$. We restrict attention to two simplifications of the former model. First, we return to the assumption of a purely loss averse agent. Second, only binary measures of performance are considered. This latter assumption seems natural in the light of the previous section: there it was shown that, when intrinsic utility is a linear function and the agent's degree of loss aversion is not too high, it is optimal for the principal to construct a binary measure of performance by offering a bonus contract.

[32] For a given r, the degree of pooling does not monotonically increase in λ. As discussed at the end of Section 3.1, a higher degree of loss aversion of the agent may help the principal to create incentives. If this is the case, a contract that contains less pooling is preferred from an incentive point of view. If this positive effect of less pooling on incentives outweighs the negative effect on the agent's loss premium, then the optimal contract consists of more distinct wage payments when λ increases. This can, however happen only locally, that is, at some point the degree of pooling increases in λ.

4.1 The Case of a Binary Measure of Performance

As before, the principal cannot observe the agent's action a or whether the benefit B was realized or not. Instead she observes a contractible binary measure of performance, i.e., $\mathcal{S} = \{1, 2\}$. For notational convenience , let $(1 - \gamma^H)$ and γ^H denote the probabilities of observing signal $s = 1$ and $s = 2$, respectively, conditional on B being realized. Accordingly, $(1 - \gamma^L)$ and γ^L are the probabilities of observing signal $s = 1$ and $s = 2$, respectively, conditional on B not being realized.[33] Thus, the unconditional probability of observing signal $s = 2$ for a given action a is $\gamma(a) \equiv a\gamma^H + (1-a)\gamma^L$. Let $\hat{\boldsymbol{\gamma}} = (\gamma^H, \gamma^L)$. We reformulate $(A1)$ for the binary case as follows.

Assumption (A4): $1 > \gamma^H > \gamma^L > 0$.

With only two possible signals to be observed, the contract takes the form of a bonus contract: the agent is paid a base wage which yields intrinsic utility u if the bad signal is observed, and he is paid the base wage plus a bonus b resulting in intrinsic utility $u + b$ if the good signal is observed. For now assume that $b \geq 0$.[34] For expositional purposes we assume that the agent's intrinsic disutility of effort is a quadratic function, $c(a) = (k/2)a^2$.[35] The agent's expected utility from choosing effort level a then is

$$E\left[U(a)\right] = u + \gamma(a)b - \frac{k}{2}a^2 - (\lambda - 1)\gamma(a)(1 - \gamma(a))b. \tag{9}$$

As before, the first component is expected net intrinsic utility from choosing effort level a, that is, expected wage payment minus effort cost. The second component is the loss premium, with $\gamma(a)(1 - \gamma(a))$ denoting the loss probability.

4.2 Invalidity of the First-Order Approach

The first derivative of expected utility with respect to effort is given by

$$E\left[U'(a)\right] = \underbrace{(\gamma^H - \gamma^L)b\left[2 - \lambda + 2\gamma(a)(\lambda - 1)\right]}_{MB(a)} - \underbrace{ka}_{MC(a)}. \tag{10}$$

While the marginal cost, $MC(a)$, obviously is a straight line through the origin with slope k, the marginal benefit, $MB(a)$, also is a positively sloped, linear function of effort a. An increase in b unambiguously makes $MB(a)$ steeper. Letting a_0 denote the intercept of $MB(a)$ with the horizontal axis, we have

$$a_0 = \frac{\lambda - 2 - 2\gamma^L(\lambda - 1)}{2(\gamma^H - \gamma^L)(\lambda - 1)}.$$

The cases for $a_0 < 0$ and $a_0 > 0$ are depicted in Figures 1 and 2, respectively. Implementation problems in our sense refer to a situation where there are actions $a \in (0, 1)$ that are not incentive compatible for any bonus payment.

[33] In the notation introduced above, we have $\gamma_1^H = 1 - \gamma^H$, $\gamma_2^H = \gamma^H$, $\gamma_1^L = 1 - \gamma^L$ and $\gamma_2^L = \gamma^L$.

[34] The assumption $b \geq 0$ is made only for expositional purposes, the results hold true for $b \in \mathbb{R}$.

[35] This functional form does not fit exactly the assumptions on $c(\cdot)$ that we imposed above, but is made for expositional convenience. Allowing for more general effort cost functions does not qualitatively change the insights that are to be obtained.

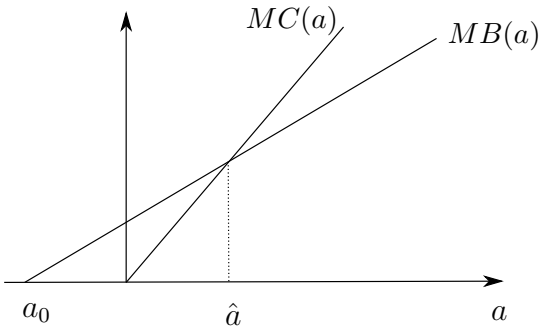

Figure 1: $MB(a)$ and $MC(a)$ for $a_0 < 0$.

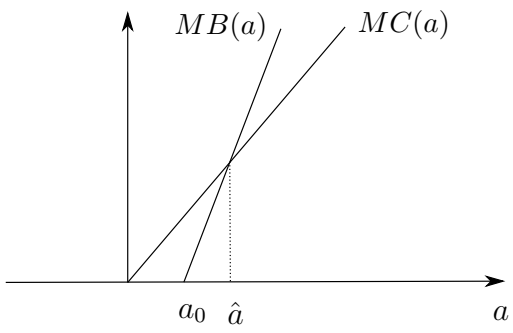

Figure 2: $MB(a)$ and $MC(a)$ for $a_0 > 0$.

Proposition 7: *Given* $(A4)$*, effort level* $\hat{a} \in (0,1)$ *is implementable if and only if* $a_0 \leq 0$.

Proof: See Appendix.

Implementation problems arise when $a_0 > 0$, or equivalently, when $\gamma^L < 1/2$ and $\lambda > 2(1-\gamma^L)/(1-2\gamma^L) > 2$. Somewhat surprisingly, this includes performance measures with $\gamma^L < 1/2 < \gamma^H$, which (possibly) are highly informative. Informative in this context means that it is more likely to observe the bad signal if benefit B was not realized , whereas it is more likely to observe the good signal if B was realized. So, why do these implementation problems arise in the first place? Remember that the agent has two targets: First, as in classic models, he seeks to maximize net intrinsic utility, $u + b\gamma(a) - (k/2)a^2$. When the agent cares only about this net intrinsic utility (e.g., he is loss neutral) then each action can be implemented by choosing a sufficiently high bonus. Due to loss aversion, however, the agent has a second target which is minimizing the expected loss. How can the agent pursue this goal? He can do so by choosing an action such that the loss probability, $\gamma(a)(1-\gamma(a))$, becomes small. The crucial point is that these two targets may conflict with each other in the sense that an increase in effort may increase net intrinsic utility but at the same time also increases the loss probability. First of all, note that implementation problems never arise when $\gamma^L \geq 1/2$ or $\lambda \leq 2$. For $\gamma^L \geq 1/2$, the loss probability is strictly decreasing in the agent's action. Consequently, with both targets of the agent being aligned, an increase in the bonus unambiguously leads to an increase in the agent's action. For $\lambda \leq 2$, the weight put on gain-loss utility, $\lambda - 1$, is lower than the weight put on intrinsic utility, so the agent is more interested in maximizing net intrinsic utility than in minimizing the loss probability. With loss aversion being not that important, an increase in the bonus therefore always leads to an increase in effort, irrespective of whether the loss probability locally increases or decreases in the agent's action. For $\gamma^L < 1/2$, on the other hand, implementation problems do arise when λ is sufficiently large. Roughly speaking, being sufficiently loss averse, the agent primarily cares about reducing the loss probability. With the loss probability being inverted U-shaped, the agent achieves this by choosing one of the two extreme actions $a \in \{0,1\}$. Therefore, the principal cannot motivate the agent to choose an action $\hat{a} \in (0,1)$ when $\gamma^L < 1/2$ and the agent's loss aversion is sufficiently severe.

4.3 Turning a Blind Eye

As we have seen in the preceding analysis, the principal faces implementation problems whenever $a_0 > 0$. One might wonder if there is a remedy for these implementation problems. The answer is "yes", there is a remedy, and in fact a surprisingly simple one. The principal can manipulate the signal in her favor by not paying attention to the signal from time to time but nevertheless paying the bonus in these cases. Formally, suppose the principal commits herself to stochastically ignoring the signal with probability $p \in [0, 1)$.[36] Thus, the overall probability of receiving the bonus is given by $\gamma(a; p) \equiv p + (1 - p)\gamma(a)$. This strategic ignorance of information gives rise to a transformed performance measure $\hat{\boldsymbol{\gamma}}(p) = (\gamma^H(p), \gamma^L(p))$. As before, $\gamma^H(p)$ denotes the probability that the bonus is paid to the agent conditional on benefit B being realized. Given that B is realized, this happens either when the performance measure is ignored, or - if the principal pays attention to the performance measure - when the good signal is realized. Hence, $\gamma^H(p) = p + (1 - p)\gamma^H$. Analogously, the probability of the bonus being paid out conditional on B not being realized is given by $\gamma^L(p) = p+(1-p)\gamma^L$. As it turns out, ignoring the whole performance measure with probability p is formally equivalent to ignoring only the bad signal with probability p.[37] For this reason, we refer to the principal not paying attention to the performance measure as turning a blind eye on bad performance of the agent. It is readily verified that under the transformed performance measure $\hat{\boldsymbol{\gamma}}(p)$ the intercept of the $MB(a)$ function with the horizontal axis,

$$a_0(p) \equiv \frac{\lambda - 2 - 2\left[p + (1 - p)\gamma^L\right](\lambda - 1)}{2(1 - p)(\gamma^H - \gamma^L)(\lambda - 1)},$$

not only is decreasing in p but also can be made arbitrarily small, in particular, arbitrarily negative. Formally, $da_0(p)/dp < 0$ and $\lim_{p\to 1} a_0(p) = -\infty$. In the light of Proposition 7 this immediately implies that the principal can eliminate any implementation problems by choosing p sufficiently high, that is, by turning a blind eye sufficiently often.

Besides alleviating possible implementation problems, turning a blind eye on the bad signal can also benefit the principal from a cost perspective. Using the definition of $\gamma(a; p)$ it can be shown that the minimum cost of implementation of action $\hat{a}$ under the transformed performance measure, $C(\hat{a}; p)$, takes the following form:

$$C(\hat{a}; p) = \overline{u} + \frac{k}{2}\hat{a}^2 + \frac{k\hat{a}(\lambda - 1)(1 - \gamma(\hat{a}))}{(\gamma^H - \gamma^L)} \frac{\gamma(\hat{a}) + p(1 - \gamma(\hat{a}))}{1 - (\lambda - 1)\left[1 - 2\gamma(\hat{a}) - 2p(1 - \gamma(\hat{a}))\right]} \quad (11)$$

Differentiating the principal's cost with respect to p reveals that $\text{sign}\{dC(\hat{a}; p)/dp\} = \text{sign}\{2 - \lambda\}$. Hence, an increase in the probability of ignoring the bad signal decreases the cost of implementing a certain action if and only if $\lambda > 2$. Hence, whenever the principal

[36] Always ignoring the signal, i.e., setting $p = 1$, would be detrimental for incentives because then the agent's monetary payoff is independent of his action. Hence, he would choose the least cost action $a = 0$. Therefore, we a priori restrict the principal to choose p from the interval $[0, 1)$.

[37] In this latter case, the agent receives the bonus either when the good signal is observed, which happens with probability $\gamma(a)$, or when the bad signal is observed but is ignored, which happens with probability $(1-\gamma(a))p$. Hence, the overall probability of the bonus being paid out is given by $\gamma(a)+(1-\gamma(a))p$.

turns a blind eye in order to remedy implementation problems, he will do so to the largest possible extent.[38,39] We summarize the preceding analysis in the following proposition.

Proposition 8: *Suppose the principal can commit herself to stochastic ignorance of the signal. Then each action $\hat{a} \in [0,1]$ can be implemented. Moreover, the implementation costs are strictly decreasing in p if and only if $\lambda > 2$.*

Proof: See Appendix.

We restricted the principal to offer non-stochastic payments conditional on which signal is observed. If the principal was able to do just that, then he could remedy implementation problems by paying the base wage plus a lottery in the case of the bad signal. For instance, when the lottery yields b with probability p and zero otherwise, this is just the same as turning a blind eye.[40] This observation suggests that the principal may benefit from offering a contract that includes randomization, which is in contrast to the finding under conventional risk aversion in Holmström (1979).[41]

4.4 Blackwell Revisited

We conclude this section by briefly pointing out an interesting implication of the above analysis. Suppose the principal has no access to a randomization device, i.e., turning a blind eye is not possible. Then the above considerations allow a straight-forward comparison of performance measures $\hat{\boldsymbol{\zeta}} = (\zeta^H, \zeta^L)$ and $\hat{\boldsymbol{\gamma}} = (\gamma^H, \gamma^L)$ if $\hat{\boldsymbol{\zeta}}$ is a convex combination of $\hat{\boldsymbol{\gamma}}$ and $\mathbf{1} \equiv (1,1)$.

Corollary 1: *Let $\hat{\boldsymbol{\zeta}} = p\mathbf{1} + (1-p)\hat{\boldsymbol{\gamma}}$ with $p \in (0,1)$. Then the principal at least weakly prefers performance measure $\hat{\boldsymbol{\zeta}}$ to $\hat{\boldsymbol{\gamma}}$ if and only if $\lambda \geq 2$.*

Proof: See Appendix.

The finding that the principal prefers the "garbled" performance measure $\hat{\boldsymbol{\zeta}}$ over performance measure $\hat{\boldsymbol{\gamma}}$ is at odds with Blackwell's theorem. To see this, let performance measures $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\zeta}}$ be characterized, respectively, by the stochastic matrices

$$\boldsymbol{P}_\gamma = \begin{pmatrix} 1-\gamma^H & \gamma^H \\ 1-\gamma^L & \gamma^L \end{pmatrix} \quad \text{and} \quad \boldsymbol{P}_\zeta = \begin{pmatrix} 1-\zeta^H & \zeta^H \\ 1-\zeta^L & \zeta^L \end{pmatrix}.$$

[38] Formally, for $\lambda > 2$, the solution to the principal's problem of choosing the optimal probability to turn a blind eye, p^*, is not well defined because $p^* \to 1$. If the agent is subject to limited liability or there is a cost of ignorance, however, the optimal probability of turning a blind eye is well defined.

[39] This is in the spirit of Becker and Stigler (1974), who show that despite a small detection probability of malfeasance, incentives can be maintained if the punishment is sufficiently severe.

[40] In this case, the agent receives the bonus when the good signal is observed, which happens with probability $\gamma(a)$, or when the bad signal is observed and the realization of the lottery is b, which happens with probability $(1-\gamma(a))p$. Hence, the overall probability of the bonus being paid out is given by $\gamma(a) + (1-\gamma(a))p$, which is nothing but $\gamma(a;p)$ from turning a blind eye on the bad signal.

[41] The finding that stochastic contracts may be optimal is not novel to the principal-agent literature. Haller (1985) shows that in the case of a satisficing agent, who wants to achieve certain aspiration levels of income with certain probabilities, randomization may pay for the principal. Moreover, Strausz (2006) finds that deterministic contracts may be suboptimal in a screening context.

According to Blackwell's theorem, any decision maker prefers information system $\hat{\boldsymbol{\gamma}}$ to $\hat{\boldsymbol{\zeta}}$ if and only if there exists a non-negative stochastic matrix $\boldsymbol{M}$ with $\sum_j m_{ij} = 1$ such that $\boldsymbol{P}_\zeta = \boldsymbol{P}_\gamma \boldsymbol{M}$.[42] It is readily verified that this matrix $\boldsymbol{M}$ exists and takes the form

$$\boldsymbol{M} = \begin{pmatrix} 1-p & p \\ 0 & 1 \end{pmatrix}.$$

Thus, even though comparison of the two performance measures according to Blackwell's theorem implies that the principal should prefer $\hat{\boldsymbol{\gamma}}$ over $\hat{\boldsymbol{\zeta}}$, the principal actually prefers the "garbled" information system $\hat{\boldsymbol{\zeta}}$ over information system $\hat{\boldsymbol{\gamma}}$. While Kim (1995) has already shown that the necessary part of Blackwell's theorem does not hold in the agency model, the sufficiency part was proven to be applicable to the agency framework by Gjesdal (1982).[43]Our findings, however, show that this is not the case anymore when the agent is loss averse.

5 Robustness, Extensions and Concluding Remarks

In this paper we explore the implications of reference-dependent preferences on contract design in an otherwise standard model of principal-agency. We find that introducing loss aversion on the agent's side leads to a reduction in the complexity of the optimal contractual arrangement. In the extreme case of a purely loss averse agent, the optimal contract takes the form of a simple bonus contract: some realizations of the performance measure are rewarded with a bonus payment, while others are not. Thus, loss aversion provides a theoretical rationale for bonus contracts, the wide application of which is hard to reconcile with obvious drawbacks – as seasonality effects or insurance fraud – that come along with this particular contractual form.

In the rest of the section we consider the robustness of our results. After a brief and semi-formal analysis of an alternative equilibrium concept, we explore the consequences of nonquadratic effort costs for implementation problems. Finally, we conclude by discussing diminishing sensitivity of the gain-loss function. Throughout the whole analysis we adopted the concept of choice-acclimating personal equilibrium (CPE). As already pointed out in Section 2, for higher degrees of loss aversion this concept has the questionable property that a decision maker may prefer stochastically dominated options. Kőszegi and Rabin (2006, 2007) provide another concept, called unacclimating personal equilibrium (UPE), under which such behavior cannot occur. The major difference between UPE and CPE is the timing of expectation formation and actual decision making. Under UPE a decision maker first forms his expectations, which determine his reference point, and thereafter, given these expectations, chooses the optimal action. To rule out that people can systematically cheat themselves, for action $\hat{a}$ to be an UPE, it must be optimal for

[42]See Blackwell (1951, 1953).

[43]In order to avoid confusion: The necessary part of Blackwell's theorem states that the principal being better off implies that she uses a more informative performance measure. The sufficiency part conversely states that making use of more informative performance measure implies that the principal is better off.

the decision maker to choose $\hat{a}$ given that he expected to do so. In the following we will argue that applying UPE instead of CPE does not change our main findings. For the sake of argumentation, consider the case of a purely loss averse agent, i.e., intrinsic utility is linear. The agent's ex-ante expected utility from choosing action a when expecting action $\hat{a}$ is

$$E[U(a|\hat{a})] = \sum_{s=1}^{S} \gamma_s(a) \left[u_s + \sum_{j=1}^{s-1} \gamma_j(\hat{a})(u_s - u_j) - \lambda \sum_{t=s+1}^{S} \gamma_t(\hat{a})(u_t - u_s) \right] - c(a) + \mu(c(\hat{a}) - c(a)) \, .$$

On the equilibrium path expectation and actual action coincide. Therefore, the agent's ex-ante expected utility, and in consequence the individual rationality constraint, takes the same form under both equilibrium concepts, CPE and UPE. The incentive compatibility constraint, on the other hand, depends on the applied equilibrium concept. Given the agent expected to choose $\hat{a}$, his marginal utility from choosing a is

$$E[U'(a|\hat{a})] = \sum_{s=1}^{S} (\gamma_s^H - \gamma_s^L) u_s + \sum_{s=1}^{S} \sum_{j=1}^{s-1} \gamma_j(\hat{a})(\gamma_s^H - \gamma_s^L)(u_s - u_j) - \lambda \sum_{s=1}^{S} \sum_{j=s+1}^{S} \gamma_j(\hat{a})(\gamma_s^H - \gamma_s^L)(u_j - u_s) - c'(a) + \mu'(c(\hat{a}) - c(a)) \, .$$

Note that either $\mu'(\cdot) = 1$ or $\mu'(\cdot) = \lambda$, depending on whether $\hat{a}$ is greater or lower than a. Even though $E[U(a|\hat{a})]$ is a strictly concave function in the agent's actual action choice a for all values of $\lambda \geq 1$, under UPE there arises the problem of multiplicity of equilibria. More precisely, for a given incentive scheme $\boldsymbol{u}$, there exists a range of actions $a \in [\underline{a}(\boldsymbol{u}), \bar{a}(\boldsymbol{u})]$ all of which constitute a UPE. This problem can easily be circumvented by assuming that the agent chooses the highest action which constitutes a UPE. In this case, there is no need to impose additional assumptions on the cost function or to assume that λ is sufficiently small.[44] By imposing this alternative assumption the incentive compatibility constraint can be rewritten as

$$\sum_{s=2}^{S} b_s \left\{ \left(\sum_{t=s}^{S} (\gamma_t^H - \gamma_t^L) \right) \left(1 + \sum_{j=1}^{s-1} \gamma_j(\hat{a}) \right) - \lambda \left(\sum_{t=s}^{S} \gamma_t(\hat{a}) \right) \left(\sum_{j=1}^{s-1} (\gamma_j^H - \gamma_j^L) \right) \right\} = 2c'(\hat{a}) \, .$$

Clearly, the incentive compatibility constraint is a linear constraint in the bonus payments $\boldsymbol{b} = (b_2, \ldots, b_S)$. Thus, our bonus contract result is robust with respect to this change of Assumptions $(A2)$ and $(A3)$.

There is another way to resolve the multiplicity problem under UPE. Kőszegi and Rabin (2006, 2007) define a preferred personal equilibrium (PPE) as a decision maker's

[44] For given expectations $\hat{a}$, let EU_g and EU_l denote the agent's expected utility given that $\mu(x) = x$ and $\mu(x) = \lambda x$, respectively. Both EU_g and EU_l are strictly concave functions, with EU_g achieving its maximum at a strictly higher action than EU_l. EU_g and EU_l intersect at $\hat{a}$. Action $\hat{a}$ is an UPE if it lies between the maximizing actions of EU_g and EU_l. Therefore, expecting to choose the action which maximizes EU_g not only constitutes an UPE, but also is the highest possible UPE.

ex-ante favorite plan among those plans he actually will follow through. Put differently, given incentive scheme $\boldsymbol{u}$, the agent chooses the action $a^{PPE} \in [\underline{a}(\boldsymbol{u}), \bar{a}(\boldsymbol{u})]$ that maximizes expected utility among those actions that constitute a UPE. If for all incentive-compatible incentive schemes we have $a^{PPE} \in (\underline{a}(\boldsymbol{u}), \bar{a}(\boldsymbol{u}))$ then PPE and CPE coincide, i.e., a^{PPE} is determined by the first-order condition that characterizes the agent's action under CPE. Thus, by imposing the PPE-analogue of $(A2)$ and $(A3)$ we can derive results identical to those under CPE. If $a^{PPE} \in \{\underline{a}(\boldsymbol{u}), \bar{a}(\boldsymbol{u})\}$ for all incentive-compatible incentive schemes, the optimal contract also is a bonus contract since both boundary actions are determined by functions linear in $\boldsymbol{b} = (b_2, \ldots, b_S)$.[45] In the intermediate case, however, where $a^{PPE} \in (\underline{a}(\boldsymbol{u}), \bar{a}(\boldsymbol{u}))$ for some incentive-compatible incentive schemes but $a^{PPE} \in \{\underline{a}(\boldsymbol{u}), \bar{a}(\boldsymbol{u})\}$ for others, the optimal contract is not necessarily a bonus scheme.

If the agent's action is characterized by PPE, for all actions $\hat{a} \in (0, 1)$ to be implementable we still need the assumption that λ is not too high. Put differently, implementation problems as discussed in Section 4 also arise under PPE. Compared to CPE, however, these implementation problems are less severe. For instance, actions close to zero are always implementable under PPE.

For the discussion of implementation problems in Section 4, we restricted attention to quadratic effort costs. The finding that implementation problems are an important issue, however, holds true for a wide variety of cost functions. Depending on the particular functional form of the corresponding marginal costs, these implementation problems may be more or less severe. For instance, the result that there are implementation problems if $a_0 > 0$ holds true for all strictly increasing and strictly convex cost functions with $c'(0) = 0$. As for strictly concave marginal costs with $c'(0) = 0$, no action $\hat{a} \in (0, 1)$ is implementable if $a_0 \geq 0$; and even for $a_0 < 0$ there may be actions, in particular actions close to 1, that are not implementable.

Moreover, we kept the whole analysis simple by ignoring diminishing sensitivity, that is, by considering a piece-wise linear gain-loss function. A more general gain-loss function makes the analysis by far more complicated: Both the incentive compatibility constraint and the individual rationality constraint are no longer linear functions in the intrinsic utility levels, and thus the Kuhn-Tucker conditions are not necessarily sufficient. Nevertheless, we expect that a reduction in the complexity of the contract may benefit the principal in this case as well. Diminishing sensitivity of the agent's utility implies that the sum of two net losses of two monetary outcomes exceeds the net loss of the sum of these two monetary outcomes. Therefore, in addition to the effects discussed in the paper, under diminishing sensitivity there is another channel through which melting two bonus payments into one "big" bonus affects, and in tendency reduces, the agent's expected net loss. There is, however, an argument running counter to this intuition. As we have shown, loss aversion may help the principal to create incentives. Therefore, setting many different wage payments, and thereby – in a sense – creating many kinks, proximity to which the

[45] The case of $a^{PPE} = \bar{a}(\boldsymbol{u})$ corresponds to the alternative assumption to $(A2)$ discussed above. If $a^{PPE} = \underline{a}(\boldsymbol{u})$, on the other hand, then a^{PPE} maximizes EU_l, as defined in the previous footnote.

agent strongly dislikes under diminishing sensitivity, may have favorable incentive effects. Exploring the effects of diminishing sensitivity in a principal-agent relationship with moral hazard is therefore an open question for future research.

A Appendix: Proofs of Propositions and Lemmas

Proof of Lemma 1

Suppose that signals are ordered according to their likelihood ratio, that is, $s > s'$ if and only if $\gamma_s^H/\gamma_s^L > \gamma_{s'}^H/\gamma_{s'}^L$. Consider a contract of the form

$$u_s = \begin{cases} \underline{u} & \text{if} \quad s < \hat{s} \\ \underline{u} + b & \text{if} \quad s \geq \hat{s} \end{cases},$$

where $b > 0$ and $1 < \hat{s} \leq S$. Under this contractual form and given that the first-order approach is valid, (IC) can be rewritten as

$$b\Bigg\{ \left[\sum_{s=\hat{s}}^{S}(\gamma_s^H - \gamma_s^L)\right] \left(1 - (\lambda-1)\sum_{s=1}^{\hat{s}-1}\gamma_s(\hat{a})\right) - (\lambda-1)\left(\sum_{s=1}^{\hat{s}-1}(\gamma_s^H - \gamma^L)\right)\left(\sum_{s=\hat{s}}^{S}\gamma_s(\hat{a})\right)\Bigg\} = c'(\hat{a}).$$

Since signals are ordered according to their likelihood ratio, we have $\sum_{s=\hat{s}}^{S}(\gamma_s^H - \gamma_s^L) > 0$ and $\sum_{s=1}^{\hat{s}-1}(\gamma_s^H - \gamma^L) < 0$ for all $1 < \hat{s} \leq S$. This implies that the term in curly brackets is strictly positive for $\lambda \leq 2$. Hence, with $c'(\hat{a}) > 0$, b can alway be chosen such that (IC) is met. Rearranging the participation constraint,

$$\underline{u} \geq \bar{u} + c(\hat{a}) - b\left(\sum_{s=\hat{s}}^{S}\gamma_s(\hat{a})\right)\left[1 - (\lambda-1)\left(\sum_{s=1}^{\hat{s}-1}\gamma_s(\hat{a})\right)\right],$$

reveals that (IR) can be satisfied for any b by choosing $\underline{u}$ appropriately. This concludes the proof. ■

Proof of Proposition 1

It is readily verified that Assumptions 1-3 from Grossman and Hart (1983) are satisfied. Thus, the cost-minimization problem is well defined, in the sense that for each action $a \in (0,1)$ there exists a second-best incentive scheme. Suppose the principal wants to implement action $\hat{a} \in (0,1)$ at minimum cost. Since the agent's action is not observable, the principal's problem is given by

$$\min_{\{u_s\}_{s=1}^{S}} \sum_{s=1}^{S}\gamma_s(\hat{a})h(u_s) \tag{MR}$$

subject to

$$\sum_{s=1}^{S} \gamma_s(\hat{a}) u_s - c(\hat{a}) \geq \bar{u}\,, \tag{IR$_R$}$$

$$\sum_{s=1}^{S} (\gamma_s^H - \gamma_s^L) u_s - c'(\hat{a}) = 0\,. \tag{IC$_R$}$$

where the first constraint is the individual rationality constraint and the second is the incentive compatibility constraint. Note that the first-order approach is valid, since the agent's expected utility is a strictly concave function of his effort. The Lagrangian to the resulting problem is

$$\mathcal{L} = \sum_{s=1}^{S} \gamma_s(a) h(u_s) - \mu_0 \left\{ \sum_{s=1}^{S} \gamma_s(a) u_s - c(a) - \bar{u} \right\} - \mu_1 \left\{ \sum_{s=1}^{S} (\gamma_s^H - \gamma_s^L) u_s - c'(a) \right\},$$

where μ_0 and μ_1 denote the Lagrange multipliers of the individual rationality constraint and the incentive compatibility constraint, respectively. Setting the partial derivative of $\mathcal{L}$ with respect to w_s equal to zero yields

$$\frac{\partial \mathcal{L}}{\partial u_s} = 0 \iff h'(u_s) = \mu_0 + \mu_1 \frac{\gamma_s^H - \gamma_s^L}{\gamma_s(\hat{a})}, \quad \forall s \in \mathcal{S}. \tag{A.1}$$

Irrespective of the value of μ_0, if $\mu_1 > 0$, convexity of $h(\cdot)$ implies that $u_s > u_{s'}$ if and only if $(\gamma_s^H - \gamma_s^L)/\gamma_s(\hat{a}) > (\gamma_{s'}^H - \gamma_{s'}^L)/\gamma_{s'}(\hat{a})$, which in turn is equivalent to $\gamma_s^H/\gamma_s^L > \gamma_{s'}^H/\gamma_{s'}^L$. Thus it remains to show that μ_1 is strictly positive. Suppose, in contradiction, that $\mu_1 \leq 0$. Consider the case $\mu_1 = 0$ first. From (A.1) it follows that $u_s = u^f$ for all $s \in \mathcal{S}$, where u^f satisfies $h'(u^f) = \mu_0$. This, however, violates (IC$_R$), a contradiction. Next, consider $\mu_1 < 0$. From (A.1) it follows that $u_s < u_{s'}$ if and only if $(\gamma_s^H - \gamma_s^L)/\gamma_s(\hat{a}) > (\gamma_{s'}^H - \gamma_{s'}^L)/\gamma_t(\hat{a})$. Let $\mathcal{S}^+ \equiv \{s | \gamma_s^H - \gamma_s^L > 0\}$, $\mathcal{S}^- \equiv \{s | \gamma_s^H - \gamma_s^L < 0\}$, and $\hat{u} \equiv \min\{u_s | s \in \mathcal{S}^-\}$. Since $\hat{u} > u_s$ for all $s \in \mathcal{S}^+$, we have

$$\begin{aligned}
\sum_{s=1}^{S} (\gamma_s^H - \gamma_s^L) u_s &= \sum_{\mathcal{S}^-} (\gamma_s^H - \gamma_s^L) u_s + \sum_{\mathcal{S}^+} (\gamma_s^H - \gamma_s^L) u_s \\
&< \sum_{\mathcal{S}^-} (\gamma_s^H - \gamma_s^L) \hat{u} + \sum_{\mathcal{S}^+} (\gamma_s^H - \gamma_s^L) \hat{u} \\
&= \hat{u} \sum_{s=1}^{S} (\gamma_s^H - \gamma_s^L) \\
&= 0,
\end{aligned}$$

again a contradiction to (IC$_R$). Hence, $\mu_1 > 0$ and the desired result follows. ■

Proof of Proposition 2

The problem of finding the optimal contract $\boldsymbol{u}^*$ to implement action $\hat{a} \in (0, 1)$ is decomposed into two subproblems. First, for a given incentive feasible ordering of signals,

we derive the optimal nondecreasing incentive scheme that implements action $\hat{a} \in (0,1)$. Then, in a second step, we choose the ordering of signals for which the ordering specific cost of implementation is lowest.

Step 1: Remember that the ordering of signals is incentive feasible if $\beta_s(\cdot) > 0$ for at least one signal s. For a given incentive feasible ordering of signals, in this first step we solve Program ML. First, note that it is optimal to set $b_s = 0$ if $\beta_s(\cdot) < 0$. To see this, suppose, in contradiction, that in the optimum (IC′) holds and $b_s > 0$ for some signal s with $\beta_s(\cdot) \leq 0$. If $\beta_s(\cdot) = 0$, then setting $b_s = 0$ leaves (IC′) unchanged, but leads to a lower value of the objective function of Program ML, contradicting that the original contract is optimal. If $\beta_s(\cdot) < 0$, then setting $b_s = 0$ not only reduces the value of the objective function, but also relaxes (IC′), which in turn allows to lower other bonus payments, thereby lowering the value of the objective function even further. Again, a contradiction to the original contract being optimal. Let $\mathcal{S}_\beta \equiv \{s \in \mathcal{S} | \beta_s(\cdot) > 0\}$ denote the set of signals for which $\beta_s(\cdot)$ is strictly positive under the considered ordering of signals, and let S_β denote the number of elements in this set. Thus, Program (ML) can be rewritten as

Program ML$^+$:

$$\begin{aligned} & \min_{\{b_s\}_{s\in\mathcal{S}_\beta}} \sum_{s\in\mathcal{S}_\beta} b_s \rho_s(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a}) \\ \text{subject to} \quad & (i) \ \sum_{s\in\mathcal{S}_\beta} b_s \beta_s(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a}) = c'(\hat{a}) \qquad (\text{IC}^+) \\ & (ii) \ \ b_s \geq 0, \ \ \forall s \in \mathcal{S}_\beta \, . \end{aligned}$$

Program ML$^+$ is a linear programming problem. It is well-known that if a linear programming problem has a solution, it must have a solution at an extreme point of the constraint set. Generically, there is a unique solution and this solution is an extreme point. Since the constraint set of Program ML$^+$, $\mathcal{M} \equiv \{\{b_s\}_{s\in\mathcal{S}_\beta} \in \mathbb{R}_+^{S_\beta} | \sum_{s\in\mathcal{S}_\beta} b_s \beta_s(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a}) = c'(\hat{a})\}$, is closed and bounded, Program ML$^+$ has a solution. Hence $\sum_{s\in\mathcal{S}_\beta} b_s \rho_s(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a})$ achieves its greatest lower bound at one of the extreme points of $\mathcal{M}$. With $\mathcal{M}$ describing a hyperplane in $\mathbb{R}_+^{S_\beta}$, all extreme points of $\mathcal{M}$ are characterized by the following property: $b_s > 0$ for exactly one signal $s \in S_\beta$ and $b_t = 0$ for all $t \in S_\beta$, $t \neq s$. It remains to determine for which signal the bonus is set strictly positive. The size of the bonus payment, which is set strictly positive, is uniquely determined by (IC$^+$):

$$b_s \beta_s(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a}) = c'(\hat{a}) \iff b_s = \frac{c'(\hat{a})}{\beta_s(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a})}. \qquad (A.2)$$

Therefore, from the objective function of Program ML$^+$ it follows that, for the signal ordering under consideration, the optimal signal for which the bonus is set strictly positive, $\hat{s}$, is characterized by

$$\hat{s} = \arg\min_{s\in\mathcal{S}_\beta} \frac{c'(\hat{a})}{\beta_s(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a})} \rho_s(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a}).$$

Step 2: From all incentive feasible signal orders, the principal chooses the one which minimizes her cost of implementation. With the number of incentive feasible signal orders being finite, this problem clearly has a solution. Let s^* denote the resulting cutoff, i.e.,

$$u_s^* = \begin{cases} u^* & \text{if} \quad s < s^* \\ u^* + b^* & \text{if} \quad s \geq s^* \end{cases} ,$$

where $b^* = c'(\hat{a})/\beta_{s^*}(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a})$ and $u^* = \bar{u} + c(\hat{a}) - b^* \left[\sum_{\tau=s^*}^{S} \gamma_\tau(\hat{a}) - \rho_{s^*}(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a})\right]$. Letting $u_L^* = u^*$, $u_H^* = u^* + b^*$, and $\mathcal{B}^* = \{s \in \mathcal{S} | s \geq s^*\}$ establishes the desired result. ■

Proof of Proposition 3

$\mathcal{B}^*$ maximizes $X(\mathcal{B}) := \left[\sum_{s\in\mathcal{B}}(\gamma_s^H - \gamma_s^L)\right] \times Y(P_\mathcal{B})$, where

$$Y(P_\mathcal{B}) := \frac{1}{(\lambda-1)P_\mathcal{B}(1-P_\mathcal{B})} - \frac{1}{P_\mathcal{B}} + \frac{1}{1-P_\mathcal{B}} .$$

Suppose for the moment that $P_\mathcal{B}$ is a continuous decision variable. Accordingly,

$$\frac{dY(P_\mathcal{B})}{dP_\mathcal{B}} = \frac{1}{P_\mathcal{B}^2(1-P_\mathcal{B})^2}\left[2P_\mathcal{B}^2 + \frac{2-\lambda}{\lambda-1}(2P_\mathcal{B}-1)\right] . \tag{A.3}$$

It is readily verified that $dY(P_\mathcal{B})/dP_\mathcal{B} < 0$ for $0 < P_\mathcal{B} < \bar{P}(\lambda)$ and $dY(P_\mathcal{B})/dP_\mathcal{B} > 0$ for $\bar{P}(\lambda) < P_\mathcal{B} < 1$, where

$$\bar{P}(\lambda) \equiv \frac{\lambda - 2 + \sqrt{\lambda(2-\lambda)}}{2(\lambda-1)} .$$

Note that for $\lambda \leq 2$ the critical value $\bar{P}(\lambda) \in [0, 1/2)$. Hence, excluding a signal of $\mathcal{B}$ increases $Y(P_\mathcal{B})$ if $P_\mathcal{B} < \bar{P}(\lambda)$, whereas including a signal to $\mathcal{B}$ increases $Y(P_\mathcal{B})$ if $P_\mathcal{B} \geq \bar{P}(\lambda)$. With these insights the next two implications follow immediately.

(i) $P_{\mathcal{B}^*} < \bar{P}(\lambda) \implies \mathcal{B}^* \subseteq \mathcal{S}^+$

(ii) $P_{\mathcal{B}^*} \geq \bar{P}(\lambda) \implies \mathcal{S}^+ \subseteq \mathcal{B}^*$

We prove both statements in turn by contradiction. (i) Suppose $P_{\mathcal{B}^*} < \bar{P}(\lambda)$ and that there exists a signal $\hat{s} \in \mathcal{S}^-$ which is also contained in $\mathcal{B}^*$, i.e., $\hat{s} \in \mathcal{B}^*$. Clearly, $\sum_{s\in\mathcal{B}^*}(\gamma_s^H - \gamma_s^L) < \sum_{s\in\mathcal{B}^*\setminus\{\hat{s}\}}(\gamma_s^H - \gamma_s^L)$ because $\hat{s}$ is a bad signal. Moreover, $Y(\mathcal{B}^*) < Y(\mathcal{B}^*\setminus\{\hat{s}\})$ because $Y(\cdot)$ increases when signals are excluded of $\mathcal{B}^*$. Thus $X(\mathcal{B}^*) < X(\mathcal{B}^*\setminus\{\hat{s}\})$, a contradiction to the assumption that $\mathcal{B}^*$ is the optimal partition. (ii) Suppose $P_{\mathcal{B}^*} \geq \bar{P}(\lambda)$ and that there exists a signal $\tilde{s} \in \mathcal{S}^+$ that is not contained in $\mathcal{B}^*$, i.e., $\mathcal{B}^* \cap \{\tilde{s}\} = \emptyset$. Since $\hat{s}$ is a good signal $\sum_{s\in\mathcal{B}^*}(\gamma_s^H - \gamma_s^L) < \sum_{s\in\mathcal{B}^*\cup\{\hat{s}\}}(\gamma_s^H - \gamma_s^L)$. $P_{\mathcal{B}^*} \geq \bar{P}(\lambda)$ implies that $Y(\mathcal{B}^* \cup \{\tilde{s}\}) > Y(\mathcal{B}^*)$. Thus, $X(\mathcal{B}^*) < X(\mathcal{B}^* \cup \{\tilde{s}\})$ a contradiction to the assumption that $\mathcal{B}^*$ maximizes $X(\mathcal{B}^*)$. Finally, since for any $\mathcal{B}^*$ we are either in case (i) or in case (ii), the desired result follows. ■

Proof of Proposition 4

Suppose, in contradiction, that in the optimum there are signals $s, t \in \mathcal{S}$ such that $s \in \mathcal{B}^*$, $t \notin \mathcal{B}^*$ and $\frac{\gamma_s^H - \gamma_s^L}{\gamma_s(\hat{a})} < \frac{\gamma_t^H - \gamma_t^L}{\gamma_t(\hat{a})}$. We derive a contradiction by showing that exchanging signal

s for signal t reduces the principal's cost, which implies that the original contract cannot be optimal. Let $\bar{\mathcal{B}} \equiv (\mathcal{B}^* \setminus \{s\}) \cup \{t\}$.It suffices to show that $X(\bar{\mathcal{B}}) > X(\mathcal{B}^*)$, where $X(\mathcal{B})$ is defined as in the proof of Proposition 3. $X(\bar{\mathcal{B}}) > X(\mathcal{B}^*)$ is equivalent to

$$\left(\sum_{j\in\mathcal{B}^*}(\gamma_j^H-\gamma_j^L)+(\gamma_t^H-\gamma_t^L)-(\gamma_s^H-\gamma_s^L)\right)\left[\frac{1-(\lambda-1)(1-2P_{\bar{\mathcal{B}}})}{(\lambda-1)P_{\bar{\mathcal{B}}}(1-P_{\bar{\mathcal{B}}})}\right] > \left(\sum_{j\in\mathcal{B}^*}(\gamma_j^H-\gamma_j^L)\right)\left[\frac{1-(\lambda-1)(1-2P_{\mathcal{B}^*})}{(\lambda-1)P_{\mathcal{B}^*}(1-P_{\mathcal{B}^*})}\right].$$

Rearranging yields

$$\left[(\gamma_t^H-\gamma_t^L)-(\gamma_s^H-\gamma_s^L)\right]\left[\frac{1-(\lambda-1)(1-2P_{\bar{\mathcal{B}}})}{(\lambda-1)P_{\bar{\mathcal{B}}}(1-P_{\bar{\mathcal{B}}})}\right] > \left(\sum_{j\in\mathcal{B}^*}(\gamma_j^H-\gamma_j^L)\right)\left[\frac{1-(\lambda-1)(1-2P_{\mathcal{B}^*})}{(\lambda-1)P_{\mathcal{B}^*}(1-P_{\mathcal{B}^*})}-\frac{1-(\lambda-1)(1-2P_{\bar{\mathcal{B}}})}{(\lambda-1)P_{\bar{\mathcal{B}}}(1-P_{\bar{\mathcal{B}}})}\right]. \quad \text{(A.4)}$$

With $Y(P_{\mathcal{B}})$ being defined as in the proof of Proposition 3, we have to consider two cases, (i) $dY(P_{\mathcal{B}^*})/P_{\mathcal{B}} \geq 0$, and (ii) $dY(P_{\mathcal{B}^*})/P_{\mathcal{B}} < 0$.

Case (i): Since $\gamma_s(\hat{a}) - \gamma_t(\hat{a}) \leq \kappa$, we have $P_{\mathcal{B}^*} \leq P_{\bar{\mathcal{B}}} + \kappa$. With $Y(P_{\mathcal{B}})$ being (weakly) increasing at $P_{\mathcal{B}^*}$, inequality (A.4) is least likely to hold for $P_{\mathcal{B}^*} = P_{\bar{\mathcal{B}}} + \kappa$. Inserting $P_{\mathcal{B}^*} = P_{\bar{\mathcal{B}}} + \kappa$ into (A.4) yields

$$\left[(\gamma_t^H-\gamma_t^L)-(\gamma_s^H-\gamma_s^L)\right]\left[\frac{1-(\lambda-1)(1-2P_{\bar{\mathcal{B}}})}{(\lambda-1)P_{\bar{\mathcal{B}}}(1-P_{\bar{\mathcal{B}}})}\right] > \left(\sum_{j\in\mathcal{B}^*}(\gamma_j^H-\gamma_j^L)\right)\left[\frac{1-(\lambda-1)(1-2P_{\bar{\mathcal{B}}}-2\kappa)}{(\lambda-1)[P_{\bar{\mathcal{B}}}(1-P_{\bar{\mathcal{B}}})+\kappa(1-2P_{\bar{\mathcal{B}}})]-\kappa^2}-\frac{1-(\lambda-1)(1-2P_{\bar{\mathcal{B}}})}{(\lambda-1)P_{\bar{\mathcal{B}}}(1-P_{\bar{\mathcal{B}}})}\right]. \quad \text{(A.5)}$$

The right-hand side of (A.5) becomes arbitrarily close to zero for $\kappa \to 0$, thus it remains to show that

$$\left[(\gamma_t^H-\gamma_t^L)-(\gamma_s^H-\gamma_s^L)\right]\left[\frac{1-(\lambda-1)(1-2P_{\bar{\mathcal{B}}})}{(\lambda-1)P_{\bar{\mathcal{B}}}(1-P_{\bar{\mathcal{B}}})}\right] > 0\,. \quad \text{(A.6)}$$

For (A.6) to hold, we must have $(\gamma_t^H-\gamma_t^L)-(\gamma_s^H-\gamma_s^L) > 0$. From the proof of Proposition 3 we know that $\mathcal{S}^+ \subseteq \mathcal{B}^*$ if $Y(P_{\mathcal{B}})$ is increasing at $\mathcal{B}^*$. Since the principal will end up including all good signals in the set $\mathcal{B}^*$ anyway, the question of interest is whether she can benefit from swapping two bad signals. Therefore, we consider case $s,t \in \mathcal{S}^-$, where $\mathcal{S}^- \equiv \{s \in \mathcal{S} | \gamma_s^H - \gamma_s^L < 0\}$. With $s,t \in \mathcal{S}^-$, we have

$$\left[(\gamma_t^H-\gamma_t^L)-(\gamma_s^H-\gamma_s^L)\right] \geq \gamma_t(\hat{a})\gamma_s(\hat{a})\left[\frac{1}{\gamma_s(\hat{a})}\frac{\gamma_t^H-\gamma_t^L}{\gamma_t(\hat{a})}-\frac{1}{\gamma_s(\hat{a})+\kappa}\frac{\gamma_s^H-\gamma_s^L}{\gamma_s(\hat{a})}\right], \quad \text{(A.7)}$$

where the inequality holds because $\gamma_t(\hat{a}) - \gamma_s(\hat{a}) \leq \kappa$. Note that for $\kappa \to 0$ the right-hand side of (A.7) becomes strictly positive, thus $(\gamma_t^H-\gamma_t^L)-(\gamma_s^H-\gamma_s^L) > 0$ for $\kappa \to 0$. Hence, for κ sufficiently small, $X(\mathcal{B}^*) < X(\bar{\mathcal{B}})$, a contradiction to $\mathcal{B}^*$ being optimal.

Case (ii): Since $\gamma_t(\hat{a})-\gamma_s(\hat{a}) \leq \kappa$, we have $P_{\mathcal{B}^*} \geq P_{\bar{\mathcal{B}}}-\kappa$. With $Y(P_{\mathcal{B}})$ being decreasing at $P_{\mathcal{B}^*}$, inequality (A.4) is least likely to hold for $P_{\mathcal{B}^*} = P_{\bar{\mathcal{B}}}-\kappa$. Inserting $P_{\mathcal{B}^*} = P_{\bar{\mathcal{B}}}-\kappa$ into (A.4), and running along the lines of case (i) allows us to establish that, for κ sufficiently small, $X(\mathcal{B}^*) < X(\bar{\mathcal{B}})$, a contradiction to $\mathcal{B}^*$ being optimal.

To sum up, for κ sufficiently small we have

$$\max_{s\in\mathcal{S}\setminus\mathcal{B}^*}\{(\gamma_s^H-\gamma_s^L)/\gamma_s(\hat{a})\} < \min_{s\in\mathcal{B}^*}\{(\gamma_s^H-\gamma_s^L)/\gamma_s(\hat{a})\}\ ,$$

or equivalently,

$$\max_{s\in\mathcal{S}\setminus\mathcal{B}^*}\{\gamma_s^H/\gamma_s^L\} < \min_{s\in\mathcal{B}^*}\{\gamma_s^H/\gamma_s^L\}\ .$$

Letting $K \equiv \min_{s\in\mathcal{B}^*}\{\gamma_s^H/\gamma_s^L\}$ establishes the desired result. ■

Proof of Proposition 5

(i) Suppose that a small change in λ leaves the optimal partition $\mathcal{B}^*$ of the set of all signals unchanged. Rearranging (IC′) yields

$$b^* = \frac{c'(\hat{a})}{\sum_{s\in\mathcal{B}^*}(\gamma_s^H-\gamma_s^L)-(\lambda-1)\left[\sum_{s\in\mathcal{B}^*}(\gamma_s^H-\gamma_s^L)\right][1-2P_{\mathcal{B}^*}]}. \tag{A.8}$$

Straight-forward differentiation reveals that

$$\frac{db^*}{d\lambda} = \frac{c'(\hat{a})\left[\sum_{s\in\mathcal{B}^*}(\gamma_s^H-\gamma_s^L)\right][1-2P_{\mathcal{B}^*}]}{\left\{\sum_{s\in\mathcal{B}^*}(\gamma_s^H-\gamma_s^L)-(\lambda-1)\left[\sum_{s\in\mathcal{B}^*}(\gamma_s^H-\gamma_s^L)\right][1-2P_{\mathcal{B}^*}]\right\}^2}.$$

Since under the second-best contract $\sum_{s\in\mathcal{B}^*}(\gamma_s^H-\gamma_s^L) > 0$, the desired result follows.

(ii) Let $\mathcal{B}^+ \equiv \left\{\mathcal{B}\subset\mathcal{S}|\sum_{s\in\mathcal{B}}(\gamma_s^H-\gamma_s^L)>0\right\}$. For any $\tilde{\mathcal{B}}\in\mathcal{B}^+$, let

$$b_{\tilde{\mathcal{B}}} = \frac{c'(\hat{a})}{\sum_{s\in\tilde{\mathcal{B}}}(\gamma_s^H-\gamma_s^L)-(\lambda-1)\left[\sum_{s\in\tilde{\mathcal{B}}}(\gamma_s^H-\gamma_s^L)\right][1-2P_{\tilde{\mathcal{B}}}]}$$

and

$$\underline{u}_{\tilde{\mathcal{B}}} = \bar{u}+c(\hat{a})-b_{\tilde{\mathcal{B}}}P_{\tilde{\mathcal{B}}}+(\lambda-1)P_{\tilde{\mathcal{B}}}(1-P_{\tilde{\mathcal{B}}})b_{\tilde{\mathcal{B}}}.$$

The cost of implementing action $\hat{a}$ when paying $\underline{u}_{\tilde{\mathcal{B}}}$ for signals in $\mathcal{S}\setminus\tilde{\mathcal{B}}$ and $\underline{u}_{\tilde{\mathcal{B}}}+b_{\tilde{\mathcal{B}}}$ for signals in $\tilde{\mathcal{B}}$ is given by

$$C_{\tilde{\mathcal{B}}} \;=\; \underline{u}_{\tilde{\mathcal{B}}}+b_{\tilde{\mathcal{B}}}P_{\tilde{\mathcal{B}}} \;=\; \bar{u}+c(\hat{a})+\frac{c'(\hat{a})(\lambda-1)P_{\tilde{\mathcal{B}}}(1-P_{\tilde{\mathcal{B}}})}{\left[\sum_{s\in\tilde{\mathcal{B}}}(\gamma_s^H-\gamma_s^L)\right][1-(\lambda-1)(1-2P_{\tilde{\mathcal{B}}})]}. \tag{A.9}$$

Differentiation of $C_{\tilde{\mathcal{B}}}$ with respect to λ yields

$$\frac{dC_{\tilde{\mathcal{B}}}}{d\lambda} = \frac{c'(\hat{a})(\lambda-1)P_{\tilde{\mathcal{B}}}(1-P_{\tilde{\mathcal{B}}})}{\left[\sum_{s\in\tilde{\mathcal{B}}}(\gamma_s^H-\gamma_s^L)\right][1-(\lambda-1)(1-2P_{\tilde{\mathcal{B}}})]^2}.$$

Obviously, $dC_{\tilde{\mathcal{B}}}/d\lambda > 0$ for all $\mathcal{B}\in\mathcal{B}^+$. Since the optimal partition of $\mathcal{S}$ may change as λ changes, the minimum cost of implementing action $\hat{a}$ is given by

$$C(\hat{a}) = \min_{\mathcal{B}\in\mathcal{B}^+} C_{\mathcal{B}}.$$

Put differently, $C(\hat{a})$ is the lower envelope of all $C_{\mathcal{B}}$ for $\mathcal{B} \in \mathcal{B}^+$. With $C_{\mathcal{B}}$ being continuous and strictly increasing in λ for all $\mathcal{B} \in \mathcal{B}^+$, it follows that also $C(\hat{a})$ is continuous and strictly increasing in λ. This completes the proof. ■

Proof of Lemma 2

We show that program (MG) has a solution, i.e., $\sum_{s=1}^{S} \gamma_s(\hat{a})h(u_s)$ achieves its greatest lower bound. First, from Lemma 1 we know that the constraint set of program (MG) is not empty for action $\hat{a} \in (0,1)$. Next, note that from (IR_G) it follows that $\sum_{s=1}^{S} \gamma_s(\hat{a})u_s$ is bounded below. Following the reasoning in the proof of Proposition 1 of Grossman and Hart (1983), we can artificially bound the constraint set – roughly spoken because unbounded sequences in the constraint set make $\sum_{s=1}^{S} \gamma_s(\hat{a})h(u_s)$ tend to infinity by a result from Bertsekas (1974). Since the constraint set is closed, the existence of a minimum follows from Weierstrass' theorem. ■

Proof of Lemma 3

Since (IR_G) will always be satisfied with equality due to an appropriate adjustment of the lowest intrinsic utility level offered, relaxing (IR_G) will always lead to strictly lower costs for the principal. Therefore, the shadow value of relaxing (IR_G) is strictly positive, so $\mu_{IR} > 0$.

Next, we show that relaxing (IC_G) has a positive shadow value, $\mu_{IC} > 0$. We do this by showing that a decrease in $c'(\hat{a})$ leads to a reduction in the principal's minimum cost of implementation. Let $\{u_s^*\}_{s\in\mathcal{S}}$ be the optimal contract under (the original) Program MG, and suppose that $c'(\hat{a})$ decreases. Now the principal can offer a new contract $\{u_s^N\}_{s\in\mathcal{S}}$ of the form

$$u_s^N = \alpha u_s^* + (1-\alpha)\sum_{t=1}^{S} \gamma_t(\hat{a})u_t^* \,, \tag{A.10}$$

where $\alpha \in (0,1)$, which also satisfies (IR_G), the relaxed (IC_G), and (OC_G), but yields strictly lower costs of implementation than the original contract $\{u_s^*\}_{s\in\mathcal{S}}$.

Clearly, for $\hat{\alpha} \in (0,1)$, $u_s^N < u_{s'}^N$ if and only if $u_s^* < u_{s'}^*$, so (OC_G) is also satisfied under contract $\{u_s^N\}_{s\in\mathcal{S}}$.

Next, we check that the relaxed (IC_G) holds under $\{u_s^N\}_{s\in\mathcal{S}}$. To see this, note that for $\alpha = 1$ we have $\{u_s^N\}_{s\in\mathcal{S}} \equiv \{u_s^*\}_{s\in\mathcal{S}}$. Thus, for $\alpha = 1$, the relaxed (IC_G) is oversatisfied under $\{u_s^N\}_{s\in\mathcal{S}}$. For $\alpha = 0$, on the other hand, the left-hand side of (IC_G) is equal to zero, and the relaxed (IC_G) in consequence is not satisfied. Since the left-hand side of (IC_G) is continuous in α under contract $\{u_s^N\}_{s\in\mathcal{S}}$, by the intermediate-value theorem there exists $\hat{\alpha} \in (0,1)$ such that the relaxed (IC_G) is satisfied with equality.

Last, consider (IR_G). The left-hand side of (IR_G) under contract $\{u_s^N\}_{s\in\mathcal{S}}$ with $\alpha = \hat{\alpha}$

amounts to

$$\begin{aligned}
&\sum_{s=1}^{S}\gamma_s(\hat{a})u_s^N-(\lambda-1)\sum_{s=1}^{S-1}\sum_{t=s+1}^{S}\gamma_s(\hat{a})\gamma_t(\hat{a})\left[u_t^N-u_s^N\right]\\
&=\sum_{s=1}^{S}\gamma_s(\hat{a})u_s^*-\tilde{\alpha}(\lambda-1)\sum_{s=1}^{S-1}\sum_{t=s+1}^{S}\gamma_s(\hat{a})\gamma_t(\hat{a})\left[u_t^*-u_s^*\right]\\
&>\sum_{s=1}^{S}\gamma_s(\hat{a})u_s^*-(\lambda-1)\sum_{s=1}^{S-1}\sum_{t=s+1}^{S}\gamma_s(\hat{a})\gamma_t(\hat{a})\left[u_t^*-u_s^*\right]\\
&=\bar{u}+c(\hat{a})\;, \qquad\qquad (A.11)
\end{aligned}$$

where the last equality follows from the fact that $\{u_s^*\}_{s\in\mathcal{S}}$ fulfills the (IR_g) with equality. Thus, contract $\{u_s^N\}_{s\in\mathcal{S}}$ is feasible in the sense that all constraints of program (MG) are met. It remains to show that the principal's costs are reduced. Since $h(\cdot)$ is strictly convex, the principal's objective function is strictly convex in α, with a minimum at $\alpha=0$. Hence, the principal's objective function is strictly increasing in α for $\alpha\in(0,1]$. Since $\{u_s^N\}_{s\in\mathcal{S}}\equiv\{u_s^*\}_{s\in\mathcal{S}}$ for $\alpha=1$, for $\alpha=\hat{\alpha}$ we have

$$\sum_{s=1}^{S}\gamma_s(\hat{a})h(u_s^*)>\sum_{s=1}^{S}\gamma_s(\hat{a})h(u_s^N),$$

which establishes the desired result. ■

Proof of Proposition 6

For the agent's intrinsic utility function being sufficiently linear, the principal's costs are approximately given by a second-order Taylor polynomial about $r=1$, thus

$$C(\boldsymbol{u}|r)\approx\sum_{s\in\mathcal{S}}\gamma_s(\hat{a})u_s+\Omega(\boldsymbol{u}|r)\;, \qquad (A.12)$$

where

$$\Omega(\boldsymbol{u}|r)\equiv\sum_{s\in\mathcal{S}}\gamma_s(\hat{a})\left[(u_s\ln u_s)(r-1)+(1/2)u_s(\ln u_s)^2(r-1)^2\right]. \qquad (A.13)$$

Relabeling signals such that the wage profile is increasing allows us to express the incentive scheme in terms of increases in intrinsic utility. The agent's binding participation constraint implies that

$$u_1=\bar{u}+c(\hat{a})-\sum_{s=2}^{S}b_s\left\{\sum_{\tau=s}^{S}\gamma_\tau(\hat{a})-(\lambda-1)\left[\sum_{\tau=s}^{S}\gamma_\tau(\hat{a})\right]\left[\sum_{t=1}^{s-1}\gamma_t(\hat{a})\right]\right\}\equiv u_1(\boldsymbol{b}) \qquad (A.14)$$

and $u_s=u_1(\boldsymbol{b})+\sum_{t=2}^{s}b_t\equiv u_s(\boldsymbol{b})$ for all $s=2,\ldots,S$. Inserting the binding participation constraint into the above cost function and replacing $\Omega(\boldsymbol{u}|r)$ equivalently by $\tilde{\Omega}(\boldsymbol{b}|r)\equiv\Omega(u_1(\boldsymbol{b}),\ldots,u_S(\boldsymbol{b})|r)$ yields

$$C(\boldsymbol{b}|r)\approx\bar{u}+c(\hat{a})+(\lambda-1)\sum_{s=2}^{S}b_s\left[\sum_{\tau=s}^{S}\gamma_\tau(\hat{a})\right]\left[\sum_{t=1}^{s-1}\gamma_t(\hat{a})\right]+\tilde{\Omega}(\boldsymbol{b}|r)\;. \qquad (A.15)$$

Hence, for a given increasing wage profile the principal's cost minimization problem is:

PROGRAM ME:

$$\min_{\boldsymbol{b}\in\mathbb{R}_+^{S-1}} \ \boldsymbol{b}'\boldsymbol{\rho}(\hat{\boldsymbol{\gamma}},\lambda,\hat{a})+\tilde{\Omega}(\boldsymbol{b}|r)$$
$$\text{subject to} \quad \boldsymbol{b}'\boldsymbol{\beta}(\hat{\boldsymbol{\gamma}},\lambda,\hat{a})=c'(\hat{a}) \tag{IC$'$}$$

If r is sufficiently close to 1, then the incentive scheme that solves program ML also solves program ME. Note that generically program ME is solved only by bonus schemes. Put differently, even if there are multiple optimal contracts for program ML, all these contracts are generically simple bonus contracts. Thus, from Proposition 2 it follows that generically for r close to 1 the optimal incentive scheme entails a minimum of wage differentiation. Note that for $\lambda = 1$ the principal's problem is to minimize $\tilde{\Omega}(\boldsymbol{b}|r)$ even for r sufficiently close to 1. ■

Proof of Proposition 7

First consider $b \geq 0$. We divide the analysis for $b \geq 0$ into three subcases.

Case 1 ($a_0 < 0$): For the effort level $\hat{a}$ to be chosen by the agent, this effort level has to satisfy the following incentive compatibility constraint:

$$\hat{a} \in \arg\max_{a\in[0,1]} \ u+\gamma(a)b-\gamma(a)(1-\gamma(a))b(\lambda-1)-\frac{k}{2}a^2 \tag{IC}$$

For $\hat{a}$ to be a zero of $dE\left[U(a)\right]/da$, the bonus has to be chosen according to

$$b^*(\hat{a})=\frac{k\hat{a}}{(\gamma^H-\gamma^L)\left[2-\lambda+2\gamma(\hat{a})(\lambda-1)\right]}.$$

For $a > a_0$, $b^*(a)$ is a strictly increasing and strictly concave function with $b^*(0) = 0$. Hence, each $\hat{a} \in [0,1]$ can be made a zero of $dE\left[U(a)\right]/da$ with a non-negative bonus. By choosing the bonus according to $b^*(\hat{a})$, $\hat{a}$ satisfies, by construction, the first-order condition. Inserting $b^*(\hat{a})$ into the $d^2E\left[U(a)\right]/da^2$ shows that expected utility is strictly concave function if $a_0 < 0$. Hence, with the bonus set equal to $b^*(\hat{a})$, effort level $\hat{a}$ satisfies the second-order condition for optimality and therefore is incentive compatible.

Case 2 ($a_0 = 0$): Just like in the case where $a_0 < 0$, each effort level $a \in [0,1]$ turns out to be implementable with a no-nnegative bonus. To see this, consider bonus

$$b_0=\frac{k}{2(\gamma^H-\gamma^L)^2(\lambda-1)}.$$

For $b \leq b_0$, $dE\left[U(a)\right]/da < 0$ for each $a > 0$, that is, lowering effort increases expected utility. Hence, the agent wants to choose an effort level as low as possible and therefore exerts no effort at all. If, on the other hand, $b > b_0$, then $dE\left[U(a)\right]/da > 0$. Now, increasing effort increases expected utility, and the agent wants to choose effort as high as possible. For $b = b_0$, expected utility is constant over all $a \in [0,1]$, that is, as long as his

participation constraint is satisfied, the agent is indifferent which effort level to choose. As a tie-breaking rule we assume that, if indifferent between several effort levels, the agent chooses the effort level that the principal prefers.

Case 3 ($a_0 > 0$): If $a_0 > 0$, the agent either chooses $a = 0$ or $a = 1$. To see this, again consider bonus b_0. For $b \leq b_0$, $dE\left[U(a)\right]/da < 0$ for each $a > 0$. Hence, the agent wants to exert as little effort as possible and chooses $a = 0$. If, on the other hand, $b > b_0$, then $d^2E\left[U(a)\right]/da^2 > 0$, that is, expected utility is a strictly convex function of effort. In order to maximize expected utility, the agent will choose either $a = 0$ or $a = 1$ depending on whether $E\left[U(0)\right]$ exceeds $E\left[U(1)\right]$ or not.

Negative Bonus: $b < 0$

Let $b^- < 0$ denote the monetary punishment that the agent receives if the good signal is observed. With a negative bonus, the agent's expected utility is

$$E\left[U(a)\right] = u + \gamma(a)b^- + \gamma(a)(1-\gamma(a))\lambda b^- + (1-\gamma(a))\gamma(a)(-b^-) - \frac{k}{2}a^2. \quad \text{(A.16)}$$

The first derivative with respect to effort,

$$\frac{dE\left[U(a)\right]}{da} = \underbrace{(\gamma^H - \gamma^L)b^-\left[\lambda - 2\gamma(a)(\lambda-1)\right]}_{MB^-(a)} - \underbrace{ka}_{MC(a)},$$

reveals that $MB^-(a)$ is a positively sloped function, which is steeper the harsher the punishment is, that is, the more negative b^- is. It is worthwhile to point out that if bonus and punishment are equal in absolute value, $|b^-| = b$, then also the slopes of $MB^-(a)$ and $MB(a)$ are identical. The intercept of $MB^-(a)$ with the horizontal axis, a_0^- again is completely determined by the model parameters:

$$a_0^- = \frac{\lambda - 2\gamma^L(\lambda-1)}{2(\gamma^H - \gamma^L)(\lambda-1)}$$

Note that $a_0^- > 0$ for $\gamma^L \leq 1/2$. For $\gamma^L > 1/2$ we have $a_0^- < 0$ if and only if $\lambda > 2\gamma^L/(2\gamma^L - 1)$. Proceeding in exactly the same way as in the case of a non-negative bonus yields a familiar results: effort level $\hat{a} \in [0,1]$ is implementable with a strictly negative bonus if and only if $a_0^- \leq 0$. Finally, note that $a_0 < a_0^-$. Hence a negative bonus does not improve the scope for implementation. ■

Proof of Proposition 8

Throughout the analysis we restricted attention to non-negative bonus payment. It remains to be shown that the principal cannot benefit from offering a negative bonus payment: implementing action $\hat{a}$ with a negative bonus is at least as costly as implementing action $\hat{a}$ with a positive bonus. In what follows, we make use of notation introduced in the paper as well as in the proof of Proposition 7. Let $a_0(p)$, $a_0^-(p)$, $b^*(\hat{a};p)$, and $u^*(\hat{a};p)$ denote the expressions obtained from a_0, a_0^-, $b^*(\hat{a})$, and $u^*(\hat{a})$, respectively, by replacing $\gamma(\hat{a})$, γ^L, and γ^H with $\gamma(\hat{a};p)$, $\gamma^L(p)$, and $\gamma^H(p)$. From the proof of Proposition 6 we

know that (i) action $\hat{a}$ is implementable with a non-negative bonus (negative bonus) if and only if $a_0(p) \leq 0$ ($a_0^-(p) \leq 0$), (ii) $a_0^-(p) \leq 0$ implies $a_0(p) < 0$. We will show that, for a given value of p, if $\hat{a}$ is implementable with a negative bonus then it is less costly to implement $\hat{a}$ with a non-negative bonus.

Consider first the case where $a_0^-(p) < 0$. The negative bonus payment satisfying incentive compatibility is given by

$$b^-(\hat{a};p) = \frac{k\hat{a}}{(\gamma^H(p) - \gamma^L(p))\left[\lambda - 2\gamma(\hat{a};p)(\lambda - 1)\right]}.$$

It is easy to verify that the required punishment to implement $\hat{a}$ is larger in absolute value than than the respective non-negative bonus which is needed to implement $\hat{a}$, that is, $b^*(\hat{a};p) < |b^-(\hat{a};p)|$ for all $\hat{a} \in (0,1)$ and all $p \in [0,1)$. When punishing the agent with a negative bonus $b^-(\hat{a};p)$, $u^-(\hat{a};p)$ will be chosen to satisfy the corresponding participation constraint with equality, that is,

$$u^-(\hat{a};p) = \bar{u} + \frac{k}{2}\hat{a}^2 - \gamma(\hat{a};p)b^-(\hat{a};p)\left[\lambda - \gamma(\hat{a},p)(\lambda - 1)\right].$$

Remember that, if $\hat{a}$ is implemented with a non-negative bonus, we have

$$u^*(\hat{a};p) = \bar{u} + \frac{k}{2}\hat{a}^2 - \gamma(\hat{a};p)b^*(\hat{a};p)\left[2 - \lambda + \gamma(\hat{a};p)(\lambda - 1)\right].$$

It follows immediately that the minimum cost of implementing $\hat{a}$ with a non-negative bonus is lower than the minimum implementation cost with a strictly negative bonus:

$$\begin{aligned}
C^-(\hat{a};p) &= u^-(\hat{a};p) + \gamma(\hat{a};p)b^-(\hat{a};p) \\
&= \bar{u} + \frac{k}{2}\hat{a}^2 - \gamma(\hat{a};p)b^-(\hat{a};p)\left[\lambda - \gamma(\hat{a};p)(\lambda - 1) - 1\right] \\
&> \bar{u} + \frac{k}{2}\hat{a}^2 + \gamma(\hat{a};p)b^*(\hat{a};p)\left[\lambda - \gamma(\hat{a};p)(\lambda - 1) - 1\right] \\
&= \bar{u} + \frac{k}{2}\hat{a}^2 - \gamma(\hat{a};p)b^*(\hat{a};p)\left[1 - \lambda + \gamma(\hat{a};p)(\lambda - 1)\right] \\
&= \bar{u} + \frac{k}{2}\hat{a}^2 - \gamma(\hat{a};p)b^*(\hat{a};p)\left[2 - \lambda + \gamma(\hat{a};p)(\lambda - 1)\right] + \gamma(\hat{a};p)b^*(\hat{a};p) \\
&= u^*(\hat{a};p) + \gamma(\hat{a};p)b^*(\hat{a};p) \\
&= C(\hat{a};p).
\end{aligned}$$

The same line of argument holds when $a_0^- = 0$: the bonus which satisfies the (IC) is

$$b_0^-(\hat{a};p) = -\frac{k}{2(\gamma^H(p) - \gamma^L(p))^2(\lambda - 1)},$$

and so $b^*(\hat{a};p) < |b_0^-(\hat{a};p)|$ for all $\hat{a} \in (0,1)$ and all $p \in [0,1)$. ■

Proof of Corollary 1

Let $p \in (0,1)$. With $\hat{\boldsymbol{\zeta}}$ being a convex combination of $\hat{\boldsymbol{\gamma}}$ and $\mathbf{1}$ we have $(\zeta^H, \zeta^L) = p(1,1) + (1-p)(\gamma^H, \gamma^L) = (\gamma^H + p(1-\gamma^H), \gamma^L + p(1-\gamma^L))$. The desired result follows

immediately from Proposition 3: Consider $\lambda > 2$. Implementation problems are less likely to be encountered under $\hat{\boldsymbol{\zeta}}$ than under $\hat{\boldsymbol{\gamma}}$. Moreover, if implementation problems are not an issue under both performance measures, then implementation of a certain action is less costly under $\hat{\boldsymbol{\zeta}}$ than under $\hat{\boldsymbol{\gamma}}$. For $\lambda = 2$ implementation problems do not arise and implementation costs are identical under both performance measures. Last, if $\lambda < 2$, implementation problems are not an issue under either performance measure, but the cost of implementation is strictly lower under $\hat{\boldsymbol{\gamma}}$ than under $\hat{\boldsymbol{\zeta}}$. ■

B Validity of the First-Order Approach

Lemma 4: *Given (A1)-(A3), the incentive constraint in the principal's cost minimization problem can be represented as* $E[U'(\hat{a})] = 0$.

Proof: The proof proceeds in two steps. Consider a contract $(u_1, \{b_s\}_{s=2}^S)$ with $b_s \geq 0$ for $s = 2, \ldots, S$, that implements action $\hat{a} \in (0, 1)$. First, we show that it is never optimal for the principal to set $b_s > 0$ if $\beta_s \leq 0$, where we write β_s instead of $\beta_s(\hat{\boldsymbol{\gamma}}, \lambda, \hat{a})$ to cut back on notation. Thereafter, it is shown that for a given contract with $b_s > 0$ if and only if $\beta_s > 0$, all actions that satisfy the first-order condition of the agent's utility maximization problem characterize a local maximum of his utility function. Since the utility function is continuous and all extreme points are local maxima, there exists a unique action that fulfills the first-order condition. This action corresponds to the unique maximum.

Step 1: Irrespective of the first-order approach being valid or not, a necessary condition for $\hat{a} \in (0, 1)$ to be incentive compatible is that (IC′) is satisfied. Note that if (IC′) holds for $\hat{a} \in (0, 1)$, then there exist at least one signal t with $\beta_t > 0$. If there exists $b_s > 0$ with $\beta_s \leq 0$, then the principal can reduce both b_s and also another bonus b_t with $\beta_t > 0$, without violating (IC′). Next, we show that increasing any spread, say b_s, always increases the principal's cost of implementation.

$$C(\boldsymbol{b}) = \sum_{s=1}^{S} \gamma_s(\hat{a}) h\left(u_1(\boldsymbol{b}) + \sum_{t=2}^{s} b_s\right), \tag{B.1}$$

$$\text{where} \qquad u_1(\boldsymbol{b}) = \bar{u} + c(\hat{a}) - \sum_{t=2}^{S} b_s \left[\sum_{\tau=s}^{S} \gamma_\tau(\hat{a}) - (\lambda - 1)\left(\sum_{\tau=s}^{S} \gamma_\tau(\hat{a})\right)\left(\sum_{t=1}^{s-1} \gamma_t(\hat{a})\right)\right].$$

The partial derivative of the cost function with respect to an arbitrary b_k is

$$\frac{\partial C(\boldsymbol{b})}{\partial b_k} = \sum_{s=1}^{k-1} \gamma_s(\hat{a}) h'\left(u_1(\boldsymbol{b}) + \sum_{t=2}^{s} b_s\right)\left[\frac{\partial u_1}{\partial b_k}\right] + \sum_{s=k}^{S} \gamma_s(\hat{a}) h'\left(u_1(\boldsymbol{b}) + \sum_{t=2}^{s} b_s\right)\left[\frac{\partial u_1}{\partial b_k} + 1\right].$$

Rearranging yields

$$\frac{\partial C(\boldsymbol{b})}{\partial b_k} = \sum_{s=1}^{k-1} \gamma_s(\hat{a}) h'(u_s) \underbrace{\left[(\lambda - 1) \left(\sum_{\tau=k}^{S} \gamma_\tau(\hat{a}) \right) \left(\sum_{t=1}^{k-1} \gamma_t(\hat{a}) \right) - \sum_{\tau=k}^{S} \gamma_\tau(\hat{a}) \right]}_{<0} + \sum_{s=k}^{S} \gamma_s(\hat{a}) h'(u_s) \underbrace{\left[(\lambda - 1) \left(\sum_{\tau=k}^{S} \gamma_\tau(\hat{a}) \right) \left(\sum_{t=1}^{k-1} \gamma_t(\hat{a}) \right) - \sum_{\tau=k}^{S} \gamma_\tau(\hat{a}) + 1 \right]}_{>0} . \quad \text{(B.2)}$$

Note $u_s \leq u_{s+1}$ which implies that $h'(u_s) \leq h'(u_{s+1})$. Thus, the following inequality holds

$$\frac{\partial C(\boldsymbol{b})}{\partial b_k} \geq \sum_{s=1}^{k-1} \gamma_s(\hat{a}) h'(u_k) \left[(\lambda - 1) \left(\sum_{\tau=k}^{S} \gamma_\tau(\hat{a}) \right) \left(\sum_{t=1}^{k-1} \gamma_t(\hat{a}) \right) - \sum_{\tau=k}^{S} \gamma_\tau(\hat{a}) \right] + \sum_{s=k}^{S} \gamma_s(\hat{a}) h'(u_k) \left[(\lambda - 1) \left(\sum_{\tau=k}^{S} \gamma_\tau(\hat{a}) \right) \left(\sum_{t=1}^{k-1} \gamma_t(\hat{a}) \right) - \sum_{\tau=k}^{S} \gamma_\tau(\hat{a}) + 1 \right] . \quad \text{(B.3)}$$

The above inequality can be rewritten as follows

$$\frac{\partial C(\boldsymbol{b})}{\partial b_k} \geq h'(u_k) \left[(\lambda - 1) \left(\sum_{\tau=k}^{S} \gamma_\tau(\hat{a}) \right) \left(\sum_{t=1}^{k-1} \gamma_t(\hat{a}) \right) \right] > 0 .$$

Since reducing any bonus lowers the principal's cost of implementation, it cannot be optimal to set $b_s > 0$ for $\beta_s \leq 0$. This completes the first step of the proof.

Step 2: The second derivative of the agent's utility with respect to a is

$$E[U''(a)] = -2(\lambda - 1) \sum_{s=2}^{S} b_s \sigma_s - c''(a) , \quad \text{(B.4)}$$

where $\sigma_s := (\sum_{i=1}^{s-1} \gamma_i^H - \gamma_i^L)(\sum_{i=s}^{S} \gamma_i^H - \gamma_i^L) < 0$. Suppose action $\hat{a}$ satisfies the first-order condition. Formally

$$\sum_{s=2}^{S} b_s \beta_s = c'(\hat{a}) \iff \sum_{s=2}^{S} b_s \frac{\beta_s}{\hat{a}} = \frac{c'(\hat{a})}{\hat{a}} . \quad \text{(B.5)}$$

Action $\hat{a}$ locally maximizes the agent's utility if

$$-2(\lambda - 1) \sum_{s=2}^{S} b_s \sigma_s < c''(\hat{a}) . \quad \text{(B.6)}$$

Under Assumption (A3), we have $c''(\hat{a}) > c(\hat{a})/\hat{a}$. Therefore, if

$$\sum_{s=2}^{S} b_s \big[-2(\lambda - 1)\sigma_s - \beta_s/\hat{a} \big] < 0 , \quad \text{(B.7)}$$

then (B.5) implies (B.6), and each action $\hat{a}$ satisfying the first-order condition of the agent's maximization problem is a local maximum of his expected utility. Inequality

(B.7) obviously is satisfied if each element of the sum is negative. Summand s is negative if and only if

$$
\begin{aligned}
&-2(\lambda-1)\left(\sum_{i=1}^{s-1}(\gamma_i^H-\gamma_i^L)\right)\left(\sum_{i=s}^{S}(\gamma_i^H-\gamma_i^L)\right)\hat{a} \\
&-\left(\sum_{\tau=s}^{S}(\gamma_\tau^H-\gamma_\tau^L)\right)\left[1-(\lambda-1)\left(\sum_{t=1}^{s-1}\gamma_t(\hat{a})\right)\right]-(\lambda-1)\left[\sum_{\tau=s}^{S}\gamma_\tau(\hat{a})\right]\left(\sum_{t=1}^{s-1}(\gamma_t^H-\gamma_t^L)\right)<0\,.
\end{aligned}
$$

Rearranging of the above inequality yields

$$
\begin{aligned}
\left(\sum_{i=s}^{S}(\gamma_i^H-\gamma_i^L)\right)\left\{\lambda+2(\lambda-1)\left[\hat{a}\sum_{i=1}^{s-1}(\gamma_i^H-\gamma_i^L)-\sum_{i=1}^{s-1}\gamma_i(\hat{a})\right]\right\} &> 0 \\
\iff \left(\sum_{i=s}^{S}(\gamma_i^H-\gamma_i^L)\right)\left\{\lambda\left(1-\sum_{i=1}^{s-1}\gamma_i^L\right)+(2-\lambda)\sum_{i=1}^{s-1}\gamma_i^L\right\} &> 0 \qquad \text{(B.8)}
\end{aligned}
$$

The term in curly brackets is positive, since $\lambda \leq 2$ and $\sum_{i=1}^{s-1}\gamma_i^L < 1$. Note that $\beta_s \leq 0$ if and only if $\sum_{i=s}^{S}(\gamma_i^H-\gamma_i^L)\leq 0$. As we have established in step 1 of this proof, in this case it is always optimal for the principal to set $b_s = 0$. Thus, if $b_s > 0$ then $\sum_{i=s}^{S}(\gamma_i^H-\gamma_i^L) > 0$, which completes the proof. ■

References

[1] **Abeler, J.**, **A. Falk**, **L. Götte**, and **D. Huffman (2008):** Reference-Dependent Preferences and Labor Supply, *mimeo*, University of Bonn.

[2] **Barberis, N.**, **M. Huang** and **T. Santos (2001):** Prospect Theory and Asset Prices, *Quarterly Journal of Economics*, Vol. 116, 1-53.

[3] **Becker, G.S.** and **G.J. Stigler (1974):** Law Enforcement, Malfeasance, and Compensation of Enforcers, *The Journal of Legal Studies*, Vol. 3, 1-18.

[4] **Bell, D.E. (1985):** Disappointment in Decision Making under Uncertainty, *Operations Research*, Vol. 33, 1-27.

[5] **Bertsekas, D. (1974):** Necessary and Sufficient Conditions for Existence of an Optimal Portfolio, *Journal of Economic Theory*, Vol. 8, 235-247.

[6] **Blackwell, D. (1951):** Comparison of Experiments, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. by J. Neyman, Berkeley: University of California Press.

[7] **Blackwell, D. (1953):** Equivalent Comparison of Experiments, *Annals of Mathematics and Statistics*, Vol. 24, 265-272.

[8] **Bowman, D.**, **D. Minehart** and **M. Rabin (1999):** Loss-Aversion in a Consumptions-Savings Model, *Journal of Economic Behavior and Organization*, Vol. 38, 155-178.

[9] **Breiter, H.C.**, **I. Aharon**, **D. Kahneman**, **A. Dale** and **P. Shizgal (2001):** Functional Imaging of Neural Responses to Expectancy and Experience of Monetary Gains and Losses, *Neuron*, Vol. 30, 619-639.

[10] **Camerer, C.F.** and **U. Malmendier (2007):** Behavioral Economics of Organizations, in *Behavioral Economics and its Applications*, P. Diamond and H. Vartiainen, editors, Princeton University Press, Princeton, 235-281.

[11] **Caron, L.** and **G. Dionne (1997):** Insurance Fraud Estimation: More Evidence from the Québec Automobile Insurance Industry, *Assurances*, Vol. 64, 567-578.

[12] **Chiappori, P.-A.**, **B. Jullien**, **B. Salanié**, and **F. Salanié (2006):** Asymmetric information in insurance: general testable implications, *RAND Journal of Economics*, Vol. 37, 783-798.

[13] **Churchill, G.A.**, **N.M. Ford**, and **O.C. Walker (1993)**: Sales Force Management, Irwin, Homewood, IL.

[14] **Daido, K.** and **H. Itoh (2007):** The Pygmalion and Galatea Effects: An Agency Model with Reference-Dependent Preferences and Applications to Self-Fulfilling Prophecy, *working paper*, Hitotsubashi University.

[15] **Dantzig, G.B. (1957):** Discrete Variable Extremum Problems, *Operations Research*, Vol. 5, 266-277.

[16] **De Meza, D.** and **D.C. Webb (2007):** Incentive Design Under Loss Aversion, *Journal of the European Economic Association*, Vol. 5, 66-92.

[17] **Demougin, D.** and **C. Fluet (1998):** Mechanism Sufficient Statistic in the Risk-Neutral Agency Problem, *Journal of Institutional and Theoretical Economics*, Vol. 154, 622-639.

[18] **Dionne, G.** and **R. Gagné (2001):** Deductible Contracts Against Fraudulent Claims: Evidence from Automobile Insurance, *Review of Economics and Statistics*, Vol. 83, 290-301.

[19] **Dittmann, I.**, **E. Maug** and **O. Spalt (2007):** Executive Stock Options When Managers Are Loss Averse, *working paper*, University of Mannheim.

[20] **Englmaier, W.** and **A. Wambach (2006):** Optimal Incentive Contracts under Inequity Aversion, *Working Paper*, University of Cologne.

[21] **Fehr, E.** and **K.M. Schmidt (1999):** A Theory of Fairness, Competition, and Cooperation, *Quarterly Journal of Economics*, Vol. 114, 817-868.

[22] **Foppert, D. (1994):** Waging War Against Fraud, *Best's Review: Property-Causality*, Ed., 94.

[23] **Gjesdal, F. (1982):** Information and Incentives: The Agency Information Problem, *Review of Economic Studies*, Vol. 49, 373-390.

[24] **Grossman, S.** and **O. Hart (1983):** An Analysis of the Principal-Agent Problem, *Econometrica*, Vol. 51, 7-45.

[25] **Gul, F. (1991):** A Theory of Disappointment Aversion, *Econometrica*, Vol. 59, 667-686.

[26] **Haller, H. (1985):** The Principal-Agent Problem with a Satisficing Agent, *Journal of Economic Behavior and Organization*, Vol. 6, 359-379.

[27] **Heidhues, P.** and **B. Kőszegi (2008):** Competition and Price Variation when Consumers are Loss Averse, *American Economic Review*, Vol. 98, 1245–1268.

[28] **Holmström, B. (1979):** Moral Hazard and Observability, *The Bell Journal Of Economics*, Vol. 10, 74-91.

[29] **Iantchev, E.P. (2005):** Optimal Contracts with Prospect Theory Preferences in the Presence of Moral Hazard, *working paper*, University of Chicago.

[30] **Joseph, K.** and **M.U. Kalwani (1998):** The Role of Bonus Pay in Salesforce Compensation Plans, *Industrial Marketing Management*, Vol. 27, 147-159.

[31] **Kahneman, D.** and **A. Tversky (1979):** Prospect Theory: An Analysis of Decision under Risk, *Econometrica*, Vol. 47, 263-291.

[32] **Kim, S.K. (1995):** Efficiency of an Information System in an Agency Model, *Econometrica*, Vol. 63 , 89-102.

[33] **Kőszegi, B.** and **M. Rabin (2006):** A Model of Reference-Dependent Preferences, *Quarterly Journal of Economics*, Vol. 121, 1133-1165.

[34] **Kőszegi, B.** and **M. Rabin (2007):** Reference-Dependent Risk Preferences, *American Economic Review*, Vol. 97, 1047-1073.

[35] **Larsen, J.T.**, **A.P. Mc Graw**, **B.A. Mellers**, and **J.T. Cacioppo (2004):** The Agony of Victory and Thrill of Defeat: Mixed Emotional Reactions to Disappointing Wins and Relieving Losses, *Psychological Science*, Vol. 15, 325-330.

[36] **Lazear, E.P.** and **K.L. Shaw (2007)**: Personnel Economics: The Economist's View of Human Resources, *Journal of Economic Perspectives*, Vol. 21, 91-114.

[37] **Lazear, E.P.** and **P. Oyer (2007):** Personnel Economics, *NBER working paper 13480*, http://www.nber.org/papers/w13480.

[38] **Loomes, G.** and **R. Sugden (1986):** Disappointment and Dynamic Consistency in Choice under Uncertainty, *Review of Economic Studies*, Vol. 53, 271-282.

[39] **MacLeod, B. (2003):** Optimal Contracting with Subjective Evaluation, *American Economic Review*, Vol. 93, 216-240.

[40] **Mellers, B.**, **A. Schwartz** and **H. Ritov (1999):** Emotion-Based Choice, *Journal of Experimental Psychology: General*, Vol. 128, 332-345.

[41] **O'Donoghue, T.** and **M. Rabin (1999):** Incentives for Procrastinators, *Quarterly Journal of Economics*, Vol. 114, 769-816.

[42] **Oyer, P. (1998):** Fiscal Year Ends and Non-Linear Incentive Contracts: The Effect on Business Seasonality, *Quarterly Journal of Economics*, Vol. 113, 149-188.

[43] **Park, E.-S. (1995):** Incentive Contracting under Limited Liability, *Journal of Economics & Management Strategy*, Vol. 4, 477-490.

[44] **Prendergast, C. (1999):** The Provision of Incentives in Firms, *Journal of Economic Literature*, Vol. 37, 7-63.

[45] **Post, T.**, **M.J. Van den Assem**, **G. Baltussen** and **R.H. Thaler (2008):** Deal Or No Deal? Decision Making Under Risk in a Large-Payoff Game Show, *American Economic Review*, Vol. 98, 38-71.

[46] **Puelz, R.** and **A. Snow (1994):** Evidence on Adverse Selection: Equilibrium Signalling and Cross-Subsidization in the Insurance Market, *Journal of Political Economy*, Vol. 102, 236-257.

[47] **Rayo, L.** and **G.S. Becker (2007):** Evolutionary Efficiency and Happiness, *Journal of Political Economy*, Vol. 115, 302-337.

[48] **Rothschild, M.** and **J.E. Stiglitz (1976):** Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information, *Quarterly Journal of Economics*, Vol. 90, 639-649.

[49] **Salanié, B. (2003):** Testing Contract Theory, *CESifo Economic Studies*, Vol. 49, 461-477.

[50] **Strausz, R. (2006):** Deterministic Versus Stochastic Mechanisms in Principal-Agent Models, *Journal of Economic Theory*, Vol. 128, 306-314.

[51] **Tversky, A.** and **D. Kahneman (1991):** Loss Aversion in Riskless Choice: A Reference-Dependent Model, *Quarterly Journal of Economics*, Vol. 106, 1039-1061.

[52] **Winter, R.A. (2000):** Optimal Insurance under Moral Hazard, in *Handbook of Insurance*, G. Dionne, editor, Kluwer Academic Publishers, Boston, 155-183.