Gaming or Shirking? On a Fundamental Trade-Off When Designing Incentives

Wendelin Schnedler*

February 22, 2019

Abstract

Examples in which agents 'game' incentives and direct effort to useless or even harmful activities abound. Clearly, gaming matters for whether to use incentive schemes and if so which. Surprisingly, the role of gaming when designing incentives has been confined to informal interpretations and claims. Gaming neither features in the standard procedure of finding optimal incentives nor in results on the value of information for designing them. This paper proposes a proper definition of gaming and uses it to render the role of gaming for designing incentives explicit.

Keywords: multi-tasking, gaming, misdirected effort, generalized sufficient statistic result, value of information, alignment, incentive scheme, LEN model

JEL-Codes: D86, M52, M41, J33

*University of Paderborn, Faculty of Economics, Warburger Straße 100, D-33098 Paderborn (wendelin.schnedler@upb.de). The author wishes to thank Bob Gibbons, Paul Grout, Bentley MacLeod, Michael Raith, Dirk Sliwka and Jan Zabojnik for insightful comments. The paper also profited from the contributions by participants of research seminars in Berlin (WZB), Bielefeld, Bristol, Cologne, Cornell, Hamburg, Heidelberg, Lyon, Mannheim, and Virginia Tech. All errors remain my own.

1 Introduction

Ignoring that economic agents may 'game' incentives has let to problems in a plethora of real-life incentive schemes.[1] For an example, consider the attempt of the U.S. Agency of International Development (USAID) to fight an infestation of Colorado potato beetles in Afghanistan by paying 5 dollar for the delivery of each bottle full of these beetles. Incentives were suppressed when locals were found breeding beetles, instead of collecting them.[2] For a less extreme example, consider strategic citations that researchers use in the hope to increase publication chances without increasing the scientific value of their manuscript.[3] In both examples, incentives are successful in getting some agent (local or academic) to exert effort but this effort is at least partially directed to a course of action that is useless or even harmful to the principal (USAID, Principal or President of the University) who introduced incentives.

Holmström and Milgrom (1991) and Baker (1992) show that fully or partially suppressing incentives (as done by USAID) can be optimal and attribute this to the agent wrongly allocating 'effort', 'time' or 'attention' across different tasks,[4] or 'gaming' incentives. All this suggests that 'dysfunctional behavior' by the agent should affect whether and which incentives to use.

The Nobel Prize Committee (2016), in honoring the work by Bengt Holmström, claims that a multi-dimensional action choice by the agent (or: multitasking) is required to formalize arguments on how 'dysfunctional behavior' affects incentive design. The formal approach of finding optimal incentives under multitasking, however, does not seem to be affected by such behavior. Just as in the single

[1] For more examples, see Stephen J. Dubner's freakonomics podcast from Nov 10, 2012. Retrieved 19 Oct 2015 from http://freakonomics.com/2012/10/11/the-cobra-effect-a-new-freakonomics-radio-podcast/. Examples of gaming are referred to in survey articles on incentives by Gibbons (1998) and Prendergast (1999) and are analyzed empirically by Oyer (1998), Dranove et al. (2003), Courty and Marschke (2004), Courty and Marschke (2008), Propper et al. (2010), Hong et al. (2013), Larkin (2014), Sloof and van Praag (2015), or Forbes et al. (2015).

[2] Ben Arnoldy, Christian Science Monitor (28th of July 2010).

[3] Frey (2003) claims that some academics respond to the strong incentives to publish by 'prostituting' themselves instead of pursuing original research. (Ironically, he provided a case in point by boosting his own publication count through submitting resembling ideas to different journals— see Frey, 2011)

[4] Holmström and Milgrom do not explicitly link their propositions to the agent's allocation decision but mentioned it in the title and first paragraph of Section 3 in their paper.

task case, the designer minimizes the loss from having to use incentives (agency costs), or equivalently, she maximizes her utility under the agent's participation and incentive constraint.

Findings on whether and which performance information to use for providing incentives under multitasking also remain mute about 'gaming'. Christensen et al. (2010) show that key results, e.g., the sufficient statistic result by Holmström (1979, 1982), carry over from single to multitasking. These results, however, imply that any independent piece of information about the agent's behavior, such as the number of beetle filled bottles, should be used to provide incentives—as if 'gaming' would not matter.[5]

Lacking a proper definition of 'gaming', 'misdirected effort', or 'dysfunctional behavior', it has so far been impossible to formally analyze the role of such behavior for incentive design. Instead, we had to content ourselves with informal interpretations and claims. The present paper proposes a proper definition of gaming and uses it to pin down its role when choosing incentives.

In the initial examples, incentives induce the agent to provide some resource, e.g., effort, time or attention, and then badly allocate it. A proper definition of gaming requires a description of the resource, to what it can be allocated, and what makes the allocation 'bad'.

In the definition that I propose, the resource is the agent's effort e, a real number that describes the agent's preferences over different courses of action in absence of incentives, or equivalently, his (psychological) costs from a given action choice. Accordingly, taking a nap requires less effort from a local than collecting beetles as long as he prefers the former to the latter. For a local who equally dislikes spending 5 minutes outside in the hot sun collecting beetles or 30 min inside breeding them, collecting and breeding would require the same effort, say $\tilde{e}$. Incentives generate effort $\tilde{e}$ in the sense that they produce displeasure: the agent gives up a course of action that he prefers (taking a nap) for something that he dislikes (breeding).

The effort $\tilde{e}$ that is generated by incentives can be allocated to different courses of action: collecting or breeding. An agent is then said to *game incentives* if the

[5]Relatedly, the set of useless signals in the multitasking LEN model characterized by Feltham and Xie (1994) in Condition (13) has measure zero in the set of all signals. Any signal that is somehow related to the principal's benefit, even if negatively (like the number of beetle-filled bottles), should be used— see also equation (3) in Baker (2002).

generated effort is not used for the action choice that is most beneficial to the principal (here: collecting) but for something else (here: breeding).

This definition of gaming offers —to my knowledge— the first formal reading of the informal claims that 'seemingly dysfunctional behavior' could have been avoided by 'basing pay on an employee's contribution to firm value' (Baker et al., 1994) or that 'rewarding for A while hoping for B' is 'foolish' (Kerr, 1975). Using a simple multitasking model, Section 2 shows that aligning incentives with the principal's benefit prevents gaming (Proposition 1). The specific definition of gaming proposed here is crucial for this result. As will become clear, aligning incentives does, for example, not ensure that the agent allocates his working time to the most beneficial course of action.

Any insight on how to avoid gaming is only useful if incentive designers care about gaming. But given that their aim is minimizing agency costs, why would they? With the proposed definition, this question can be answered. Gaming contributes to agency costs just as the agent's unwillingness to exert effort (shirking) and both are thus implicitly taken into account when incentive designers try to reduce these costs (Proposition 2).

For the intuition, consider as a (hypothetical) benchmark how much the principal would have gained had the agent used his effort to the principal's largest benefit. Adding and subtracting this benchmark to agency costs (a trick similar to that in the Hicks-decomposition into income and substitution effect) reveals that for *any* incentive scheme, whether optimally chosen or not, these costs have two components: the principal's loss from gaming and shirking.

When minimizing agency costs, incentive designers will exploit any chance to reduce shirking costs without increasing gaming costs and vice versa. Ultimately, these opportunities are exhausted and losses from shirking and gaming need to be traded off (Proposition 3). In particular, the incentive designer may want to trade more gaming for less shirking. When publications are the only measure of research output, one has to accept strategic citations (gaming) in order to get academics to strive harder (reduce shirking). Rather than 'foolish', incentives that are not aligned with the principal's benefit can be optimal although they are gamed by the agent (see Section 2).

Incentives can be suppressed because they (i) lead the agent to reduce some

valuable input or (ii) misdirect the agent's effort. While the first argument has been formalized in many ways,[6] the lack of a proper definition of 'misdirected effort' meant that the second had to remain verbal. Using the definition of gaming, it is formalized here for the first time (Corollary 1).

All these insights on how 'dysfunctional behavior' affects incentives hold relatively generally. In contrast to the claim by the Nobel Prize Committee (2016), a multidimensional choice is not required to formalize them. Section 3 introduces a framework that embeds most single and multitasking moral-hazard models. In this framework, a multi-dimensional action space is neither necessary nor sufficient for gaming (Lemma 1). Still, aligning incentives with the principal's benefit prevents gaming (Proposition 4), agency costs are due to gaming and shirking (Proposition 5) and both have to be traded-off (Corollary 2). Moreover, gaming continues to be the only reason to ignore performance information and suppress incentives (Corollary 3). Since the model by Holmström and Milgrom (1991) is a special case of this framework, the reason for incentives to be suppressed in their model is also gaming as defined here.

Agency costs could, of course, be decomposed in many ways. The proposed decomposition, however, exactly identifies what invalidates standard results when moving beyond the traditional moral-hazard model. Section 4 revisits Holmström's sufficient statistic result (1979) and Kim's ranking of information systems (1995) when effort can be misdirected. Both results are shown to fully apply to shirking costs (Corollary 4 and 5) but not to agency costs (Proposition 6). It is thus the precise definition of gaming here, which overturns these results. This also clarifies that any generalizations of such results, for example those by Christensen et al. (2010), have to remain partial because they ignore the effect of gaming.

The trade-off between gaming and shirking is not the first attempt to capture Holmström and Milgrom's notion (1991) that it matters for incentive design how well the agent's activity at different tasks is measured. Feltham and Wu (2000) and Baker (2000, 2002) propose that more congruity of a performance measure with the benefit comes at the price of less precision. While this implicitly assumes

[6] See Seabright (2009); Bénabou and Tirole (2003, 2006); Sliwka (2007); Herold (2010); Friebel and Schnedler (2011); Schnedler (2011); Schnedler and Vadovic (2011); van der Weele (2012); Schnedler and Vanberg (2014).

that the principal prefers more to less congruent measures of the same precision, she actually prefers measures that emphasize tasks that the agent likes (Schnedler, 2008, Proposition 2). Section 5 uses the gaming-shirking approach to re-visit this finding. While congruity between the performance measure and the principal's benefit is desirable in the sense that it eliminates gaming (Corollary 6), it increases the loss from shirking (Corollary 7).

2 Gaming and Shirking: A Multitasking Example

The aim of this section is to explicitly identify the role of gaming for the design of incentives in a multitasking example.

First, I introduce a simple multitasking principal-agent model and describe the standard approach of finding optimal incentives. Then, I propose a definition of gaming and show that it seems to capture at least what Baker et al. (1994) mean by 'seemingly dysfunctional behavior' by formalizing their verbal claim that aligning incentives prevents such behavior (Proposition 1). I also derive the costs of gaming and contrast them with the costs of shirking. Then, agency costs are decomposed into gaming and shirking costs (Proposition 2). This illustrates *that* gaming affects incentive design in the standard approach of finding optimal incentives. The resulting trade-off between gaming and shirking (Proposition 3) then explains *how* gaming affects incentive design. Finally, the reason to suppress incentives is identified to be gaming (Corollary 1).

2.1 The model

Consider the principal of a university (she) who wants an academic (agent, he) to engage in research but also in marketing this research. The academic can allocate one unit of time between thinking about research, $a_1 \in [0;1]$, marketing, $a_2 \in [0;1]$, or something of no value to the principal (like his next holiday): $a_1 + a_2 \leq 1$

The principal benefits if the academic spends time on research and marketing: $b(a) = \beta a_1 + (1-\beta)a_2$, where $\beta \in (0,1)$ describes the beneficial effect of research relative to marketing. The principal also has control over some good that can be used as a reward: teaching reductions, more office space, etc. In order to reflect the

opportunity costs of these resources to the principal, let her utility be a function of benefit b and reward r: $v(b,r) = b - r$.

The academic likes reward r but dislikes spending time on either research or marketing. His utility is: $u(a,r) = r - a_1^2 - a_2^2$. He has the outside option of spending no time on research or marketing, $a^0 = (0,0)$ and receiving no reward $u(a^0,0) = 0$. Finally, he is assumed to be risk-neutral. This assumption is not crucial and can be generalized. Imposing it, helps to make the point that the gaming-shirking trade-off (to be identified later) applies even in the absence of insurance problems.

There is very little objective information that can be used in this example to provide incentives. A court of law cannot verify what the academic is thinking or which benefit the principal derives from it (only the principal knows this). However, the principal can reward publication success ($Y = 1$) or failure ($Y = 0$). Success is more likely when the academic thinks longer about research or spends more time contemplating how to market this research. For simplicity, assume $Prob(Y = 1|a_1,a_2) = \rho a_1 + (1-\rho)a_2$, where $\rho \in (0,1)$ describes the relative importance of research for publication success. The relative weighting ρ is exogenously determined by the taste of referees and editors, not by the university principal. These tastes may differ from the preferences of the princpal, $\rho \neq \beta$. Contracts take the form of a salary $\underline{\pi}$ that is independent from performance and a premium π that is only paid in case of success.

2.2 Standard approach to find optimal incentives

If the course of action could be stipulated in a contract, the principal would pick the action choice a^* that maximizes her benefit while ensuring with an appropriate compensation r^* that the academic is not worse off. The solution to this first-best

problem is:[7]

$$a^* = (a_1^*, a_2^*) = \left(\frac{\beta}{2}, \frac{1-\beta}{2}\right) \text{ and } r^* = (a_1^*)^2 + (a_2^*)^2 = \frac{1}{4}(\beta^2 + (1-\beta)^2). \quad (1)$$

If the action choice cannot be stipulated but is induced by incentives $(\underline{\pi}, \pi)$, the agent selects a course of action that maximizes his expected utility:

$$a^\pi \in \arg \max_{a \in \{(a_1,a_2) | a_1 + a_2 \leq 1\}} \quad \underline{\pi} + \pi \cdot (\rho a_1 + (1-\rho) a_2) - a_1^2 - a_2^2.$$

The induced action can be computed using first-order conditions for $0 \leq \pi \leq 2$:[8]

$$a^\pi = (a_1^\pi, a_2^\pi) = (\rho, (1-\rho)) \cdot \frac{\pi}{2}. \quad (2)$$

The larger this premium π, the more time the academic will spend thinking about research and marketing. The more emphasis ρ referees place on research, the less time the academic will spend on marketing.

Incentive are then designed to maximize the principal's benefit while ensuring that the agent is not worse off (which is represented by a participation constraint PC), but also keeping in mind that the agent is only willing to engage in certain choices (which is represented by the incentive constraint IC):[9]

$$\begin{aligned} &\max_{\underline{\pi}, \pi \in [0,2], a^\pi} \beta a_1 + (1-\beta) a_2 - r^\pi \\ &\text{such that } u(a^\pi, r^\pi) \geq 0 && \text{(PC)} \\ &\qquad \text{and } a^\pi = (a_1^\pi, a_2^\pi) = (\rho, (1-\rho)) \cdot \frac{\pi}{2}, && \text{(IC)} \end{aligned}$$

where r^π is the expected compensation that the principal pays the academic.

[7] Formally, the problem is:

$$\begin{aligned} &\arg \max_{a,r} \beta a_1 + (1-\beta) a_2 - r \\ &\text{such that } u(e(a), r) \geq 0 && \text{(PC)} \end{aligned}$$

[8] The problem is concave with corner solutions at $a^0 = (0,0)$ for $\pi \leq 0$ and $(\rho, (1-\rho))$ for $\pi \geq 2$.

[9] Without loss of generality, attention can be limited to $\pi \in [0,2]$.

By decreasing the salary $\underline{\pi}$, the principal can reduce the expected compensation r^{π} until the agent is not worse off:

$$r^{\pi} = \left(\rho^2 + (1-\rho)^2\right)\left(\frac{\pi}{2}\right)^2. \tag{3}$$

Mathematically, the problem of finding the optimal incentives is equivalent to minimizing agency costs, i.e., the principal's loss from being unable to directly contract on the agent's action:[10]

$$min_{\pi\in[0,2]} \underbrace{b(a^*) - r^* - (b(a^{\pi}) - r^{\pi})}_{=:\alpha^{\pi}}, \tag{4}$$

where a^{π} is the agent's choice given premium π from equation (2) and r^{π} is the expected compensation from equation (3) that ensures the agent's participation.

Notice that the role of 'dysfunctional behavior' for finding optimal incentives is neither apparent from this minimization program nor from the principal's maximization program on page 8.

2.3 Defining gaming and shirking

The definition of gaming proposed here combines the term 'gaming' that was first pioneered in economics by Baker (1992) with the notion by Holmström and Milgrom (1991) that some resource can be badly allocated. The resource to be allocated here is effort, or equivalently, the academics's displeasure from an action choice, $e(a) = a_1^2 + a_2^2$.

Effort e is distinct from the agent's activity choice a which differs from Holmström (1979) who finds it 'convenient to think of a as effort' and uses effort interchangeably with action. If effort and action were the same, then discussing to which action a effort is allocated would not be meaningful. For reasons that will become clear, the resource is also not taken to be 'total effort' or total work time, $t(a) = a_1 + a_2$ as in Holmström and Milgrom (1991), Bond and Gomes (2009) or, more recently, in Sliwka and Manthei (2013).

Any incentive premium π that induces an action choice a^{π} different from the

[10] This follows from replacing the participation constraint by the respective choices of r^{π} and r^*

initial choice a^0 requires a positive effort from the academic. This effort can be computed as:

$$e^{\pi} := e(a^{\pi}) = (a_1^{\pi})^2 + (a_2^{\pi})^2 = \left(\rho^2 + (1-\rho)^2\right) \cdot \left(\frac{\pi}{2}\right)^2. \tag{5}$$

While incentives induce effort, i.e., create displeasure, this displeasure may or may not be associated with a benefit for the principal.

The action choice a^* that would have generated the largest benefit to the principal given an effort level of e^{π} is:[11]

$$a^*(e^{\pi}) = (\beta, 1-\beta) \cdot \frac{\sqrt{e^{\pi}}}{\sqrt{\beta^2 + (1-\beta)^2}}. \tag{7}$$

Whenever effort e^{π} is used for an action a^{π} different from $a^*(e^{\pi})$, the dis-utility incurred by the academic is at least partially 'wasted'. Switching to action $a^*(e^{\pi})$ would have been more beneficial to the principal but equally displeasing to the agent. The extent of 'waste' can be computed by comparing the actual benefit, $b(a^{\pi})$, with the hypothetical benefit had the induced effort e^{π} been used in the most beneficial way, $b(a^*(e^{\pi}))$. These ideas motivate the following definition.

Definition 1. *The academic games incentives π whenever he exerts effort $e^{\pi} > 0$ and uses it for an action a^{π} that does not employ this effort in the most beneficial way: $a^{\pi} \neq a^*(e^{\pi})$. Gaming costs amount to the loss from effort not being used in the most beneficial way: $b(a^*(e^{\pi})) - b(a^{\pi})$.*

In order to show that this definition might capture what is referred to as 'seemingly dysfunctional behavior', I want to use it to formalize the claim by Baker et al. (1994) that such behavior could have been prevented by aligning incentives and

[11] The choice $a^*(e^{\pi})$ that maximizes the principal's benefit among all choices that require the same effort e^{π} is:

$$a^*(e^{\pi}) = \arg\max_a \beta a_1 + (1-\beta)a_2 \text{ such that } a_1^2 + a_2^2 = e^{\pi}. \tag{6}$$

Since $\beta \in (0,1)$, the benefit-maximizing way of using effort e^{π} is unique and can be found by solving for a_2 in the side-constraint of (6), substituting a_2 in the objective function, and determining a_1^* from the respective first-order condition. This approach captures all solutions because $a_1 \geq 0$ and $a_2 \geq 0$.

rewards, or, as they put it 'basing pay on an employee's contribution to firm value'. For this, a definition of alignment is needed.

Definition 2. *Incentives are aligned with the benefit if (i) publication success is rewarded,* $\pi > 0$*, and (ii) the relative weight that referees assign to research and marketing also describes the relative importance to the principal,* $\rho = \beta$.

According to this definition, the iso-benefit lines, which describe all action choices leading to the same benefit for the principal, and the iso-reward lines, which describe all action choices that lead to the same expected reward, are literally aligned; they have the same slope. The condition $\pi > 0$ then ensures that the respective better sets are also equivalent.

Proposition 1. *The academic games incentives* $\pi \neq 0$ *unless they are aligned with the principal's benefit.*

Proof. The actual use a^π of effort e^π from (2) coincides with the choice that maximizes the principal's benefit from (7) if and only if:

$$a^\pi = a^*(e^\pi) \overset{(2),(5),(7)}{\Longleftrightarrow} (\rho, (1-\rho))\frac{\pi}{2} = (\beta, (1-\beta))\sqrt{\frac{\rho^2 + (1-\rho)^2}{\beta^2 + (1-\beta)^2}}\frac{\pi}{2}. \tag{8}$$

If incentives are aligned, $\rho = \beta$, then $a^\pi = a^*(e^\pi)$ and there can be no gaming. If there is no gaming, either (i) $e^\pi = 0$ or (ii) $e^\pi > 0$ and $a^\pi = a^*(e^\pi)$. Since $\pi \neq 0$, (i) can be ruled out and hence $e^\pi > 0$ by(5) and $\pi > 0$. Then, equation 8 implies $\rho = \beta$ and incentives are aligned. □

The specific definition of gaming proposed here thus offers a precise description in what sense the response to misaligned incentives is 'dysfunctional' or why 'rewarding for A while hoping for B' is 'foolish' (Kerr, 1975). It also provides a formal justification for the use of balanced scorecards, a device employed in practice to align incentives (Kaplan and Norton, 1996) but only if the aim is to avoid gaming.[12]

As an alternative to the suggested definition, one might consider the most beneficial use of the academic's total working time, $t(a) = a_1 + a_2$. Aligning

[12]For an intriguing alternative justification, see Gibbons and Kaplan (2015).

incentives with the benefit, however, does not ensure that the academic uses this time to maximize the principal's benefit—see Proposition 7 in the Appendix.

A natural counterpart to the loss from badly used effort are the costs to the principal because the academic may shirk and choose an action that requires less effort than the first-best choice. On the one hand, there is the principal's benefit loss that results from any deviation of the first-best from the actual effort e^* even if effort were used optimally: $b(a^*) - b(a^*(e^\pi))$. On the other hand, this loss has to be corrected for any gains because compensation in the first-best r^* may exceed that under incentives r^π, $r^* - r^\pi$. Combining both terms yields the shirking costs for incentives π, which reflect the losses due to shirking in comparison with the first best:

$$S^\pi = b(a^*) - b(a^*(e^\pi)) - (r^* - r^\pi)\,.$$

2.4 Gaming and optimal incentives

This section shows that the standard approach of finding optimal incentives, as described above, implicitly accounts for gaming.

Consider any incentives with premium π (whether optimal or not) that use the minimal compensation r^π, lead to an action a^π, and hence require effort e^π from the academic. Then, the benefit that would have been created had effort e^π been used optimally for $a^*(e^\pi)$ is $b(a^*(e^\pi))$. We can use this benefit as a benchmark to decompose the agency costs of incentives π.

Proposition 2. *The costs from running incentives (agency costs) are composed of shirking and gaming costs:*

$$\alpha^\pi = \underbrace{b(a^*) - b(a^*(e^\pi)) - (r^* - r^\pi)}_{\text{shirking costs } S^\pi} + \underbrace{b(a^*(e^\pi)) - b(a^\pi)}_{\text{gaming costs } G^\pi}.$$

For a graphical depiction of the decomposition, see Figure 1. This decomposition reveals for the first time that designers using the standard approach of finding optimal incentives by minimizing agency costs implicitly care about gaming because it contributes to them.

The decomposition can be exploited to identify a trade-off that underpins the optimal choice of the success premium.

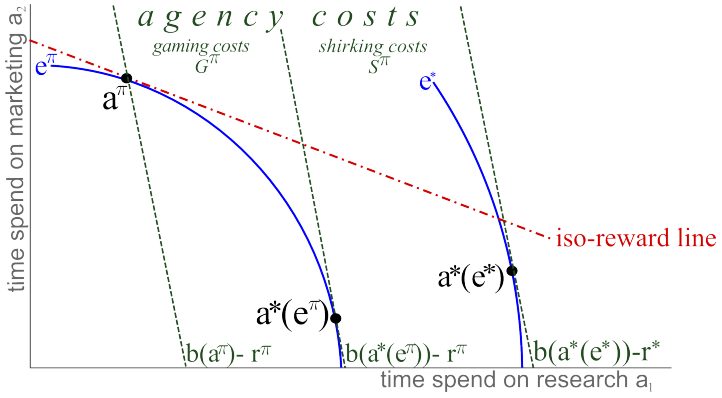

Figure 1: The costs of incentives π (agency costs) are due to gaming (effort e^{π} being misdirected to a^{π} instead of $a^*(e^{\pi})$) and shirking, (effort e^{π} instead of e^*).

Proposition 3. *For finding the optimal success premium, π^*, the principal has to weigh the marginal gains from reduced shirking costs against the marginal loss from larger gaming costs:*

$$-\left.\frac{dS^{\pi}}{d\pi}\right|_{\pi=\pi^*} = \left.\frac{dG^{\pi}}{d\pi}\right|_{\pi=\pi^*}.$$

Proof. The optimal success premium minimizes agency costs α^{π}. Since α^{π} is convex in the success premium, the optimal premium can be found by the first-order condition: $\left.\frac{d\alpha^{\pi}}{d\pi}\right|_{\pi=\pi^*} = 0$, or equivalently, $\left.\frac{dS^{\pi}}{d\pi}\right|_{\pi=\pi^*} + \left.\frac{dG^{\pi}}{d\pi}\right|_{\pi=\pi^*} = 0$. □

The principal increases the success premium until marginal gains from less shirking are exactly offset by the marginal losses from more gaming—see Figure 2. In our example, where shirking costs decrease faster than gaming costs increase, the optimal premium is positive. In other words, optimal incentives are gamed.

In this model, the gaming-and-shirking trade-off determines whether incentives are high-powered (large π) or low-powered (small π). Hence, the gaming-shirking trade-off answers the same question as the well-known incentive-insurance trade-off in the more traditional moral-hazard model but now for a moral-hazard model where the agent, here the academic, faces a multi-dimensional decision.

In previous proposals to explain the choice of optimal incentives, whether single- or multitasking, insurance issues play a crucial role, either in the form of risk-attitude or limited liability (Holmström, 1979; Holmström and Milgrom, 1991; Datar et al., 2001; Feltham and Wu, 2000; Baker, 2000, 2002). In contrast, these issues are not essential here: the gaming-shirking trade-off is operational although

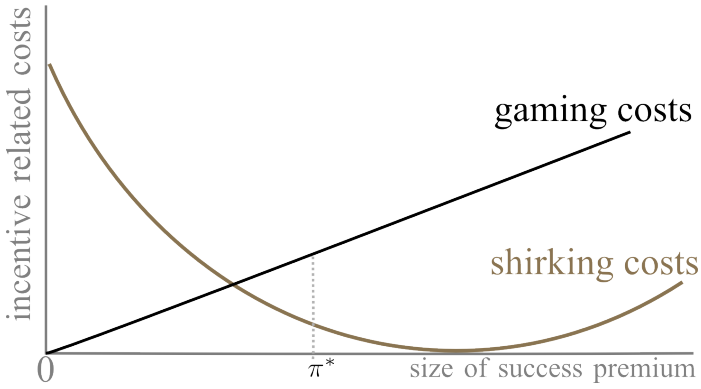

Figure 2: For the optimal publication premium, π^*, the marginal gains from preventing shirking have to be weighed against marginal losses from more gaming.

academic and principal are assumed to be risk-neutral and their liability is not limited.

Insurance issues could easily be incorporated into the model by limiting the academic's liability. Then, more compensation would be needed and shirking costs would increase.[13] Still, lowering shirking costs would come at the price of higher gaming costs. This suggests that the trade-off between both costs is of a more fundamental nature; a suggestion that is later confirmed in Corollary 2.

2.5 Gaming and suppressed incentives

This section uses the specific definition of effort as 'displeasure' to formalize the vague notion that badly allocated effort is the reason why performance information is not used to provide incentives.

Consider a variation of the simple multitasking model in which marketing is harmful to the principal, $1-\beta<0$. 'Marketing' could, for example, just be a euphemism for bad scientific conduct such as forging significance levels.[14]

In this variation of the model, focusing all effort on research maximizes the

[13] If $\underline{\pi}\geq 0$ and $\pi\geq 0$, the principal would optimally set $\underline{\pi}=0$ and the expected transfer would double: $r^{\pi}=\frac{(\pi)^2}{2}\left(\rho^2+(1-\rho)^2\right)$. Gaming costs would be unaffected while shirking costs would be: $S^{\pi}=\frac{\beta^2+(1-\beta)^2}{4}-\left(\frac{\pi}{2}\sqrt{\rho^2+(1-\rho)^2}\sqrt{\beta^2+(1-\beta)^2}-\frac{(\pi)^2}{2}\left(\rho^2+(1-\rho)^2\right)\right)$, leading to a U-shape function very similar to that in Figure 2.

[14] A rather dramatic example is that of Ulrich Lichtenthaler who was ranked best German-speaking researcher in business economics in 2005 before retracting 16 of his already published articles, for example, because significance levels were wrongly reported—see retractionwatch.com/2014/06/16/ulrich-lichtenthaler-retraction-count-rises-to-16, accessed on 6th of March 2015.

principal's benefit. $a^*(e^\pi) = \left(\sqrt{e^\pi}, 0\right)$ and $a^*(e^*) = \left(\sqrt{e^*}, 0\right) = \left(\frac{\beta}{2}, 0\right)$. A publication premium of π, however, continues to induce: $a^\pi = \left(\rho\frac{\pi}{2}, (1-\rho)\frac{\pi}{2}\right)$ because neither the referee's nor the academic's preferences have changed. If referees are very susceptible to marketing, the damage caused by gaming can be so large that it is better not to use incentives.

Corollary 1. *Suppose that marketing is damaging to the principal,* $1-\beta < 0$. *Then, performance information is not used to provide incentives because of gaming whenever the referee places too little weight* ρ *on research:*

$$\pi^* = 0 \text{ if and only if } \left.\frac{dG^\pi}{d\pi}\right|_{\pi=0} \geq \left|\left.\frac{dS^\pi}{d\pi}\right|_{\pi=0}\right| \Leftrightarrow \rho \leq \frac{\beta-2}{2\beta-1}.$$

Proof. Since agency costs are convex in the publication premium, a positive publication premium cannot minimize them iff $\left.\frac{d\alpha^\pi}{d\pi}\right|_{\pi=0} \geq 0 \Leftrightarrow \left.\frac{dG^\pi}{d\pi}\right|_{\pi=0} \geq \left|\left.\frac{dS^\pi}{d\pi}\right|_{\pi=0}\right|$, which by Lemma 3 in Appendix A is true if and only if $\rho \leq \frac{\beta-2}{2\beta-1}$. □

Like in the home contractor model by Holmström and Milgrom (1991), performance information is optimally ignored. Rather than vaguely attributing this observation to the allocation of 'time', 'effort' or 'attention', it is explicitly linked to the proposed specific definition of gaming.

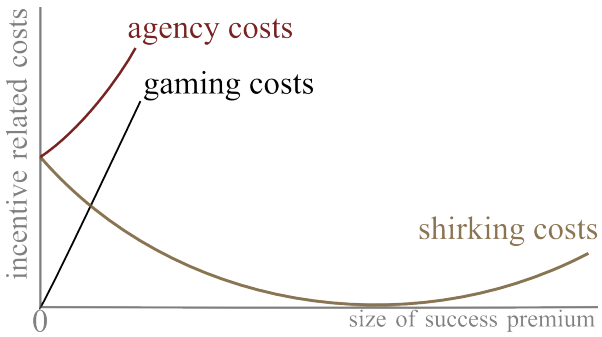

Figure 3: If marketing is very detrimental to the principal's benefit (e.g. because it means that the academic fabricates results), rewarding publication success increases gaming costs by more than it reduces shirking costs and using no incentives is optimal.

Incentives are not used here although various assumptions were deliberately chosen to differ from the Holmström and Milgrom (1991). First, the academic was not intrinsically motivated. Second, the two tasks were not perfect substitutes from

the academic's perspective (Instead, he preferred a mixed use of his work time). Third, the success signal reflected both dimensions. Finally, the second dimension (forging results) was detrimental to rather than essential for generating a benefit. The specific assumptions of the home contractor model are thus not necessary to explain why incentives are suppressed. Indeed gaming may prevent the use of performance information much more generally (see Corollary 3, later).

3 Generalization

This section takes the insights from the simple multi-tasking model without insurance problems to a very general framework that allows for insurance to matter and captures most moral-hazard models in which a principal (she) uses rewards to influence the choice of an agent (he).

3.1 General Model

Consider an agent who can take a course of action that affects some principal.

Agent's choice and preferences. The agent chooses an action a from some (separable metric) space $\mathcal{A}$. The principal controls the amount r of some reward good and the agent prefers having more of this good. Preferences over actions and rewards (a,r) are complete, transitive, continuous, and separable in a and r and strongly monotonic in r. Separability means that agent's preferences over action choices a do not depend on the level r of the reward good. This assumption is typically imposed in moral hazard models—even in the very general treatments of the problem by Gjesdal (1982) or Grossman and Hart (1983).[15] While this excludes, for example, that the academic's relative preference for research and marketing changes as he gets richer, it does allow for income effects, e.g., the agent may prefer to exert less effort when he gets richer.[16]

Agent's effort and utility. Given separability, the agent's preferences over actions can be represented using a (continuous) effort function $e: \mathcal{A} \to \mathbb{R}$, where

[15] See Assumption A1 in Grossman and Hart (1983) and Assumption 4 in Gjesdal (1982) in Section 2-4; a notable exception is his Section 5, where this assumption is dropped.

[16] For an example, see Hermalin (1992) who features income effects while agent's preferences are additively separable in reward and action.

$a \mapsto e(a)$.[17] For simplicity (but slightly abusing notation), e is also used to refer to the real number describing the agent's effort for a given action. Using the effort function, preferences over actions and rewards together can be represented by a continuous (real-valued) utility function $u : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with $(e,r) \mapsto u(e,r)$ that is strictly falling in e and strictly increasing in r. Standardize the utility when the agent does not engage with the principal to zero: $u(0,0) = 0$.

Principal's benefit and utility. The principal cares about the agent's action choice and dislikes giving up the reward good. Her preferences over (a,r) are complete, transitive, continuous and separable in a and r,[18] where r is a 'bad': she strictly prefers rewarding less. Preferences over action choices can be represented by a (continuous) benefit function $b : \mathcal{A} \to \mathbb{R}$, where $a \mapsto b(a)$. Using the benefit function, preferences over both, action choices and rewards, can be expressed by a (real-valued) continuous utility function $v : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, where $(b,r) \mapsto v(b,r)$ that is strictly increasing in b and strictly falling in r. In many applications, the benefit does not only depend on the agent's choice but also on luck, i.e., factors beyond the control of agent or principal. This could be captured by defining benefit as a function of a as well as some random variable. In order to keep notation simple, the principal's benefit here is written as a function of the action, only.[19]

Incentives. Incentives are a credible promise of a distribution of rewards by the principal for each action that the agent engages in. The credibility of the principal's promise may result from an explicit contract, which is enforced by a court of law, or from a a relational contract, which is enforced in a repeated interaction in the tradition of MacLeod and Malcomson (1989, 1998), Levin (2003), and Halac (2012). Formally, incentives are represented by a function I: $\mathcal{A} \to \mathcal{C}$, where $\mathcal{C}$ is the set of cumulative distribution functions. The function maps a given choice a by the agent to a cumulative distribution function of rewards: $a \mapsto F^I(r|a)$. This distribution can be degenerate in order to reflect that the agent has full control over rewards through his action choice.

[17] This follows from Theorem 1 in Bergstrom (2015).

[18] This assumption holds in almost all moral-hazard models. Indeed preferences are often assumed to be additively separable. A notable exception is the model by Raith (2008).

[19] In a model, where benefits were random, $b(a)$ can be regarded as the principal's certainty equivalent of the lottery generated by choice a. The existence of the certainty equivalent can be ensured by an argument similar to that in footnote 21.

Contractual Environment. Denote by $\mathcal{J}$ the set from which the principal can choose incentives: $I \in \mathcal{J}$. This set describes the contractual environment and is typically restricted, i.e., not all mappings from action choices to distributions are possible. The academic could, for example, only be rewarded on the basis of publications and was more likely to receive π if he spent more time researching. Being precise about the contractual environment is crucial in order to describe which incentives are feasible and is hence necessary for finding optimal incentives. As will become clear, the following insights on the fundamental forces and trade-offs of incentive design do not depend on the details of the contractual environment.

Induced behavior. Suppose that incentives $I \in \mathcal{J}$ result in some behavior, or equivalently, that the agent's utility maximization problem given incentives I has a solution. Denote the choice by the agent given incentives I by a^I and the real number describing the effort required for this choice by $e^I := e(a^I)$.[20] The agent's choice a^I results in a cumulative distribution $F^I(r|a^I)$. The certainty equivalent of this distribution from the principal's perspective is denoted by r^I, where r^I is defined implicitly by $v(b(a^I), r^I) = \int v(b(a^I), r) dF^I(r|a^I)$.[21] Practically, r^I describes the principal's costs for offering the reward distribution associated with incentives I to the agent when the latter chooses a^I.

First-best. Suppose that any promise by the principal to reward a behavior were credible. Denote some Pareto-optimal choice in this case by a^* and the associated reward by r^*. Formally, a^* and r^* are a solution to $\max_{a,r} v(b(a), r)$ such that $u(e(a), r) \geq 0$.[22] Denote the first-best effort by $e^* := e(a^*)$. The first-best serves as a benchmark that may or may not be reached with incentives $I \in \mathcal{J}$.

Agency costs. The agent never loses from the introduction of incentives (otherwise his participation constraint would not be met).[23] Agency costs, the loss α^I to the principal from running an incentive scheme I rather than stipulating the agent's

[20] If incentives leave the agent indifferent between multiple action choices, a^I is assumed to be a choice suggested by the principal.

[21] The existence of r^I follows from continuity of v in r and the intermediate value theorem for integrals. Uniqueness from the fact that v is strictly increasing in r.

[22] Existence of a^* and r^* can be ensured as follows. For any given $\tilde{a}$, there is a $\tilde{r}$ such that $u(e(\tilde{a}), \tilde{r}) = 0$ because of continuity of u. Since v decreases and u increases in r, the maximizer (a^*, r^*) has to come from the set $\{(\tilde{a}, \tilde{r}) | u(e(\tilde{a}), \tilde{r}) = 0\}$, which is compact because it is the domain of continuous functions mapping to the compact set $\{0\}$. Since v is a continuous function on that compact set, it has a maximum.

[23] If the participation constraint does not bind, the agent gains from the introduction of incentives.

course of action in a contract, can be written as

$$\alpha^I = v(b(a^*), r^*) - v(b(a^I), r^I). \tag{9}$$

Scope. The framework covers most hidden action problems—even those that were originally not placed in a principal-agent setting such as Averch and Johnson (1962).[24] Since the framework embeds single as well as multitasking moral-hazard models as special cases, it can be used to tease out whether a multi-dimensional action space is necessary for discussing dysfunctional behavior as suggested by the Nobel Prize Committee (2016).

3.2 Choosing and using Effort

The agent's ability to state preferences over actions independently from the level of rewards, which can then be represented by $e(a)$, allows us to distinguish between two aspects of the agent's choice. The agent decides, on the one hand, on how much effort e he exerts, and, on the other hand, on how this effort is used. Formally, the agent's choice set can be (dis-jointly) decomposed into subsets that require the same effort e: $\mathcal{A} = \bigcup_e \{a|e(a) = e\}$, where different effort levels e are associated with distinct indifference sets $\{a|e(a) = e\}$. 'Choosing effort' means selecting one of these indifference sets and 'using effort' means selecting an action from this set. This notion of 'choosing' and 'using' effort offers a formally precise interpretation of what Raith (2008) refers to as 'how much' and 'what' the agent does.

We can employ the idea of 'using effort' to distinguish between two model classes in the literature. Since any action choice is associated with a different effort level, the early moral-hazard literature (see e.g. Gjesdal, 1976; Holmström, 1979, 1982; Kim, 1995) but also Baker (1992) are special cases of the presented framework which are only concerned with the agent's choice of effort. Multitasking models such as Holmström and Milgrom (1991) and the ensuing literature in accounting and labor economics (see e.g. Feltham and Xie, 1994; Datar et al., 2001; Feltham and Wu, 2000; Baker, 2000, 2002; Schnedler, 2008; Christensen et

[24] The regulator in their model is the principal here; the monopolist is the agent; the choice of labor and capital intensity is the action choice; production costs are the effort; and the social gains from production are the benefit.

al., 2010) are special cases in which the effort choice does not fully determine the agent's action choice. In these models, the agent chooses effort and decides how to use it.

3.3 Gaming and dimensionality of action space

Let us define the set of benefit-maximizing ways of 'using effort' e:

$$A^*(e) := \arg\max_a b(a) \text{ such that } e(a) = e.$$

This set is not empty because $b(\cdot)$ is a continuous function and the set $\{a|e(a) = e\}$ is compact, which in turn follows from $e(\cdot)$ being a continuous function. Refer to some element from $A^*(e)$ as $a^*(e)$.

Definition 3 (Gaming)**.** *The agent games incentives if he uses the effort induced by incentives $e^I = e(a^I) > 0$ in a way that does not maximize the principal's benefit—although he would not care on where he directs effort in the absence of incentives: $a^I \notin A^*(e^I)$.*

Next, I introduce a measure of the extent of the negative consequences of gaming by comparing the benefit actually obtained with incentives I, $b(a^I)$, to the benefit that the principal would have obtained had effort e^I been used Pareto-optimally, $b(a^*(e^I))$, where $b(a^*(e^I))$ is well-defined because all courses of action $a^*(e^I) \in A^*(e^I)$ entail the same benefit. For this comparison, the certainty equivalent of rewards r^I is held constant.

Definition 4 (Gaming costs)**.** *The gaming costs of incentives I are:*

$$G^I := v(b(a^*(e^I)), r^I) - v(b(a^I), r^I).$$

Gaming costs are never negative because the benefit from optimally used effort cannot be exceeded $b(a^I) \leq b(a^*(e^I))$. They are zero if and only if effort is used in the most beneficial way, $a^I \in A^*(e^I)$. For example, if there is a unique way of using any given effort e, $\left|\{a\middle|e(a) = e\}\right| = 1$, the agent's choice is by definition the most beneficial one, he cannot game incentives and gaming costs are zero. This is, for example, the case in traditional (single-task) moral-hazard models, where

the agent dislikes larger action choices. In extant multitasking models, there are multiple ways to use effort, $\left|\{a\middle|e(a)=e\}\right|>1$, and gaming becomes possible.

This may wrongly suggest that a multi-dimensional action space is necessary for gaming. For a counter-example, where gaming occurs according to the definition but the choice space is single-dimensional, consider a nurse who decides on how many minutes $a\in\mathbb{R}^+$ to spend with a patient. Suppose that this nurse is intrinsically motivated to spend a^0 minutes with the patient. Spending more or less time creates displeasure $e(a)=(a-a^0)^2$. The patient, on the other hand, benefits from more time with the nurse: $b(a)<b(\tilde{a})$ for $a<\tilde{a}$. Suppose that all that can be observed and used in contracts is whether the time spent is below or above some threshold κ, where $\kappa<a^0$. Then, a (rather dysfunctional) incentive scheme I that pays the nurse a sufficiently large premium for doing *at most* κ would induce exactly $a=\kappa$ and require effort $e(a)=(\kappa-a^0)^2$. The same effort could have been used for $\tilde{a}=a^0+a^0-\kappa$ because $e(\tilde{a})=(a^0+a^0-\kappa-a^0)^2=e(a)$. Observe that $\kappa<a^0$ implies $a=\kappa<a^0+a^0-\kappa=\tilde{a}$, so that using effort for $\tilde{a}$ rather than a would have been more beneficial for the patient. The agent thus games the dysfunctional incentives I according to our definition: he exerts effort $e(a)=(\kappa-a^0)^2>0$ but does not use it for the most beneficial action $\tilde{a}$. A multi-dimensional choice space is thus not necessary for gaming.

A multi-dimensional choice space is not sufficient for gaming incentives, either. Even if his choice is multi-dimensional, the agent cannot game incentives if effort is a one-to-one function of the action choice. Consider a sales agent who can inform himself about the customer's preferences and be particularly nice to her: $\mathcal{A}=\{\text{don't inform},\text{inform}\}\times\{\text{not nice},\text{nice}\}$. Suppose he finds being nice easier than informing himself: $e((\text{don't inform},\text{not nice}))<e((\text{don't inform},\text{nice}))<e((\text{inform},\text{not nice}))<e((\text{inform},\text{nice}))$. The resulting mapping between action choices and effort is one-to-one, so that effort can only be used in one way, and this choice is by definition the most (as well as the least) beneficial.[25]

Summarizing the insights from the two counter-examples results in the following lemma.

[25]Even for a multidimensional but not countable action set, say $\mathcal{A}=\mathbb{R}\times\mathbb{R}$, effort may theoretically uniquely determine the action because $a\in\mathbb{R}\times\mathbb{R}$ can be mapped one-to-one to $\mathbb{R}$ using Peano curves. However, I am not aware of any economically application in which such a relationship would be meaningful.

Lemma 1. *A multi-dimensional choice space of the agent is neither sufficient nor necessary for the agent to game incentives.*

Recall the claim by the Nobel Prize Committee (2016) that a multi-dimensional choice is necessary to formalize arguments on 'dysfunctional behavior' and its effect on incentives. This would suggest that the definition of gaming proposed here is unsuitable to formalize such arguments. In the following, I want to examine the two most prominent arguments about 'dysfunctional behavior', namely, (i) that it can be avoided by aligning incentives with the benefit and (ii) that it is the reason why incentives are suppressed. I will then show that both arguments can be formalized with the definition of gaming although it does not necessitate a multi-dimensional action space.

First, however, I turn to the second key aspect addressed by incentives in the example earlier, namely, shirking.

3.4 Shirking and its costs

Incentives are only needed because the agent does not like the behavior that the principal wants to implement; it requires effort and he wants shirk and settle on an action choice with less effort which is in his better-set. Incentives may thus not only be imperfect in the sense that the agent misuses effort but also in the sense that the agent shirks, $e^I < e^*$, or that the compensation necessary to prevent shirking exceeds that in the first-best, $r^I > r^*$. The loss for given incentives I due to shirking can be isolated by eliminating the effect of gaming and using the utility of the principal *had the agent used effort in the most-beneficial way* and subtracting it from her utility in the first-best.

Definition 5 (Shirking costs)**.** *The costs of eliciting effort e^I with incentives I, or in short, shirking costs, are:*

$$S^I := v(b(a^*(e^*)), r^*) - v(b(a^*(e^I)), r^I).$$

Shirking costs are never negative because $I \in \mathcal{I}$ imposes a restriction on the choices of effort e^I and reward r^I, while e^* and r^* maximize the principal's utility without

such a restriction.[26]

3.5 Aligned incentives and gaming

For examining whether aligning incentives with the benefit prevents gaming, the respective notion from the academic example needs to be generalized. Fundamentally, alignment in this example meant that larger rewards are associated with larger benefits. Since the link between the agent's action and rewards is not necessarily deterministic anymore, the idea of 'larger rewards' has to be interpreted in some stochastic sense. This is exactly what the following definition does.

Definition 6 (Aligned Incentives)**.** *Incentives I are aligned (with the principal's benefit) if a change in the agent's action choice that increases the principal's benefit also (weakly) increases the agent's rewards (in the sense of first-order stochastic dominance):*

$$\textit{for all } a, \tilde{a} \in \mathcal{A} \textit{ with } b(a) > b(\tilde{a}) \textit{ and for all } r: \quad F^I(r|a) \leq F^I(r|\tilde{a}).$$

What is attractive about alignment is that it can be checked by examining the relationship between rewards and benefits, both of which are typically known to a principal who intends to use incentives.[27] With this definition in place, the proposed notion of gaming still supports the claim by Baker et al. (1994) that alignment prevents 'seemingly dysfunctional responses'.

The intuition is simple. Alignment implies that the agent cannot obtain larger rewards by decreasing the principal's benefit. Absent incentives, the agent is indifferent when choosing between actions that require the same effort. His choice is thus entirely driven by rewards. The most rewarded choice, however, is also the most beneficial one if incentives and benefits are aligned.

Conversely, if effort is employed most beneficially, benefit and rewards have to be aligned 'locally'; otherwise, the agent could profitably deviate to another action

[26] Notice that the benefit may well be smaller $b(a^*(e^*)) < b(a^*(e^I))$ but then $r^* < r^I$.

[27] There are, of course, plenty of examples in which incentives seem aligned before they are put in place but then turn out to be grossly misaligned: the legendary Havard Case about Lincoln Electric's secretaries who upped their pay by unnecessary hitting keys on their typewriters (Berg, 1983), Sears policy of reducing base pay and increasing commissions which led to unnecessary repairs (Paine, 1994), or rewards for test scores that led to cheating Jacob and Levitt (2003).

that requires the same effort. The principal can, of course, only infer whether or not the agent has reason to deviate if she knows the agent's preferences. If she lacks this information, gaming can only be avoided if rewards are higher whenever benefits are larger: incentives have to be aligned with the benefit. This intuition is behind the following proposition.

Proposition 4 (Aligned Incentives and Gaming)**.** *(i) Incentives that are aligned with the benefit cannot be gamed. (ii) For gaming to be prevented irrespective of the agent's preferences, incentives have to be aligned with the benefit.*

Proof. Part (i) works by contradiction. Suppose that aligned incentives are gamed: the agent chooses an action $\tilde{a}$ rather than a, although $b(a) > b(\tilde{a})$, while $e(a) = e(\tilde{a})$. Then, $F^I(r|a) \leq F^I(r|\tilde{a})$, because incentives are aligned. Consequently, $\int u(e^I, r)dF^I(r|a) \geq \int u(e^I, r)dF^I(r|\tilde{a})$ and the agent does not lose from deviating to action choice a. This, however, contradicts the assumption that the agent chooses $\tilde{a}$ rather than a.[28] Hence, the assumption that incentives are gamed cannot be true.

Part (ii). Suppose there are two arbitrary action choices, a and $\tilde{a}$, with $b(a) > b(\tilde{a})$ and consider preferences such that $e(a) = e(\tilde{a}) = e^I$. Effort is allocated to generate the largest benefit if and only if the agent cannot profitably deviate from a to $\tilde{a}$. This means $\int u(e^I, r)dF^I(r|a) \geq \int u(e^I, r)dF^I(r|\tilde{a})$ for any u that increases in r, which is equivalent to $F^I(r|a) \leq F^I(r|\tilde{a})$. In summary, for any $a, \tilde{a} \in \mathcal{A}$ with $b(a) > b(\tilde{a})$, one gets $F^I(r|a) \leq F^I(r|\tilde{a})$, which is the definition of alignment. □

So, even though the definition of gaming here applies beyond multi-tasking, it can still be used to formalize the claim that rewarding according to 'firm value' can avoid 'seemingly dysfunctional behavior' (Baker et al., 1994). As seen before, the notion of total working time is generally not suited to formalize the idea that aligning benefit and incentives prevents 'seemingly dysfunctional responses'. An exception are models in which total working time and effort coincide because the agent dislikes working but does not care on how he spends his time—see Holmström and Milgrom (1991) or Bond and Gomes (2009) for examples.

Before examining whether the claim that gaming prevents the use of performance information in incentive schemes can be formalized with the proposed

[28] Recall that in case of indifference the agent selects the action preferred by the principal—see footnote 20.

notion of gaming, I want to link gaming to agency costs and identify its role in shaping incentives in the more general framework.

3.6 Role of gaming when designing incentives

The idea that incentives are about preventing gaming and shirking carries over to the more general framework.

Proposition 5 (The Forces Determining the Value of Incentives)**.** *The principal's costs from running an incentive scheme (agency costs) are composed of shirking and gaming costs:* $\alpha^I = S^I + G^I$.

Proof. By definition, agency costs are: $\alpha^I = v(b(a^*(e^*)), r^*) - v(b(a^I), r^I)$. Subtracting and adding the utility from Pareto-optimal used effort, $v(b(a^*(e^I)), r^I)$, yields:

$$v(b(a^*(e^*)), r^*) - v(b(a^I), r^I) = \begin{array}{c} v(b(a^*(e^*)), r^*) - v(b(a^*(e^I), r^I) \\ + v(b(a^*(e^I), r^I) - v(b(a^I), r^I) \end{array} = \begin{array}{c} S^I \\ + G^I \end{array}.$$ □

Since the principal's aim is to minimize agency costs, she will automatically take gaming costs into account when pursuing this aim. An exception are, of course, situations in which all incentives entail the same gaming costs, for example, because they all induce the same action or because all incentives that the principal can choose from are aligned.

If incentives vary in gaming and shirking costs, the principal will reduce gaming costs as long as this is possible without increasing shirking costs. Eventually, all respective opportunities are exhausted and she faces a trade-off: gaming costs can only be reduced further if shirking costs are increased. This argument directly leads to the following corollary.

Corollary 2 (Effort Elicitation and Direction Trade-Off)**.** *Suppose incentives* $\hat{I}$ *minimize agency costs. Then, (i) lower gaming costs can only be achieved at the price of higher shirking costs: for all* $I \in \mathcal{I}$ *with* $G^I < G^{\hat{I}}$: $S^I > S^{\hat{I}}$. *(ii) Lower shirking costs are only possible by increasing gaming costs: for all* $I \in \mathcal{I}$ *with* $S^I < S^{\hat{I}}$: $G^I > G^{\hat{I}}$.

Proof. The proof for (i) works by contradiction. Take incentives that minimize agency costs $\hat{I}$ and suppose no trade-off exists. Then, for some I : $G^I < G^{\hat{I}}$ but $S^I \leq S^{\hat{I}}$. This, however, contradicts the fact that $\hat{I}$ minimizes agency costs. Claim (ii) can be proven completely analogously. □

This gaming-shirking trade-off underpins incentive design as long as gaming costs vary between schemes. A notable exception is the traditional moral hazard model, where gaming costs are zero for all feasible incentive schemes.

Next, use the decomposition idea to formalize that gaming is the reason why information about the agent's effort is optimally discarded.

3.7 Suppressed incentives and gaming

More information about effort ultimately means that eliciting effort becomes cheaper, which suggests the following general definition.

Definition 7. *A contractual environment $\tilde{\mathcal{J}}$ is more informative (about effort) than $\mathcal{J}$, if it permits incentives $\tilde{I}$ with lower shirking costs than any incentives in environment $\mathcal{J}$: for some $\tilde{I} \in \tilde{\mathcal{J}}$ and all $I \in \mathcal{J}$: $S^{\tilde{I}} < S^I$. The additional information about effort in $\tilde{\mathcal{J}}$ relative to $\mathcal{J}$ is discarded if the principal does not use such incentives $\tilde{I}$.*

This definition captures the notion of an 'informative' signal by Holmström (1979) as well as that of a 'more efficient' information system by Kim (1995)—see next section.

Using additional information about effort reduces shirking costs and hence agency costs. The only reason not to exploit a more informative environment are consequently increased gaming costs.

Corollary 3 (Optimally Discarding Performance Information)**.** *Discarding the information about effort in $\tilde{\mathcal{J}}$ relative to $\mathcal{J}$ is optimal if and only if the gains from lower shirking costs are more than outweighed by larger gaming costs: although there is some $\tilde{I} \in \tilde{\mathcal{J}}$, with $S^{\tilde{I}} < S^I$ for all $I \in \mathcal{J}$, for at least one $I \in \mathcal{J}$: $S^I - S^{\tilde{I}} < G^{\tilde{I}} - G^I$.*

Proof. Since $\tilde{\mathcal{J}}$ is more informative than $\mathcal{J}$, there is some $\tilde{I} \in \tilde{\mathcal{J}}$ with $S^{\tilde{I}} < S^I$ for all $I \in \mathcal{J}$. Any such incentives $\tilde{I} \in \tilde{\mathcal{J}}$ are optimally discarded if and only if $\alpha^{\tilde{I}} > \alpha^I$ for some $I \in \mathcal{J}$, or equivalently, if and only if $S^{\tilde{I}} + G^{\tilde{I}} > S^I + G^I$ for some $I \in \mathcal{J}$. □

Since the home contractor model by Holmström and Milgrom (1991) is a special case of the considered framework, this corollary justifies the verbal notion that incentives are suppressed in this model because of 'badly allocated effort'—as long as the latter is identified with the notion of gaming proposed here.

This section has shown that the value from exploiting a more informative environment is affected by gaming. The next section examines whether gaming is the reason why classical results on the value of information for designing incentive collapse outside the traditional single-task model.

4 Gaming and the Value of Information

The aim of this section is to show that the decomposition of agency costs into gaming and shirking costs is not arbitrary but exactly identifies why results on the value of information for incentives do not hold beyond the traditional moral-hazard model.

Since its infancy, the formal analysis of hidden action problems has been interested in the value of information (Gjesdal, 1976, 1982; Harris and Raviv, 1979; Holmström, 1979, 1982; Shavell, 1979). Probably one of the most important insights from the early moral hazard literature is that freely available and independent information about performance is valuable as long as it is not yet reflected in incentives. The intuition is that with such information more effort can be obtained without having to increase the agent's exposure to risk. The sufficient statistic result by Holmström (1979, 1982) prominently captures this idea. Gjesdal (1982) affirms the sufficient statistic result for a general class of models, including those where the agent's choice is multidimensional (see also Christensen et al., 2010).

The key criterion that Holmström (1979) introduces for the sufficient statistic result is informativeness. A signal Y is *informative* about the agent's choice a in relation to a signal already in use, say X, if the joint density of Y and X, $f(x,y,a)$, cannot be decomposed into an effect of the action a on X, $h(x,a)$, and 'noise',

$g(x,y)$: $f(x,y,a) = g(x,y) \cdot h(x,a)$—see condition (17) in Holmström (1979). Starting with optimal incentives using X, Holmström finds that there are incentives using also Y with strictly lower agency costs if and only if Y is informative about a given X (Proposition 3 in Holmström, 1979).

In Holmström (1979), the agent prefers smaller action choices a. We can thus identify the action choice a in his model with the effort choice e in the general framework examined here. His model then becomes a reduced form of this general framework that is only about choosing effort but does not explain how effort is used. This means that agency costs in his model become shirking costs in the more general framework. Accordingly, the sufficient statistic result becomes a statement about shirking costs rather than agency costs.

Corollary 4 (Sufficient Statistic and Shirking Costs)**.** *Under the assumption of Holmström (1979), there are incentives using a signal Y in addition to signal X with lower shirking costs than all incentives using X if and only if the signal Y is informative about effort e in relation to X (in the sense of Holmström, 1979).*

Proof. The proof follows directly from Proposition 3 in Holmström (1979) by identifying the agent's decision in his model with the choice of effort e here. □

The sufficient statistic result compares two sets of feasible incentives, one which includes incentives that condition on signal Y, say $\tilde{\mathcal{J}}$, and one which doesn't, $\mathcal{J}$. If the signal is informative in the sense of Holmström (1979; 1982), there are incentive schemes in $\tilde{I} \in \tilde{\mathcal{J}}$ with lower shirking costs than any incentives in $I \in \mathcal{J}$. Informativeness about effort in the sense of Holmström is thus a special case of a more informative environment as defined in the previous section. Applying Corollary 3 reveals that signals that are informative in the sufficient statistic sense are only valuable if the increase in gaming costs is negligible.

Holmström (1979, 1982) examines the value of an additional piece of information (or signal), or equivalently, he compares two nested information systems. In order to compare the relative noisiness of non-nested information systems, Kim (1995) defines an MPS criterion[29] and shows that less noisy signals according to

[29]Kim (1995) computes the likelihood ratio, which essentially describes how much probability mass is shifted by the agent's choice. He then compares the mean-preserving spread of the likelihood ratio distribution functions, which is different from the mean-preserving spread of the signal. In particular, less noise is associated with a *larger* spread of the distribution functions.

this criterion can be used to generate the same effort lower expected transfers r^I (see his Proposition 1).

Like Holmström (1979), Kim's model assumes that the agent prefers smaller action choices (see his Assumption 1) and we can regard his model to be a reduced version that is only about the agent's effort choice and not about the action to which effort is allocated. His result is thus only about the level and price of effort.

Corollary 5 (Signal Comparison & Shirking Costs)**.** *Under the assumptions of Kim (1995), optimal incentives using signal Y lead to lower shirking costs than for any incentives using signal $\tilde{Y}$ if Y is less noisy according to Kim's MPS criterion.*

Proof. The proof follows from Proposition 1 in Kim (1995) by identifying the agent's decision a in his model with effort e here and by observing that for given effort level e, lower expected compensation is equivalent to lower shirking costs because the benefit from optimally used effort $b(a^*(e))$ remains constant. □

Both corollaries generalize established criteria from the traditional moral hazard model, where each action is associated with a different cost, to the general framework, where the agent is allowed to be indifferent between choices, which includes multitasking settings like that of Christensen et al. (2010). The same approach can be used to translate any result about action choices and their consequences on agency costs in the traditional moral hazard model to effort choices and their consequences on shirking costs in the general model without having to re-analyze the complete problem as, for example, in Christensen et al. (2010).

The limits of any such generalization also become clear. Whatever criterion is proposed in the traditional model, it only applies to shirking costs in the general model. The ranking of signals (or information systems) can hence be overturned by gaming costs. In order to see this, consider the criterion that signal $\tilde{Y}$ is a more noisy version of Y about e in the sense of a Blackwell garbling. This criterion is sufficient for Y being more informative than $\tilde{Y}$ in the sense of Holmström (1979) if both systems are nested and for Y being less noisy than $\tilde{Y}$ according to Kim's criterion (1995)—see Gjesdal (1982). If $\tilde{Y}$ is a more noisy version of Y in the sense of a Blackwell garbling, the principal thus prefers Y to $\tilde{Y}$—see Proposition 2 in Gjesdal (1982) or Proposition 3 in Christensen et al. (2010).

The principal prefers Y to $\tilde{Y}$ because she can get *the same action* while saving on insurance costs. This logic assumes that the information system does not affect to which action the agent allocates effort or that it does not matter where he allocates effort. Once the system affects where effort is directed and the principal cares about this, the principal may prefer the more noisy signal about effort because the respective effort is better used.

Proposition 6 (More noisy but preferred)**.** *Even if a signal $\tilde{Y}$ is more noisy than Y in the sense of $\tilde{Y}$ being a Blackwell garbling of Y for e or in the sense of Kim's MPS criterion, the principal may still prefer $\tilde{Y}$ to Y.*

Proof. The proof works by counter-example and can be found in Appendix B. □

Taken together, the findings from this section show that gaming is the reason why the results by Holmström (1979) and Kim (1995) do not hold more generally. Again, the proposed specific definition of gaming is crucial.

If effort is defined as the sum of task-wise choices, the single-task results cannot be generalized. The reason is that more total time is not necessarily associated with larger costs for the agent. The choice of total time can thus not be identified with the effort choice in the single-tasking literature.

The notion of 'gaming' implicit in the interpretation of Baker (1992) cannot explain why information results fail, either. In Baker (1992), larger action choices entail larger costs for the agent, effort cannot be misdirected, so that gaming costs are zero. As a result, agency costs always amount to shirking costs and results from the single-tasking literature fully apply and are never overturned by 'gaming' in the sense of Baker.

5 Re-visiting Congruity-Precision-Trade-Off

This section uses the developed tools to shed light on why extant attempts to capture the trade-offs involved in incentive design under multi-tasking are problematic.

Feltham and Xie (1994); Feltham and Wu (2000); Baker (2000, 2002) stipulate that incentive designers have to trade-off the congruity (or congruence) between performance measure and benefit with the precision of the performance measure.

This implicitly assumes that *more* congruity reduces agency costs. On the other hand, Schnedler (2008, Proposition 2) shows that, holding precision constant, the lowest agency costs are achieved with a performance measure that over-emphasizes tasks that the agent likes.

The Section confirms that congruity may well be regarded as 'desirable' in the sense of entailing no gaming costs (Corollary 6). Consequently, less congruent performance measures must lead to lower agency costs because they involve less shirking (Corollary 7).

The congruity-precision trade-off has been proposed for the multitasking linear normal model, or short: LEN model, which assumes that performance measure, Y, and benefit, B, are linear in action choices $a = (a_1, \ldots, a_n)' \in \mathbb{R}^n$, and noise: $Y(a, \eta) = \mu_1 a_1 + \ldots + \mu_n a_n + \eta$, and $B(a, \varepsilon) = b_1 a_1 + \ldots + b_n a_n + \varepsilon$, where ε and η are normally distributed error terms. Moreover, rewards are assumed to be linear in the performance measure, $r(Y) = \underline{\pi} + \pi Y$, and effort to be quadratic in action choices: $e(a) = a'Ea$, with E being a positive-definite matrix.

A performance measure is said to be *congruent* if the relative effect of choices on this measure along any two dimensions ('tasks') is the same as on the benefit. Formally, the marginal effects vector $\mu = (\mu_1, \ldots, \mu_n)$ is a multiple of that of the benefit $b = (b_1, \ldots, b_n) : \mu = \lambda b$, for some $\lambda > 0$.[30] Otherwise, the performance measure is dis-congruent.

For a trade-off between congruity and precision to be meaningful, congruity must be attractive. It could, for example, prevent 'bad' choices by the agent. This idea can be justified by appealing to gaming. Rewarding the realization of a performance measure that is congruent with the benefit means that it is not possible to (stochastically) increase measured performance without increasing the benefit. In other words, incentives are aligned and entail no gaming (by Proposition 4). Moreover, using a dis-congruent performance measure induces a different and non-optimal use of effort. The following corollary summarizes these considerations.

Corollary 6. *Congruity is desirable in the LEN model in the sense that incentives I*

[30] A measure Y with $\lambda < 0$ measures 'bad performance' and can be turned into a congruent performance measure $\tilde{Y}$ with some positive λ by flipping the scale: $\tilde{Y} = -1 \cdot Y$. Without loss of generality, we can thus assume for a congruent measure that $\lambda > 0$.

are gamed unless the underlying performance measure is congruent:

$$G^I = 0 \Leftrightarrow \mu = \lambda b, \textit{ for some } \lambda > 0.$$

Proof. For the 'if' part, take any pair $a, \tilde{a}$, with $b(a) > b(\tilde{a})$. Using this and that the performance measure is congruent with the benefit $\mu = \lambda b$ for $\lambda > 0$ and rewarded ($\pi > 0$), one gets $\text{Prob}(\underline{\pi} + \pi\lambda b a + \eta \leq r) \leq \text{Prob}(\underline{\pi} + \pi\lambda b\tilde{a} + \eta \leq r)$ for some $\lambda > 0$. This, however, means that the distribution of rewards gets (weakly) stochastically larger when moving from $\tilde{a}$ to a. Incentives are thus aligned and by Proposition 4 incentives are not gamed.

For the 'only if' part, recall that e is strictly convex and b linear, so that b has a unique maximizer on $\mathcal{E} := \{a | e(a) = \underline{e}\}$, say $\hat{a}$, at which the derivative of b on $\mathcal{E}$ disappears. Now assume that performance is not rewarded congruently with the benefit, i.e., there is no $\lambda > 0$ such that $\mu = \lambda b$. Then, the derivative of μa on $\mathcal{E}$ at $\hat{a}$ does not disappear. There is hence some $\tilde{a}$ with $e(\tilde{a}) = e(\hat{a})$ and $\mu\tilde{a} > \mu\hat{a}$, which implies that $\text{Prob}(\underline{\pi} + \pi\mu\tilde{a} + \eta \leq r) < \text{Prob}(\underline{\pi} + \pi\mu\hat{a} + \eta \leq r)$, or equivalently, that the agent prefers $\tilde{a}$ to the Pareto-optimal way $\hat{a}$ of using $\underline{e}$. Assuming that performance is not rewarded congruently thus implies that the agent games incentives. □

This corollary establishes that congruent performance measures are clearly superior to dis-congruent ones in terms of directing effort. Still, Schnedler (2008) finds that performance measures which emphasize tasks that the agent likes lead to lower agency costs than congruent measures of the same precision. Measuring performance dis-congruently rather than congruently can only result in lower agency costs if either shirking costs, gaming costs, or both are lower (by Proposition 5). Congruent performance measures, however, do not entail gaming costs (by Corollary 6). It can thus be excluded that gaming costs are lower: agency costs must be smaller because of lower shirking costs.

Corollary 7. *The reason why some dis-congruent performance measures entail lower agency costs in the LEN model than congruent ones with the same precision are lower shirking costs.*

Proof. Let I be incentives that reward a signal with $\mu = \lambda b$, for some $\lambda > 0$, and

incentives $\tilde{I}$ have lower agency costs, $\alpha^{\tilde{I}} < \alpha^{I}$, which is equivalent to $S^{\tilde{I}} + G^{\tilde{I}} < S^{I} + G^{I}$ by Proposition 5. Using that $G^{I} = 0$ by Corollary 6 and $G^{\tilde{I}} \geq 0$, one gets: $S^{\tilde{I}} < S^{I}$. □

Shirking costs in the LEN model are partially driven by the agent's need for insurance. In this sense, the corollary confirms Schnedler's claim (2008) that insurance issues are the reason why dis-congruent measures are superior to congruent ones.

6 Conclusion

Despite Gibbons' observation (1998) that other issues are 'at least as important' as the 'tenuous' (Prendergast, 2002) incentive-insurance trade-off, formal models used for the education of future economic advisers and business consultants still heavily focus on this trade-off.[31] Given this focus and given that even scholars of incentives themselves neglect gaming (Christensen et al., 2010), the seemingly infinite supply of real-life examples of dysfunctional incentives is perhaps not so surprising.

This paper hopes to redress the balance with a theory of incentives that explicitly incorporates gaming. Intuitions are simple and the theory's generality reflects the importance of gaming for most applications. In summary, the paper offers an alternative to the standard paradigm which (if taught) can hopefully resolve misconceptions about incentive design and prevent dysfunctional incentives.

References

Averch, Harvey and Leland L. Johnson, "Behavior of the Firm Under Regulatory Constraint," *The American Economic Review*, 1962, *52* (5), pp. 1052–1069.

Baker, George, "Incentive Contracts and Performance Measurement," *Journal of Political Economy*, June 1992, *100* (3), 598–614.

[31] Major textbooks devote whole chapters to the trade-off, while multitasking is discussed in much shorter sections (Bolton and Dewatripont, 2005; Macho-Stadler and Perez-Castrillo, 1997; Laffont and Martimort, 2002) that are often to be skipped by first-time readers; gaming is not even indexed. Daron Acemoglu's MIT graduate course on labor economics (2014) formally discusses the traditional model on 30 slides, while gaming is only dealt with informally on four slides.

_ , "The Use of Performance Measures in Incentive Contracting," *American Economic Review*, May 2000, *90* (2), 415–420.

_ , "Distortion and Risk in Optimal Incentive Contracts," *Journal of Human Resources*, Fall 2002, *37* (4), 728–751. Special Issue on Designing Incentives to Promote Human Capital.

_ **, Robert Gibbons, and Kevin J. Murphy**, "Subjective Performance Measures in Optimal Incentive Contracts," *Quarterly Journal of Economics*, November 1994, *109* (4), 1125–1156.

Bénabou, Roland and Jean Tirole, "Intrinsic and Extrinsic Motivation," *Review of Economic Studies*, 2003, *70*, 489–520.

_ **and** _ , "Incentives and Prosocial Behavior," *American Economic Review*, December 2006, *96* (5), 1652–1678.

Berg, Norman A., "Lincoln Electric Company," Case 376-028, Harvard Business School 1983. revised from original August 1975 version.

Bergstrom, Ted, "Lecture Notes on Separable Preferences," Fall 2015. distributed as part of Economics 210A at the University of California, Santa Barbara.

Bolton, P. and M. Dewatripont, *Contract theory*, MIT Press, 2005.

Bond, Philip and Armando Gomes, "Multitask principal–agent problems: Optimal contracts, fragility, and effort misallocation," *Journal of Economic Theory*, 2009, *144* (1), 175–211.

Christensen, Peter O, Florin Sabac, and Jie Tian, "Ranking Performance Measures in Multi-task Agencies," *The Accounting Review*, 2010, *85* (5), 1545–1575.

Courty, Pascal and Gerald Marschke, "An Empirical Investigation of Gaming Responses to Explicit Performance Incentives," *Journal of Labor Economics*, January 2004, *22* (1), 23–56.

_ **and** _ , "A General Test for Distortions in Performance Measures," *The Review of Economics and Statistics*, 2008, *90* (3), 428–441.

Datar, Srikant, Susan Cohen Kulp, and Richard A. Lambert, "Balancing Performance Measures," *Journal of Accounting Research*, 2001, *39* (1), 75–92.

Dranove, David, Daniel Kessler, Mark McClellan, and Mark Satterthwaite, "Is More Information Better? The Effects of 'Report Cards' on Health Care Providers," *The Journal of Political Economy*, 2003, *111* (3), 555–588.

Feltham, Gerald A. and Jim Xie, "Performance Measure Congruity and Diversity in Multi-Task Principal/Agent Relations," *The Accounting Review*, July 1994, *69* (3), 429–453.

Feltham, Gerald .A. and Martin G. H. Wu, "Public Reports, Information Acquisition by Investors, and Management Incentives," *Review of Accounting Studies*, 2000, *5* (2), 155–190.

Forbes, Silke J., Mara Lederman, and Trevor Tombe, "Quality Disclosure Programs and Internal Organizational Practices: Evidence from Airline Flight Delays," *American Economic Journal: Microeconomics*, April 2015, *7* (2), 1–26.

Frey, Bruno S., "Publishing as Prostitution? – Choosing Between One's Own Ideas and Academic Success," *Public Choice*, 2003, *116* (1), 205–223.

_ , "Correspondence with David Autor," *Journal of Economic Perspectives*, September 2011, *25* (3), 239–240.

Friebel, Guido and Wendelin Schnedler, "Team Governance: Empowerment or Hierarchical Control," *Journal of Economic Behavior and Organization*, 2011, *78*, 1–13.

Gibbons, Robert, "Incentives in Organizations," *Journal of Economic Perspectives*, Fall 1998, *12* (4), 115–132.

_ **and Robert S. Kaplan**, "Formal Measures in Informal Management: Can a Balanced Scorecard Change a Culture?," *American Economic Review*, May 2015, *105* (5), 447–51.

Gjesdal, Froystein, "Accounting in Agencies," 1976. Graduate School of Business, Stanford University.

_ , "Information and Incentives: The Agency Information Problem," *Review of Economic Studies*, 1982, *49*, 373–390.

Grossman, Sanford J. and Oliver D. Hart, "An Analysis of the Principal-Agent Problem," *Econometrica*, January 1983, *51* (1), 7–46.

Halac, Marina, "Relational contracts and the value of relationships," *The American Economic Review*, 2012, *102* (2), 750–779.

Harris, Milton and Artur Raviv, "Optimal Incentive Contracts with Imperfect Information," *Journal of Economic Theory*, 1979, *20* (2), 231–259.

Hermalin, Benjamin E, "The effects of competition on executive behavior," *The RAND Journal of Economics*, 1992, pp. 350–365.

Herold, Florian, "Contractual Incompleteness as a Signal of Trust," *Games and Economic Behavior*, 2010, *68*, 180–191.

Holmström, Bengt, "Moral Hazard and Observability," *Bell Journal of Economics*, 1979, *10* (1), 74–91.

_ , "Moral Hazard in Teams," *Bell Journal of Economics*, 1982, *13* (2), 324–340.

_ **and Paul Milgrom**, "Multitask Principal-Agent-Analysis: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization*, 1991, *7*, 24–52. special issue.

Hong, Fuhai, Tanjim Hossain, John A List, and Migiwa Tanaka, "Testing the Theory of Multitasking: Evidence from a Natural Field Experiment in Chinese Factories," Technical Report, National Bureau of Economic Research 2013.

Jacob, Brian A and Steven D Levitt, "Rotten apples: An investigation of the prevalence and predictors of teacher cheating," *The Quarterly Journal of Economics*, 2003, *118* (3), 843–877.

Kaplan, R.S. and D.P. Norton, "Using the Balanced Scorecard as a Strategic Management System," *Harvard Business Review*, 1996, *74*, 75–85.

Kerr, Steven, "On the Folly of Rewarding A While Hoping for B," *Academy of Management Journal*, 1975, *18* (4), 769–783.

Kim, Son Ku, "Efficiency of Information System in an Agency Model," *Econometrica*, January 1995, *63* (1), 89–102.

Laffont, Jean-Jacques and David Martimort, *The Theory of Incentives: The Principal-Agent Model*, Princeton: Princention University Press, 2002.

Larkin, Ian, "The cost of high-powered incentives: Employee gaming in enterprise software sales," *Journal of Labor Economics*, 2014, *32* (2), 199–227.

Levin, Jonathan, "Relational Incentive Contracts," *American Economic Review*, August 2003, *93* (3), 835–847.

Macho-Stadler, Ines and D. Perez-Castrillo, *An Introduction to the Economics of Information: Incentives and Contracts*, Oxford: Oxford University Press, 1997.

MacLeod, W. Bentley and James M. Malcomson, "Implicit Contracts, Incentive Compatibility, and Involuntary Unemployment," *Econometrica*, March 1989, *57* (2), 227–480.

_ **and** _ , "Motivation and Markets," *American Economic Review*, June 1998, *88* (3), 388–411.

Nobel Prize Committee, "Oliver Hart and Bengt Holmström: Contract Theory. Scientific Background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel," oct 2016. The Royal Swedish Academy of Science.

Oyer, Paul, "Fiscal Year Ends and Nonlinear Incentive Contracts: The Effect on Business Seasonality," *Quarterly Journal of Economics*, 1998, pp. 149–185.

Paine, Lynn Sharp, "Managing for organizational integrity," *Harvard business review*, 1994, *72* (2), 106–117.

Prendergast, Canice, "The Provision of Incentives in Firms," *Journal of Economic Literature*, March 1999, *XXXII*, 7–63.

_ , "The Teneous Trade-Off Between Risk and Incentives," *Journal of Political Economy*, 2002, *110* (5), 1071–1102.

Propper, Carol, Matt Sutton, Carolyn Whitnall, and Frank Windmeijer, "Incentives and Targets in Hospital Care: Evidence from a Natural Experiment," *Journal of Public Economics*, 2010, *94* (3), 318–335.

Raith, Michael, "Specific Knowledge and Performance Measurement," *The RAND Journal of Economics*, Winter 2008, *39* (4), 1059–1079.

Schnedler, Wendelin, "When is It Foolish to Reward for A While Benefiting from B?," *Journal of Labor Economics*, 2008, *26* (4), 595–619.

_ , "You Don't Always Get What You Pay For," *German Economic Review*, February 2011, *12* (1), 1–10.

_ **and Christoph Vanberg**, "Playing 'Hard to Get': An Economic Rationale for Crowding Out of Intrinsically Motivated Behavior," *European Economic Review*, 2014, *68*, 106–115.

_ **and Radovan Vadovic**, "Legitimacy of Control," *Journal of Economics and Management Strategy*, 2011, *20* (4), 985–1009.

Seabright, Paul B., "Continuous Preferences and Discontinuous Choices: How Altruists Respond to Incentives," *The B.E. Journal of Theoretical Economics*, April 2009, *9* (1, Contributions), Article 14. Contributions.

Shavell, Steven, "Risk-Sharing and Incentives in the Principal-Agent Relationship," *Bell Journal of Economics*, 1979, *10* (1), 55–73.

Sliwka, Dirk, "On the Hidden Costs of Incentive Schemes," *American Economic Review*, 2007, *97* (3), 999–1012.

_ **and Kathrin Manthei**, "Multitasking and the Benefits of Objective Performance Measurement-Evidence from a Field Experiment," 2013.

Sloof, Randolph and Mirjam van Praag, "Testing for Distortions in Performance Measures: An Application to Residual Income-Based Measures like Economic Value Added," *Journal of Economics & Management Strategy*, 2015, *24* (1), 74–91.

van der Weele, Joel, "The Signaling Power of Sanctions in Social Dilemmas," *Journal of Law, Economics, and Organization*, 2012, *28* (1), 103–125.

A Additional Results for Multitasking Example

Proposition 7 (Aligning does not ensure most beneficial use of time)**.** *Aligning incentives with the principal's benefit leads the academic to spend a total time, say* $t^\pi := t(a^\pi)$*, on marketing and research but does not ensure that he allocates* t^π *in the most beneficial way for the principal:*

$$\textit{for some } \rho = \beta \textit{ and } \pi > 0: \quad a^\pi \neq \arg \max_{\{a|t(a)\leq t^\pi\}} b(a).$$

Proof. The proof works by counter-example. Suppose the principal prefers research to marketing, say $1 > \beta > \frac{1}{2}$. Since incentives are aligned, $1 > \rho > \frac{1}{2}$ and by (2), the total time spent by the agent on both tasks is $a_1 + a_2 = \frac{\pi}{2}$, where some time is spend on marketing: $a_2 = (1-\rho)\frac{\pi}{2} > 0$. Next, we show that this way of using work time does not maximize the principal's benefit. By shifting all time spent on marketing to research, the principal's benefit from researcher increases by $(1-\rho)\frac{\pi}{2}\beta$, while he loses $(1-\rho)\frac{\pi}{2}(1-\beta)$ from less time being spent on marketing. Overall, the principal's benefit increases because $\beta > \frac{1}{2}$. The induced action thus did not maximize the principal's benefit. □

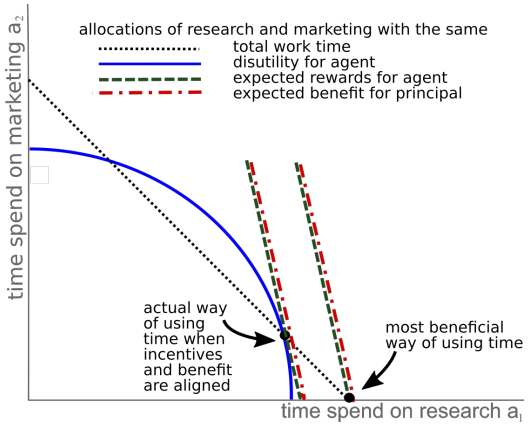

Figure 4: The academic produces expected rewards with the least effort (lowest dis-utility). Although incentives are aligned with the benefit, total work time is not used in the most beneficial way for the principal.

Lemma 2. *Given* $\beta > 1$*, gaming and shirking costs amount to:*

$$G^{\pi} = \frac{\pi}{2}\left(\beta\sqrt{\rho^2+(1-\rho)^2} - (\rho\beta+(1-\rho)(1-\beta))\right),$$

$$S^{\pi} = \frac{\beta^2}{4} - \left(\beta\frac{\pi}{2}\sqrt{\rho^2+(1-\rho)^2} - \frac{(\pi)^2}{4}\left(\rho^2+(1-\rho)^2\right)\right).$$

Proof. Observe that for $\beta < 1$, all effort should be focused on research, so that $G^{\pi} = b((\sqrt{e^{\pi}},0)) - b(a^{\pi})$, then plug in a^{π} from (2) and a^{π} from (5). For shirking costs, recall that $S^{\pi} = b(a^*) - b((\sqrt{e^{\pi}},0)) - (r^* - r^{\pi})$ and plug in a^* from (1), e^{π} from (5) and r^{π} from (3). □

Lemma 3. *Given* $\beta > 1$*, the increase in gaming costs outweighs the reduction in shirking costs at* $\pi = 0$*,* $\frac{dG^{\pi}}{d\pi} \geq \left|\frac{dS^{\pi}}{d\pi}\right|_{\pi=0}$ *if and only if* $\rho \leq \frac{\beta-1}{2\beta-1}$*.*[32]

Proof.

$$\frac{dG^{\pi}}{d\pi} \geq \left|\frac{dS^{\pi}}{d\pi}\right|_{\pi=0}$$

$$\Leftrightarrow \frac{1}{2}\left(\beta\sqrt{\rho^2+(1-\rho)^2} - (\rho\beta+(1-\rho)(1-\beta))\right) \geq \frac{1}{2}\beta\sqrt{\rho^2+(1-\rho)^2}$$

$$\Leftrightarrow -(\rho\beta+(1-\rho)(1-\beta)) \geq 0 \Leftrightarrow -(1-\beta) \geq \rho(\beta-1+\beta).$$

□

[32]For a numerical example fulfilling the condition take $\rho = \frac{1}{4}$ and $\beta = \frac{3}{2}$.

B Proof for Proposition 6

Proof. In order to construct the counter-example, we use a variation of the earlier academic example, in which insurance issues matter. In line with Kim (1995), assume that the agent's utility is $w(r)-e$, where w is a concave function so that the agent is risk-averse. In order to vary the 'noise' of signals, suppose that the publication signal is 'polluted'. The true realization of this signal is only observed with probability γ but with probability $1-\gamma$, the observed signal shows the opposite of the actual realization:

$$\tilde{Y} = \begin{cases} Y & \text{with probability} \quad \gamma \\ (1-Y) & \text{with probability} \quad 1-\gamma \end{cases},$$

where $\gamma \geq \frac{1}{2}$, measures the degree to which the original signal matters. In other words, signals with larger γ are less noisy. In particular, $\tilde{Y}$ is a Blackwell-garbled version of Y, where the degree of garbling can be controlled by γ.

Apart from the change in the agent's utility and the class of signals, all assumptions are identical to Section 2. Before computing the agent's choice given incentives based on $\tilde{Y}$, we compute the probability of success of the new signal:

$$\begin{aligned} P(\tilde{Y}=1) &= (\rho a_1 + (1-\rho)a_2)\cdot\gamma + (1-(\rho a_1+(1-\rho)a_2))(1-\gamma) \\ &= (1-\gamma) + (\rho a_1 + (1-\rho)a_2)(2\gamma-1). \end{aligned}$$

By slightly abusing notation and denoting with π the agent's gain in utility resulting from larger rewards: $\pi = w(r_1)-w(r_0)$, where r_y is the reward in case of $\tilde{Y}=y$, we obtain the agent's choice as: $a^{\pi} = \frac{\pi}{2}(2\gamma-1)(\rho,(1-\rho))$. This formula allows us to re-interpreted the signals $\tilde{Y}$ and Y as a signals about effort rather than the agent's action. This is possible because every action choice requires some effort $e = a_1^2 + a_2^2$, so that conversely, any effort level e is used for a specific action choice given the signal:

$$(a_1(e), a_2(e)) = (\rho,(1-\rho))\frac{\sqrt{e}}{\sqrt{\rho^2+(1-\rho)^2}}.$$

Now, use this description of the action choice to re-compute the probability of a successful signal $\tilde{Y}=1$:

$$P(\tilde{Y}=1) = (1-\gamma) + (\rho^{\rho} + (1-\rho)^2)\frac{\sqrt{e}}{\sqrt{\rho^2+(1-\rho)^2}}(2\gamma-1) \tag{10}$$

$$= (1-\gamma) + \sqrt{e\cdot(\rho^2+(1-\rho)^2)}(2\gamma-1). \tag{11}$$

From this expression, we see that both, Y (with $\gamma = 1$) as well as its Blackwell garbling $\tilde{Y}$, are signals about effort e in the sense that their distributions only depends on effort e. Since a less garbled signal is a sufficient statistic for the more garbled signal, Proposition 2 in Kim (1995) ensures that Y is a mean-preserving spread of the likelihood ratio distribution of $\tilde{Y}$. Using Proposition 1 in Kim (1995) the expected payment, which is required to obtain effort e with $\tilde{Y}$, say $r(e,\gamma)$, is higher than that for Y, say $r(e,1)$. Signal Y is thus preferable to $\tilde{Y}$ in terms of shirking costs.

Next, we show that signal Y is not necessarily preferable in terms of agency costs. The reason should by now be obvious: higher gaming cost may outweigh reductions in shirking costs. Suppose that $\beta = 1$ and that $\rho = 0$ for signal Y and $\rho = 1$ for signal $\tilde{Y}$. Despite this parameter choice $\tilde{Y}$ remains a Blackwell garbling of Y with respect to effort: the probability of success computed in (11) is $(1-\gamma)+\sqrt{e}(2\gamma-1)$ and for $\tilde{Y}$ and $\sqrt{e}$ for Y. Eliciting effort with Y remains cheaper than with $\tilde{Y}$ but the effort induced with Y is completely useless, while the effort with $\tilde{Y}$ is efficiently employed. When γ approaches one, the difference between expected payments $r(e,1) - r(e,\gamma)$ becomes smaller, while the difference in generated benefit remains constant at $\sqrt{e}$. If we compare agency costs for sufficiently large γ, we get $\Pi_Y = 0 - r(e,1) < \Pi_{\tilde{Y}} = \sqrt{e} - r(e,\gamma)$. So, $\tilde{Y}$ is preferred to Y, although $\tilde{Y}$ is 'worse' in the sense of being a Blackwell garbling of Y and hence also in the sense of Y being a sufficient statistic for $\tilde{Y}$ and its likelihood ratio distribution function being a mean-preserving spread of the latter. □