Institutional Members: CEPR, NBER and Università Bocconi

WORKING PAPER SERIES

Costly Contracting in a Long-Term Relationship

Pierpaolo Battigalli and Giovanni Maggi

Working Paper n. 249

November 2003

IGIER – Università Bocconi, Via Salasco 5, 20136 Milano –Italy
http://www.igier.uni-bocconi.it

Costly Contracting in a Long-Term Relationship*

Pierpaolo Battigalli
Bocconi University
pierpaolo.battigalli@uni-bocconi.it

Giovanni Maggi
Princeton University
maggi@princeton.edu

November 2003

Abstract

We examine a model of contracting where parties interact repeatedly and can contract at any point in time, but writing enforceable contracts is costly. A contract can describe contingencies and actions at a more or less detailed level, and the cost of writing a contract is proportional to the amount of detail. We consider both formal (externally enforced) and informal (self-enforcing) contracts. The presence of writing costs has important implications both for the optimal structure of formal contracts, particularly the tradeoff between contingent and spot contracts, and for the interaction between formal and informal contracting. Our model sheds light on these implications and generates a rich set of predictions about the determinants of the optimal mode of contracting.

JEL classification: D23, C73.

KEYWORDS: writing costs, contingent *vs* spot contracting, formal *vs* informal contracts.

*Corresponding author: Pierpaolo Battigalli, Bocconi University, IEP and IGIER, Via Salasco 5, 20136 Milano (Italy). Pierpaolo Battigalli thanks Bocconi University for financial support, as well as Princeton University for its hospitality. Giovanni Maggi gratefully acknowledges financial support from the National Science Foundation. We thank Luca Anderlini, Arnaud Costinot, Avinash Dixit, Kfir Eliaz, Leonardo Felli, Bentley McLeod, Ludovic Renou, Joel Watson, the participants in conferences at Siena, Venezia and Gertzensee, and the participants in seminars at Northwestern University, PUC Rio de Janeiro, the European University Institute, Bocconi University, DELTA, INSEAD and Helsinki for helpful comments and discussions.

1. Introduction

Contracting in a long-term relationship may come in a variety of modes. Contracts may be formal (i.e. externally enforced), informal (i.e. self-enforcing),[1] or a combination of the two. Formal contracts in turn may be of the contingent type, of the "spot" type, or a mixture of the two.[2] In this paper we examine the optimal mode of contracting in the presence of writing costs. We will present a simple model that sheds light on the implications of writing costs and yields rich predictions about the determinants of the mode of contracting.

We consider a multi-task, principal-agent setting with verifiable contingencies and actions, where parties interact repeatedly and can write contracts at any point in time (this includes the possibility of spot contracting, as contracts can be written after observing the state of nature and before actions are taken). A contract can describe contingencies and actions at a more or less detailed level, and the cost of writing a contract is increasing in the amount of detail. In each period, parties can save on writing costs by modifying the previous contract rather than drafting a whole new contract.

In the first part of the paper we focus on situations where the parties rely entirely on formal contracting. A simple way to model these situations is to consider the finite-horizon version of the game, which does not admit "reputational" equilibria. We examine the optimal structure of formal contracts, and in particular the choice between contingent and spot contracts. At the intuitive level, it is not obvious whether the presence of writing costs should favor contingent or spot contracting: on the one hand, spot contracting avoids the cost of describing contingencies; on the other hand, spot contracts must describe the agent's *behavior* repeatedly, and this may push in favor of a contingent contract. The idea that transaction costs might favor long-term contingent contracts has already been expressed informally, for example by Hart and Holmstrom

[1] We will use interchangeably the expressions "informal" and "self-enforcing". We refrain from using the terminology "explicit" vs. "implicit" contracts – which is common in the literature – because contracts may be quite explicit even though they cannot be enforced in court.

[2] The difference between contingent and spot contracting is nicely explained by Williamson (1985, p.20). He states that the writing of contracts "*...can be done with a great deal of care, in which case a complex document is drafted in which numerous contingencies are recognized, and appropriate adaptations by the parties are stipulated and agreed to in advance. Or the document can be very incomplete, the gaps to be filled by the parties as the contingencies arise. Rather, therefore, than contemplate all conceivable bridge crossings in advance, which is a very ambitious undertaking, only actual bridge-crossing choices are addressed as events unfolds.*"

(1987, p. 130), who write: "*if a relationship is repetitive, it may save on transaction costs to decide in advance what actions each party should take rather than to negotiate a succession of short term contracts.*" Our model allows us to explore this idea more systematically.[3]

Absent writing costs, the model has no predictive power, because there is a plethora of optimal contracting plans, including a contingent contract, a sequence of spot (noncontingent) contracts, and a host of intermediate solutions. But in the presence of (even very small) writing costs, the model yields a unique optimum. Each task is optimally handled in one of three ways: (i) a contingent clause, written once and for all; (ii) a sequence of spot clauses; or (iii) a noncontingent clause that is replaced at some point in time by a contingent clause (we refer to this as the "enrichment" approach). A key parameter that affects the optimal mode of contracting is the cost of describing contingencies relative to the cost of describing actions. A contingent approach is optimal when this ratio is low, a spot approach is optimal when it is high, and an enrichment approach may be optimal when it takes intermediate values.

If tasks are sufficiently heterogeneous, so that the enrichment approach is optimal for some of the tasks, the model predicts that the number of contingent clauses in the contract will increase over time. This is consistent with an empirical phenomenon that has been documented by Meihuizen and Wiggins (2000): in the natural gas industry in the United States, they find that gas supply contracts become gradually more contingent over time.

The optimal contracting mode is also affected by the degree of uncertainty and by the discount factor. Tasks that are characterized by a higher degree of uncertainty are more likely to be regulated by contingent clauses, whereas lower-uncertainty tasks are more likely to be regulated by a spot approach. The intuition is that, when external shocks are more frequent, the writing costs associated with a spot approach are higher, while the cost of writing contingent clauses is not affected. A higher discount factor tends to favor contingent contracting over spot contracting. This is a consequence of the fact that the cost of writing contingent clauses is incurred upfront, while the cost of a spot approach is spread out over time.

[3]In our model contingent contracting may be preferred to spot contracting because it is costly to describe actions. Another type of transaction cost that can generate a preference for contingent contracts is given by fixed recontracting costs. This point is made in the context of macroeconomic models by Gray (1976, 1978) and Dye (1985b).

We emphasize that the predictions of our model would be radically different if writing costs were modeled in a more conventional way, and in particular along the lines of Dye's (1985a) well-known model of costly contracting. For example, we show that, with writing costs *à la* Dye, a contingent contract is typically dominated by a sequence of spot contracts. This contrasts with our model, where a contingent contract may well be optimal.

Our model also provides an interesting insight concerning Maskin and Tirole's (1999) well-known irrelevance result. They have argued, within a static setting, that the possibility of unforeseen contingencies does not create inefficiencies, provided that an appropriate message-based mechanism can be played *ex post*. Our analysis highlights that this argument breaks down if parties interact repeatedly and transaction costs take the form of writing costs, rather than unforeseen contingencies.

In the second part of the paper we consider the possibility of self-enforcing contracts, i.e. contracts that are enforced by "reputation" mechanisms rather than by external courts. We introduce this possibility by considering the infinite-horizon version of the game. The advantage of a self-enforcing contract is that it can be communicated informally, rather than being written formally, because it need not be enforceable in court, and this saves on writing costs. On the other hand, the absence of an external enforcement mechanism may limit the effectiveness of a self-enforcing contract. Hence there is a trade-off between formal and informal contracting. When we allow for both formal and informal contracting in the model, we find that they tend to be used jointly, with some tasks regulated formally and others regulated informally. In particular, tasks characterized (other things equal) by a lower cost for the agent are regulated by informal contracting, while higher-cost tasks are regulated by formal contracting. This is because the agent has a stronger incentive to cheat for higher-cost tasks. The presence of writing costs can thus contribute to explain the fact that long-term relationships are often managed by a combination of formal and informal contracting.

Formal and informal contracts coexist provided that (i) the discount factor is not too high, so that a fully informal contract cannot be sustained, and (ii) at least one task has a relatively low cost, so that at least one task can be regulated informally. Interestingly, this is true even if writing costs are very small. Moreover, we find that the fraction of tasks regulated informally

may *in*crease as writing costs decrease.

Another interesting finding is that the relative importance of informal contracting tends to be higher when the potential surplus from the relationship, or the gains from contracting, are larger. The intuitive reason is that, when the surplus is higher, the temptation to cheat on the informal contract is lower. This is also the intuition behind the result mentioned in the above paragraph: a decrease in writing costs leads to an increase in the net surplus, which relaxes the incentive constraint and thus may increase the number of informally-regulated tasks.

Most of our analysis focuses on the case in which writing costs are relatively small, and more specifically, small enough that it is optimal to implement the first best outcome. In other words, we focus on the case in which contracting is "complete."[4] The emphasis on this case is useful for two reasons. First, all the main insights can be brought out with small writing costs, and the exposition is considerably simpler than in the case of large writing costs. Second, this helps to clarify that contractual incompleteness *per se* is not essential to our results. In the final part of the paper we show how to extend the analysis to the case of large writing costs. If writing costs are large, the main difference in results is that contracting may be incomplete. Contractual incompleteness can take two forms: rigidity and discretion. Other things equal, tasks characterized by low surplus are left to the agent's discretion, intermediate-surplus tasks are regulated by rigid rules, and high-surplus tasks are regulated in a first-best way (by formal or informal contracting). This parallels a result derived in Battigalli and Maggi (2002), which focuses on a static model of contracting with writing costs.[5]

The interaction between formal and informal contracts has been analyzed also by Baker *et al.* (1994) and Pearce and Stacchetti (1998).[6] These papers propose a different explanation for the combined use of the two types of contract. They consider a repeated principal-agent model where formal contracts can be based only on verifiable signals of the agent's action, whereas informal contracts can be based on unverifiable signals. In these models it may be optimal to

[4] By "complete" contracting we mean simply that the first-best outcome is implemented, so this may include contingent formal, spot formal, and informal contracting.

[5] It is important to mention that there exist a number of other papers on complexity costs as a cause of contractual incompleteness in a static setting, for example Dye (1985a), Anderlini and Felli (1994, 1999), MacLeod (2000), Krasa and Williams (2001) and Al Najjar *et al.* (2002).

[6] There is also a vast literature on purely self-enforcing contracts. Bull (1987) and Mac Leod and Malcomson (1989) are two prominent examples of this literature.

combine a formal wage and an informal 'bonus'. Our model, on the other hand, explains why it may be efficient to regulate some tasks formally and some others informally (in our model there is no need for bonuses). Perhaps more importantly, the rationale for mixing formal and informal contracting in our model is not the presence of verifiable and non-verifiable signals, but the interaction between writing costs and self-enforcement constraints.

Our model also yields contrasting predictions on the interplay between formal and informal contracting. One key result in Baker *et al.* (1994) is that, if the imperfection in formal contracting is sufficiently small, informal contracts cannot be sustained. Therefore their model predicts that, if the formal-contracting system becomes more efficient, informal contracting may be undermined as a consequence. In our model, even if formal contracting is close to perfect (i.e. if writing costs are close to zero), the optimum typically involves a mix of formal and informal contracting or a fully informal contract. Thus, our analysis suggests that informal contracting need not disappear as the formal-contracting system becomes more efficient.[7]

Finally, our model yields interesting predictions on the choice between spot and contingent contracts. The models by Baker *et al.* (1994) and Pearce and Stacchetti (1998) are silent about this issue because the nature of the contractual imperfection is different. In these models the imperfection is given by the non-verifiability of signals about the agent's performance. This imperfection is exogenous and is present in every period, while in our model the imperfection (writing costs) is endogenous and need not be incurred in every period.

2. The contracting environment

We analyze a repeated multi-task principal-agent game where parties can contract in each period. Payoffs depend on actions and external contingencies, which are both verifiable in court but costly to describe in a written contract. Thus, the only contractual imperfection

[7] The reason for this divergence in results lies in the punishment strategy. Baker *et al.* assume that, if a player cheats, parties revert to the optimal formal contract, with all the surplus from this contract accruing to the principal. This implies that, if formal contracting is close to perfect, it completely fails to deter the principal from cheating. However, in general this is not the most severe credible punishment strategy. Our approach, on the other hand, is to characterize the worst credible punishment strategies, which we are able to do under the condition that the discount factor is not too low. If this condition is satisfied, each player can be punished with his/her maxmin payoff, hence changes in writing costs do not affect the severity of the punishment.

is the presence of writing costs. We model payoffs and writing costs in a similar way as in Battigalli and Maggi (2002). We start by describing the language used to write contracts.

$\Pi^e = \{e_1, e_2, ..., e_N\}$ is a finite collection of *primitive sentences*, each of which describes an *elementary event* concerning the external environment. For example, e_1: "the passenger has a moustache", e_2: "the passenger's bag is red".

$\Pi^a = \{a_1, a_2, ..., a_N\}$ is a finite collection of primitive sentences describing *elementary actions* (behavioral events), for example, a_1 : "check the passenger's passport", a_2 : "search the passenger's bag".

With a slight abuse of terminology, we will use the notation e_k (resp. a_k) to indicate both an elementary event (resp. action) and the primitive sentence that describes it.

We assume that this language is the (only) common-knowledge language for the parties and the courts. This ensures that there are no problems of ambiguous interpretation of the contract.

A *state* is a complete description of the exogenous environment, represented by a valuation function $s : \Pi^e \to \{0, 1\}$, where $s(e_k) = 1$ means that primitive sentence e_k is true at state s and $s(e_k) = 0$ means that primitive sentence e_k is false at state s.[8] In other words, $s(e_k)$ is a dummy variable that takes value one if elementary event e_k occurs and zero otherwise, and a state is a realization of the vector of dummy variables $(s(e_1), s(e_2), ...)$.

Similarly, a *behavior* is a complete description of all elementary actions, represented by a valuation function $b : \Pi^a \to \{0, 1\}$, where $b(a_k) = 1$ means that elementary action a_k is executed, and $b(a_k) = 0$ that a_k is not executed.

We assume a very simple payoff structure. There is a one-to-one correspondence between elementary actions and elementary events. The principal wants action a_k to be performed if and only if elementary event e_k occurs. In our airport example, the principal wants the agent to check the passenger's passport if and only if the passenger has a moustache, and to search his bag if and only if the bag is red.

[8] To simplify the exposition we describe the basic notation omitting time subscripts. We will introduce time subscripts later in this section, when we describe the dynamic aspects of the game.

Principal and agent are risk neutral. The principal gets an incremental benefit of π_k from "matching" $s(e_k)$ with $b(a_k)$, while he gets zero incremental benefit if there is a "mismatch". Formally, the principal's per-period payoff gross of writing costs is:

$$\pi(s, b, m) = \sum_{k=1}^{N} \pi_k I_k(s, b) - m \tag{2.1}$$

where m is the payment to the agent and $I_k(s, b) = s(e_k)b(a_k) + (1 - s(e_k))(1 - b(a_k))$ is a dummy variable that takes value one if there is a match between event e_k and action a_k, and zero if there is a mismatch.

The agent's interests are always in conflict with the principal's, in the sense that his preferred actions are always opposite the principal's preferred actions. Formally, the agent's one-period utility is:

$$U(s, b, m) = m - \sum_{k=1}^{N} d_k I_k(s, b). \tag{2.2}$$

We will often refer to the job of matching action k with event k as "task" k. The parameter d_k thus captures the agent's disutility from performing task k. We let $\mathbf{N} = \{1, ..., N\}$ denote the set of tasks and we use bold capital letters to denote subsets of tasks, as in $\mathbf{K} \subseteq \mathbf{N}$.

The parties' reservation payoff is zero. Assuming $0 < d_k < \pi_k$ for all k, the parties' joint surplus (gross of writing costs) is maximized when the agent performs all tasks $k \in \mathbf{N}$. We will refer to this as the *first best* outcome.

Payoffs are common knowledge to the contracting parties, and the state and the parties' behavior are verifiable in court. Thus, there are no issues of moral hazard or adverse selection. We assume that preferences and realized payoff levels are not verifiable in court, and that the principal cannot "sell the activity" to the agent (i.e., the agent cannot be made the recipient of the revenue π).[9]

Next we define a contract. A contract is a pair $(\mathbf{g}, m)$ where $\mathbf{g} = (g_k)_{k \in \mathbf{N}}$ is a set of N

[9] If preferences were verifiable, the first-best outcome could trivially be achieved by a contract of the form "The agent's behavior must maximize the sum of the parties' utilities." On the other hand, if realized payoff levels were verifiable, the first-best outcome could be achieved by offering the agent a transfer that increases one-for-one with the principal's realized payoff level. And selling the activity to the agent would be equivalent to specifying a contingent transfer as in the previous point.

clauses and m is a transfer from the principal to the agent (wage). Each clause g_k regulates a task. Given our simple matching structure between actions and events, we can restrict our attention to four types of clause: (i) a contingent clause, constraining the agent to do a_k if and only if e_k occurs, $C_k : [a_k \leftrightarrow e_k]$; (ii) a noncontingent positive clause, constraining the agent to do a_k whatever happens, $R_k : [a_k]$; (iii) a noncontingent negative clause, constraining the agent to do *not* a_k whatever happens, $\bar{R}_k : [\neg a_k]$; (iv) the empty clause, D (for discretion), that imposes no constraint on the agent. For example, if $N = 3$, the set of clauses (R_1, D, C_3) constrains the agent to do a_1 whatever happens and to do a_3 if and only if elementary event e_3 occurs, leaving the agent free with regard to task 2. Note that, since we include the empty clause among the possible clauses, there is no loss of generality in assuming that the number of clauses in the contract is N.[10]

The parties interact for T periods (where T may be infinite) and have common discount factor $\delta \in [0, 1)$. The parameter δ can also be interpreted as capturing the stability of the relationship.[11] Within each period t, the timing is the following: the state of nature s_t is observed; then the principal offers a contract $(\mathbf{g}_t, m_t)$ to the agent, incurring the associated writing costs; if the contract is accepted, the principal makes payment m_t and then the agent acts (being constrained by the contract).

We assume that elementary events are independent of each other, and that each elementary event is governed by a Markov process. Let the transition probabilities for event e_k be $\Pr[s_t(e_k) = 1 | s_{t-1}(e_k) = 1] = p_k$ and $\Pr[s_t(e_k) = 0 | s_{t-1}(e_k) = 0] = q_k$. Without loss of generality we define labels so that $p_k \geq q_k$ for all k. We assume $p_k \geq 1/2$ for all k. Note that an *i.i.d.* process corresponds to the special case $p_k + q_k = 1$. Another special case of interest is that of *symmetric persistence:* $p_k = q_k$. Finally we assume that the initial state is $s_1(e_k) = 1$ for all k; the extension to the more general case in which also the initial state is random is straightforward but tedious.

If at time t the principal wants to offer a different contract than at time $t - 1$, he can

[10] We note that in this model there is nothing to gain from making wages contingent on the state or on the agent's behavior.

[11] The parameter δ can be interpreted as the composition of two parameters, $\delta = \rho\delta'$, where ρ is the probability that the game will continue and δ' is the 'true' discount factor.

save on writing costs (whose exact nature will be specified below) by *modifying* the existing contract rather than writing a whole new contract. Contract modifications can take two forms: (i) *amendments*, that is, permanent modifications of the contract; or (ii) *exceptions*, that is, temporary modifications applied only for the current period.[12]

To capture this idea, we distinguish between the *effective* contract (i.e. the contract actually enforced at time t) and the *default* contract. The default contract at t is given by the default contract at $t-1$ modified by the amendments at t, and the effective contract at t is given by the default contract at t modified by the exceptions at t. The default contract will be a key state variable of our problem, while the set of amendments and exceptions will be the control variable.

More formally, the set of *default* clauses is

$$\tilde{\mathbf{g}}_t = (\tilde{g}_{k,t})_{k\in\mathbf{N}},$$

where $\tilde{g}_{k,t} \in \{C_k, R_k, \bar{R}_k, D\}$. The set of default clauses at time t is given by:

$$\tilde{\mathbf{g}}_t = \tilde{f}(\tilde{\mathbf{g}}_{t-1}, \mathbf{g}_t^A) = \left((\tilde{g}_{k,t-1})_{k\in\mathbf{N}\setminus\mathbf{K}_t^A}, (\alpha_{k,t})_{k\in\mathbf{K}_t^A}\right),$$

where $\alpha_{k,t} \in \{C_k, R_k, \bar{R}_k, D\}$ $(\alpha_{k,t} \neq \widetilde{g}_{k,t-1})$ is the amendment for task k and $\mathbf{g}_t^A = (\alpha_{k,t})_{k\in\mathbf{K}_t^A}$ is the set of amendments. For each task k, the default clause at $t=0$ is the empty clause: $\tilde{g}_{k,0} = D$.

The set of *effective* clauses at time t is given by:

$$\mathbf{g}_t = f(\tilde{\mathbf{g}}_t, \mathbf{g}_t^E) = \left((\tilde{g}_{k,t})_{k\in\mathbf{N}\setminus\mathbf{K}_t^E}, (\varepsilon_{k,t})_{k\in\mathbf{K}_t^E}\right),$$

where $\varepsilon_{k,t} \in \{C_k, R_k, \bar{R}_k, D\}$ $(\varepsilon_{k,t} \neq \widetilde{g}_{k,t-1})$ is the exception for task k and $\mathbf{g}_t^E = (\varepsilon_{k,t})_{k\in\mathbf{K}_t^E}$ is the set of exceptions (it may be assumed without loss of generality that $\mathbf{K}_t^A \cap \mathbf{K}_t^E = \emptyset$).

We can now describe the costs of writing contracts. Writing primitive sentences is costly. We allow for a simple form of dynamic writing economies: writing a given primitive sentence

[12] The key assumption here is that, if at time t nothing is written about task k, the existing clause (if any) about task k applies. We do not allow parties to save on writing costs by recalling contracts from earlier dates. For example, we do not allow contract $\mathbf{g}_t$ to say "contract $\mathbf{g}_{t-2}$ applies with the following modifications...". A more general model would allow for richer 'recalling' possibilities, but we conjecture that, if there is a costs of recalling more remote contracts, the key insights of the analysis would not change.

$\xi \in \Pi^a \cup \Pi^e$ for the first time is (weakly) more costly than writing it the subsequent times. If $c(\xi)$ denotes the cost of writing ξ for the first time and $c^r(\xi)$ denotes the cost of writing ξ each subsequent time (r stands for "recall"), we are assuming $c^r(\xi) \leq c(\xi)$. Specifying the wage and writing logical connectives (such as $\neg$ or $\leftrightarrow$) in the contract is costless.

In each period t the principal incurs the costs of modifying the previous contract, that is, the costs of writing the amendments and exceptions, $(\mathbf{g}_t^A, \mathbf{g}_t^E)$.[13] These costs (which are history-dependent, due to the recalling economies) can be derived using the assumptions above. Focusing on task k, there are only a few relevant possibilities that we need to consider:

(i) Writing a contingent clause C_k at time $t = 1$ costs $c(a_k) + c(e_k)$;

(ii) Writing a noncontingent clause (R_k or $\bar{R}_k$) at time $t = 1$ costs $c(a_k)$.

(iii) Replacing clause R_k with clause $\bar{R}_k$ or *viceversa* costs $c^r(a_k)$, since this involves recalling an already-described action.

(iv) Replacing a noncontingent clause (R_k or $\bar{R}_k$) with a contingent clause (C_k) costs $c^r(a_k) + c(e_k)$, since this involves describing a new elementary event and recalling an already-described action.

(v) The empty clause D of course involves no cost. Also, removing a clause (i.e. replacing it with the empty clause D) is costless.[14]

Note that we do not consider the possibility of multiperiod contracts (the current contract constrains players only for the current period), and the principal is not allowed to make payments in excess of the wage specified in the formal contract. At the end of section 4 we will argue that both of these restrictions are without loss of generality in this model. We will also argue that our results would not change if the payment m were made after the agent acts.

We emphasize that in each period the contract is written *after* the state is observed. Thus, writing a contingent contract would not make sense if the parties interacted for just one period.

[13] Note that writing a clause at $t = 1$ can be seen as a modification of the empty clause D, which is the default clause at $t = 0$.

[14] Introducing a cost of removing clauses would make our notation heavier without changing our results.

But with repeated interaction, as we will see shortly, a contingent contract may be efficient.

A final note before we start with the analysis. In the literal interpretation of the game, the parties sign a one-period contract in each period. However, we could equivalently assume that contracts are open-ended, in the sense that the existing contract is automatically renewed unless modified. Thus, since a contingent contract is never modified, we can interpret it as a contract that is signed once and for all at $t = 1$.

3. Formal contracting

In this section we focus on situations where parties rely entirely on formal contracting. A simple way to capture these situations is to assume that T is finite, because in this case backward induction implies that the agent "shirks" on any task that is not covered by the formal contract, and hence informal contracting is ruled out.

To simplify computations we will consider the limit of the game as T approaches infinity.[15] However, we note that the qualitative results of this section would be preserved for any game with at least three periods, therefore they can be interpreted as applying also to relationships that have a relatively short duration. The analysis of this section can also be interpreted as applying to relationships where the horizon is indefinite but the parties for some reason cannot coordinate on more efficient "reputational" equilibria, where informal contracts are sustained by the threat of reverting to a worse equilibrium.

Another reason for analyzing the finite-T game is that it allows us to focus sharply on the tradeoffs between different modes of formal contracting (e.g. spot vs. contingent contracting) without the confounding effects due to the interaction between formal and informal contracting.

In our model, the limit of the sequence of subgame perfect equilibria as $T \to \infty$ is the Markov perfect equilibrium of the infinite-horizon game. A Markov perfect equilibrium (MPE) is a subgame perfect equilibrium (SPE) where strategies depend only on the payoff-relevant state variable (see Fudenberg and Tirole, 1991, Ch. 13). The payoff-relevant state variable has

[15] In our game the subgame perfect equilibrium of the finite-horizon game is generically unique, and if there are more than one they are payoff-equivalent. For this reason we speak of "the" equilibrium.

four components: the current state of the environment s_t, the set of default clauses $\tilde{\mathbf{g}}_{t-1}$, the set of tasks $\mathbf{M}^a_{t-1}$ for which the corresponding elementary action (a_k) has been described in the past, and the set of tasks $\mathbf{M}^e_{t-1}$ for which the corresponding elementary event (e_k) has been described in the past.[16] We will denote the overall state variable by $X_t = (s_t, \tilde{\mathbf{g}}_{t-1}, \mathbf{M}^a_{t-1}, \mathbf{M}^e_{t-1})$. Notice that, since the agent's current choice has no impact on future states, Markov strategies do not depend on past actions of the agent. This implies that in the MPE the agent takes the inefficient action for every task not covered by the contract.

In the MPE, the wage m_t is set at the minimum level that induces the agent to accept the proposed contract. Since the determination of the wage is a trivial aspect of the analysis, we will focus on the set of clauses. Solving for the MPE boils down to maximizing the expected discounted value of the surplus net of writing costs.

To state the problem formally, define a *formal contracting policy* as a function of the form $X_t \mapsto (\mathbf{g}^A_t, \mathbf{g}^E_t) = \psi(X_t)$. This function induces, for each (t, τ, X_t) $(t, \tau \geq 0)$, a random value for the surplus net of writing costs at date $t+\tau$, which we denote $\boldsymbol{\sigma}^{\psi}_{t+\tau}|X_t$. The problem can then be stated as

$$\forall X_t, \forall \tau \geq 0, \quad \max_{\psi} \mathrm{E}\left[\sum_{\tau=0}^{\infty} \delta^{\tau} \boldsymbol{\sigma}^{\psi}_{t+\tau} | X_t\right] \tag{3.1}$$

The *optimal* contracting policy is the solution to problem (3.1).

Most of the interesting points can be brought out by considering strictly positive but *small* writing costs. Even with small writing costs the model generates rich predictions on the optimal structure of contracts. We assume that writing costs are sufficiently small that the optimal contracting policy implements the first best outcome, or in other words, "complete" contracting is optimal. A largely sufficient condition for this is:[17]

$$\forall k, \ c(a_k) < \pi_k - d_k \tag{S}$$

Later in the paper we will discuss how results change when writing costs are large, so that

[16] Formally, $\mathbf{M}^a_t$ is defined as follows: $\mathbf{M}^a_0 = \emptyset$ and $\mathbf{M}^a_t = \mathbf{M}^a_{t-1} \cup \{k : \tilde{g}_{k,t} \neq D$, or $g_{k,t} \neq D\}$, and $\mathbf{M}^e_t$ is defined as follows: $\mathbf{M}^e_0 = \emptyset$ and $\mathbf{M}^e_t = \mathbf{M}^e_{t-1} \cup \{k : \tilde{g}_{k,t} = C_k$, or $g_{k,t} = C_k\}$. We will sometimes refer to these as the "memory sets".

[17] Under this condition, in each period it is better to write a new noncontingent clause (regardless of previous clauses) than letting the agent take the inefficient action, therefore the optimal contracting plan will implement the first best.

"incomplete" contracting may be optimal.

Given our assumptions, we can derive the optimal contracting policy by looking separately at each task k. For this reason, to keep the exposition simple we will drop the task subscript k for the remainder of this section. We will reintroduce it when we examine informal contracting, as there will be nontrivial interactions among tasks.

It can be shown that there is no loss of generality in focusing on the following candidate *rules* for a given task:

(i) At time $t = 1$ a C clause is written, and it is never modified. We will refer to this as a *contingent* rule, and denote it $\mathcal{C}$.

(ii) At time $t = 1$ a default clause R is written, and it is amended every time the realization of s_t changes. We refer to this as a *default-cum-amendments* rule, and denote it $\mathcal{DA}$.

(iii) At time $t = 1$ a default clause R is written, and an exception $\bar{R}$ is introduced whenever $s_t(e) = 0$. We will refer to this as a *default-cum-exceptions* rule, and denote it $\mathcal{DE}$.

(iv) At $t = 1$ a clause R is written, and it is permanently replaced by a clause C the first time $s_t(e) = 0$ occurs. We refer to this as an *enrichment* rule, and denote it $\mathcal{E}$.

One can show that no other rule can be strictly optimal. If we dropped our assumption that the initial state is $s_1(e) = 1$, other rules could be strictly optimal, but the essence of the results would not change.[18]

We think of the $\mathcal{DA}$ and $\mathcal{DE}$ rules as forms of "spot" contracting, whereas the $\mathcal{C}$ rule represents a purely contingent approach. The enrichment rule $\mathcal{E}$ does not fall squarely into either category of spot or contingent contracting, because it starts with a noncontingent clause and switches to a contingent clause when a low-probability event occurs.

The following result characterizes the optimal rule for a given task. The proof of this and

[18] Note that a *rule* as specified above does not completely determine a *policy*, because it specifies exceptions and amendments only for reacheable states. For example, the rule saying that clause C has to be written at time $t = 1$ and kept thereafter does not specify what to do after C is removed. This partial description of the optimal policy is sufficient for our purposes.

all other results is relegated to the appendix.

Proposition 1. *Let* $\gamma = \frac{c(e)}{c^r(a)}$. *Then, under assumption (S):*
(i) If $q > 1/2$, *the optimal rule is* $\mathcal{C}$ *if* γ *is small,* $\mathcal{E}$ *if* γ *is intermediate, and* $\mathcal{DA}$ *if* γ *is high.*
(ii) If $q < 1/2$, *the optimal rule is* $\mathcal{C}$ *if* γ *is small,* $\mathcal{E}$ *if* γ *is intermediate, and* $\mathcal{DE}$ *if* γ *is high.*
(The set of γ *values for which* $\mathcal{E}$ *is optimal is nonempty if and only if* $q + p \geq 1$.*)*

This proposition highlights the key tradeoffs involved in the choice of formal contracts. The first, simple tradeoff is between the two forms of spot contracting, $\mathcal{DA}$ and $\mathcal{DE}$. It is intuitive that the $\mathcal{DA}$ rule dominates the $\mathcal{DE}$ rule if $q > 1/2$, and viceversa if $q < 1/2$: if the *ex ante* less likely state $[s(e) = 0]$ is not persistent, it is efficient to use exceptions, but if it has a tendency to persist then amendments are more efficient.

The parameter γ captures the cost of describing an elementary event relative to the cost of recalling an already-described action. Intuitively, if γ is sufficiently low a pure contingent approach is optimal, and if it is sufficiently high a pure spot approach ($\mathcal{DA}$ or $\mathcal{DE}$) is optimal. As a consequence, if $\mathcal{E}$ is ever optimal, this will be the case for intermediate values of γ. Proposition 1 also states that this set of γ values is nonempty if and only if $q + p \geq 1$. The sum $q + p$ captures the overall tendency of the current state to persist, therefore $\mathcal{E}$ is more likely to be optimal when the state is persistent. To gain intuition for this result, consider first the extreme case in which the (s_t) process is *i.i.d.* $(q + p = 1)$. In this case, since the environment is stationary also the optimal rule is stationary: non-stationary rules like $\mathcal{E}$ are dominated. Now focus on the opposite case in which the process is highly autocorrelated ($q + p$ close to 2, which implies p close to 1): in this case, $\mathcal{E}$ dominates $\mathcal{C}$, because it implies lower cost at $t = 1$ while the additional cost $(c^r(a) + c(e))$ is expected to be incurred far in the future; and if γ is relatively low, $\mathcal{E}$ will also dominate the spot rules ($\mathcal{DA}$ and $\mathcal{DE}$).

Notice that, absent writing costs, the model has little predictive power, because there is a vast multiplicity of optimal contracting plans. Any contracting plan that implements the first best is optimal. These include a complete long-term contingent contract, a sequence of complete spot contracts, and a whole host of intermediate solutions. However, as the analysis

makes clear, an arbitrarily small writing cost is sufficient to pin down a unique optimum and deliver strong predictions on the optimal structure of formal contracts.

An interesting prediction generated by the model is that, if there is sufficient heterogeneity across tasks, so that the enrichment rule $\mathcal{E}$ is optimal for a subset of the tasks, the contract becomes gradually more complex over time, as noncontingent clauses get replaced by contingent clauses. In other words, given enough task heterogeneity, the number of contingent clauses increases over time. This matches a phenomenon that has been documented empirically, for example by Meihuizen and Wiggins (2000) for the case of supply contracts in the natural gas industry in the United States. The interpretation of this phenomenon suggested by our model is that it may be efficient to introduce a new contingent clause in the contract when a new event occurs that was previously considered unlikely.

The model also generates interesting comparative-statics predictions on the impact of changes in the contractual environment on the choice between spot and contingent contracting. The relative importance of spot contracting can be captured by the fraction of tasks regulated by a pure spot rule ($\mathcal{DA}$ or $\mathcal{DE}$). Recall that, even though we dropped the task index k from the notation, tasks may be heterogenous with respect to any of the relevant parameters, so the optimal contracting plan may regulate different tasks in different ways.

First we examine the impact of a change in the degree of uncertainty in the environment. We can think of two simple ways to parametrize the degree of uncertainty in this model. The first is to consider the case of symmetric persistence, that is $p = q$. The second is to focus on the *i.i.d.* case, that is $p = 1 - q$. In both cases, an increase in p decreases the degree of uncertainty (recalling the restriction $p \geq 1/2$).[19] If the p parameter varies across tasks, one can increase (weakly) all the p_ks.

We find that, both in the case of symmetric persistence and in the case of *i.i.d.* shocks, an increase in uncertainty implies a decrease in the fraction of tasks regulated by spot contracting. The intuition is simple: if uncertainty is higher, the external state is expected to fluctuate more over time, and this increases the expected cost of using a spot approach, while the cost of a

[19] There is also a third way to increase uncertainty: one can simply decrease p_k holding q_k constant for each k. Our result holds also in this case.

contingent contract is not affected by uncertainty.

Another parameter that affects the tradeoff between contingent and spot contracting is the discount factor δ. We find that the fraction of tasks regulated by spot contracting is decreasing in δ. The intuitive reason is that a spot approach saves on the cost of describing external events in the first period at the price of paying the cost of describing (recalling) actions in future periods. The following proposition summarizes these comparative statics results:

Proposition 2. *Under assumption (S):*
(i) Both in the symmetric-persistence case and in the i.i.d. *case, an increase in uncertainty leads to a decrease in the fraction of tasks regulated by spot contracting.*
(ii) The fraction of tasks regulated by spot contracting is decreasing in δ.

It is worth highlighting an implication of the argument made above about the impact of uncertainty: if uncertainty varies across tasks, then tasks characterized by a higher degree of uncertainty are more likely to be regulated by contingent clauses, while lower-uncertainty tasks are more likely to be regulated by a spot approach.

3.1. The role of language

Next we argue that the predictions of the model depend on our language-based approach, and could differ radically if writing costs were modeled in a different way. To make this point, we consider a more 'traditional' specification of writing costs, similar in spirit to Dye (1985a).

Let $\{s^1, ..., s^M\}$ be the set of states and $\{b^1, ..., b^M\}$ the set of behaviors (where $M = 2^N$), and assign indices so that it is efficient to do b^j if and only if the state is s^j, for all j. Now assume that, unlike our model, it is not possible to break down the description of a state or behavior into its elementary constituents. Let c^s be the cost of describing a state and c^b the cost of describing a behavior, and suppose that c^s and c^b are small, so that it is optimal to implement the first best outcome. Keep all other assumptions of our model unchanged.

In this version of the model, if N is not too small, a sequence of spot contracts is optimal, and in particular it dominates a contingent contract. To see this, note that a complete contingent

contract must specify the efficient behavior b^j for each state s^j and therefore its cost (paid once and for all) is $2^N(c^s + c^b)$; while a sequence of spot contracts costs at most c^b in each period, therefore its discounted cost does not exceed $\frac{c^b}{1-\delta}$.

Thus, this alternative specification of writing costs implies that spot contracting is typically optimal. This is in stark contrast with our model, where a contingent contract may well be optimal even if N is large; indeed, with our specification of writing costs the optimal Markovian policy is essentially independent of N (recall that the problem is separable in the N tasks). For example, suppose $N \geq \log_2(\frac{1}{1-\delta})$. Then, under the specification *à la* Dye, contingent contracting is dominated for any c^s and c^b, whereas under our specification contingent contracting is optimal for a whole region of parameters (in particular, if $c(a_k) < \pi_k - d_k$ and $\frac{c(e_k)}{c^r(a_k)}$ is sufficiently small for all k).

This should clarify our statement that the nature of language matters for the predictions of the theory. The question of which type of language is more relevant is an open one, but we have argued elsewhere (Battigalli and Maggi, 2002) that a language of the type considered in this paper is likely to be more efficient than a Dye-type language, and that it is closer to the languages that we observe in reality.[20]

3.2. The Maskin-Tirole argument

In a well-known 1999 paper, Maskin and Tirole (henceforth, MT) have argued that unforeseen contingencies (or, by a straightforward extension of their argument, the costs of describing contingencies) do not imply inefficiencies in contracting, provided that an appropriate mechanism is played after contingencies are observed and before actions are taken. In this section we argue

[20] Another remark about language is in order. We assumed that the language described at the outset is the only common-knowledge language. In principle, the parties could construct a new language, for example by attaching a new primitive sentence to each state and to each behavior, and write a contract with the new language. Note that the parties would have to attach a vocabulary that translates the new language into the original one, in order for the courts to be able to interpret the contract. If the relationship is one-shot, the new language cannot be more efficient than the original one, because the cost of writing the vocabulary in the contract is at least as large as its benefits. In a repeated relationship, however, this approach might in principle be efficient. (We thank Leonardo Felli and Luca Anderlini for bringing this point to our attention.) A more general model would allow for this kind of recoding of the language, but we conjecture that the main qualitative results would not be affected. We already allow for the possibility of "recalling" the description of an action a at low cost; the effect of coding would probably be very similar.

that (i) the results of our model would not change if such mechanisms were available, and (ii) at a more general level, the MT irrelevance argument tends to break down if the contractual imperfection takes the form of writing costs (rather than unforeseen contingencies) and the parties interact repeatedly.

That our results are unaffected by the availability of MT mechanisms is straightforward. The key observation is that a MT mechanism offers no advantage relative to a spot contract. To see this, first note an important difference between our contractual setting and MT's setting: in our model there is no ex-ante, unverifiable investment action. For this reason, in our model spot contracts are sufficient to implement the efficient behavior. The second observation is that in our setting it is costly to describe behavior, even ex post, and MT mechanisms cannot help to reduce these costs: the cost of implementing a given behavior b_t through a MT mechanism is at least as high as the cost of implementing b_t through a spot contract.[21] It follows immediately that MT mechanisms are redundant in our setting.

As a premise to point (ii), we highlight a conclusion of our model that is in apparent contrast with MT's argument: both the cost of describing behavior and the cost of describing contingencies matter for welfare. An increase in the cost of describing behavior may reduce welfare because it increases the cost of using spot contracts. More interestingly, an increase in the cost of describing contingencies may reduce welfare, because under some conditions it is efficient to write contingent clauses even though it is feasible to write spot contracts.

This conclusion is more general than our model, and applies also to richer contractual settings with ex-ante investments. For example, if MT's model is modified by replacing the assumption of unforeseen contingencies with the assumption that it is costly to describe contingencies and actions, and by allowing for repeated interaction, the irrelevance argument will no longer hold. In particular, both the costs of describing contingencies and the costs of describing actions will matter for welfare. The reason is similar to the one explained in the previous paragraph. In a richer setting of this type, it may well be efficient to use a MT mechanism, but as we argued above, an increase in the cost of describing actions increases the cost of using a MT mechanism, and hence may decrease welfare. Moreover, if the interaction is repeated,

[21] Recall that, in a MT mechanism, at the ex-post stage the behavior to be implemented must be described formally, i.e. in such a way that it can be enforced in court.

it may be efficient to write a contingent contract even though MT mechanisms are available, because the cost of using MT mechanisms is incurred repeatedly, whereas the cost of writing a contingent contract is not. Therefore, also the costs of describing contingencies matter for welfare.

4. Formal and informal contracting

In reality, long-term relationships are often governed by informal contracts, or by a combination of formal and informal contracts. Informal contracts have an important advantage over formal contracts, namely that they can be based on informal communication (i.e. communication for the only purpose of reciprocal understanding), as opposed to formal communication (i.e. communication for the purpose of making the contract enforceable in court). Arguably, the cost of the latter is higher than the cost of the former, because for the contract to be enforced by courts it must be written according to the commonly accepted legal standards, which may be quite cumbersome to meet. In particular, it is not sufficient that the language used in the formal contract be common knowledge to the contracting parties; it has to be common knowledge to the parties *and* the courts, and this may require effort and skills (or lawyers).

The shortcoming of informal contracts, on the other hand, is the absence of an external enforcement mechanism. Since an informal contract must be self-enforcing, if players are not sufficiently patient it may not be possible to implement the first best outcome with a fully informal contract. In what follows we will examine more closely this trade-off between formal and informal contracting.

We assume that there is no cost associated with informal contracting. We could introduce a cost of informal contracting, but this would change the main results in an obvious direction.

Consider the game of section 2. The way we allow for informal contracts is by focusing on the constrained Pareto-efficient SPE of the infinite-horizon game. In such equilibria cooperation may be sustained by formal rules and/or by informal rules: formal rules are enforced by external courts, while informal rules are enforced by the threat of reverting to a worse equilibrium. The assumption that players focus on a constrained Pareto-efficient SPE is standard in the literature

on self-enforcing contracts. This assumption seems reasonable for situations in which there is sufficient mutual understanding between players, so that they are able to coordinate on a "good" equilibrium.

Our objective is to understand under what conditions it is efficient to govern the relationship by formal contracting, by informal contracting, or by a combination of the two, and in the latter case, which tasks are regulated formally and which are regulated informally. Finally, we will explore how some key parameters affect the optimal mix of formal and informal contracting.

The first step in the analysis of efficient SPE is to characterize the optimal punishment strategies. In this type of game it is efficient to punish a deviator with his maxmin payoff, if this punishment is credible. We now show that, if players are sufficiently patient, maxmin punishments are indeed credible. We already know that the MPE keeps the agent at his maxmin. The following lemma identifies a critical level of δ (which is close to $\frac{1}{2}$ when writing costs are small) above which there exists also a credible punishment strategy that keeps the principal at his maxmin.

Lemma 1. *There exists δ^* (function of other parameters) such that for $\delta \geq \delta^*$ there is a SPE that keeps the continuation payoff of the principal at his maxmin (zero) in every subgame starting with a move by the principal. (The critical level δ^* approaches $\frac{1}{2}$ as writing costs become negligible.)*

To convey the basic intuition we focus on the case in which writing costs are negligible and argue that the principal can be kept at her maxmin provided δ is slightly above $\frac{1}{2}$. Consider a punishment strategy with the following structure: after a deviation, the informal contract is abandoned and parties revert to the optimal formal contracting plan (the MPE), with the surplus going entirely to the player that has not deviated. If the principal offers a formal contract that she is not supposed to offer, the agent is supposed to reject it unless it gives him the full current net surplus. The key incentive-compatibility condition that this punishment strategy has to satisfy concerns a situation where the principal is being punished. In this situation, the principal has the option of tempting the agent with a "sweet deal" that gives the agent part of the current net surplus. According to the candidate equilibrium strategies, by accepting

this offer the agent would lose all the future net surpluses. Neglecting writing costs, a sufficient condition under which the agent will not accept this offer is $\sum_{k\in\mathbf{N}}(\pi_k - d_k) < \frac{\delta}{1-\delta}\sum_{k\in\mathbf{N}}(\pi_k - d_k)$. The left hand side of this inequality is an upper bound to the benefit of accepting a "sweet deal". The right hand side is the present value of future surpluses, which is the opportunity cost of accepting (assuming that the candidate equilibrium strategies are followed from the next period). If $\delta > \frac{1}{2}$ this condition is satisfied, and the proposed punishment strategy is credible.

The condition $\delta \geq \delta^*$ is essential for Lemma 1. Consider the extreme case of δ equal to zero: then there is a unique SPE (the Markovian equilibrium) in which the principal offers a formal contract and makes a positive profit in every period. However we suspect this condition is not essential for our qualitative results. If it is not satisfied, it may not be possible to keep the principal at his maxmin payoff in the punishment phase, in which case the principal's incentive constraints will be more stringent, and this is likely to result in fewer tasks being regulated informally, but our qualitative results are still likely to hold.

Having discussed the off-equilibrium-path strategies, we can now focus on the equilibrium path. We continue to assume that writing costs are small, in the sense that condition (S) holds. Under this condition, it can be shown that an efficient equilibrium must implement the first best, just as in the formal-contracting analysis. It is then natural to focus on simple equilibrium paths where (i) each task k is either regulated by one of the first-best formal rules described in the previous section ($\mathcal{C}_k$, or $\mathcal{E}_k$ or $\mathcal{DE}_k$ or $\mathcal{DA}_k$), or by an *informal* contingent rule prescribing the first-best action, and (ii) the agent always accepts the proposed formal contract $(\mathbf{g}_t, m_t)$ and takes the first-best action for each task regulated by an informal rule.[22]

Such an equilibrium path is described by a tuple $(\mathbf{I}, \mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA}, (m_t)_{t\geq 1})$, where bold capital letters denote subsets of tasks, and m_t is a random variable, that is, $m_t : H_t \to \mathbb{R}$, where is H_t the set of histories of shocks $h_t = (s_1, ..., s_t)$.[23] $(\mathbf{I}, \mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA})$ is an ordered partition of the set of tasks $\mathbf{N} = \{1, ..., N\}$ specifying that tasks in subset $\mathbf{I}$ are agreed-upon informally and tasks in $\mathbf{C}$ (respectively, $\mathbf{E}, \mathbf{DE}, \mathbf{DA}$) are regulated formally *via* a contingent

[22] Recall that, given the assumed rules of the game, the principal must pay the exact wage specified in the formal contract, m_t. As we argue at the end of this section, in this model there is nothing to gain from paying informal bonuses to the agent.

[23] It should be clear that, even if the wage is state-dependent, it is not written as a contingent wage in the formal contract, but it is written period by period after observing the state s_t, so it involves no writing costs.

(respectively, an enrichment, default cum exception, default cum amendment) rule. We will refer to a tuple $(\mathbf{I}, \mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA})$ as a *task partition.*

We say that a task partition is *incentive compatible* if it is part of a SPE. We say that a task partition is *efficient* if it is incentive compatible and there is no incentive compatible task partition that yields a higher expected present value of the net surplus. Notice that in our model this is equivalent to the standard definition of constrained Pareto-efficiency.[24]

Next we derive a necessary condition for the incentive compatibility of a task partition. If some tasks are regulated by informal rules, the agent has the opportunity to shirk on those tasks (i.e. increase his current utility by not taking the efficient action). The most effective way to prevent the agent from shirking on the informal tasks is to give him all the net surplus from the second period onward if he does not shirk and to keep him at his maxmin if he shirks. Then the agent will not shirk only if the present value of future expected net surpluses is at least as large as the current benefit from shirking. To express this condition formally, it is convenient to introduce some more notation. Let $\pi(\mathbf{K}) = \sum_{k \in \mathbf{K}} \pi_k$, $d(\mathbf{K}) = \sum_{k \in \mathbf{K}} d_k$, and let $\widehat{c}_{t+1}(\mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA} | h_\tau)$ denote the expected present value of writing costs from time $t+1$ conditional on history h_t. Then it is clear that a task partition is incentive compatible only if the following set of aggregate incentive constraints hold:

$$\forall t \geq 1, \forall h_t \in H_t, \qquad \frac{\delta}{1-\delta}[\pi(\mathbf{N}) - d(\mathbf{N})] - \delta \widehat{c}_{t+1}(\mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA} | h_t) \geq d(\mathbf{I}) \qquad \text{(IC)}$$

The left hand side of (IC) is the expected present value of future net surpluses conditional on history h_t. The right hand side of (IC) is the agent's disutility from performing the informal tasks, which is also the benefit from shirking. A task partition is incentive compatible only if the latter does not exceed the former, in any period and given any history.

If $\delta \geq \delta^*$, so that maxmin punishments are credible, condition (IC) is also sufficient for

[24] The expected net surplus is simply the sum of the expected payoffs of principal and agent. Therefore, constrained maximization of the PDV of expected net surplus implies constrained Pareto-efficiency. To see why also the converse holds, suppose that an incentive compatible task partition $\mathbf{P}$ yields a lower PDV of expected net surplus than incentive compatible partition $\mathbf{Q}$ and pick SPE paths $\rho_{\mathbf{P}}$ and $\rho_{\mathbf{Q}}$ respectively implementing $\mathbf{P}$ and $\mathbf{Q}$. Then it is possible to redistribute the surplus in $\rho_{\mathbf{Q}}$ modifying the first period wage so that both parties are better off than with $\rho_{\mathbf{P}}$ and both get a positive PDV of expected payoff. This does not alter the incentives to deviate in future periods, therefore the modified path $\rho'_{\mathbf{Q}}$ is also a SPE path. It follows that $\rho_{\mathbf{P}}$ is not constrained Pareto-efficient.

incentive compatibility. To see this, suppose that a task partition satisfies (IC) and $\delta \geq \delta^*$. Then a SPE inducing this task partition can be constructed as follows. As long as no deviation occurs, each task is regulated according to the partition. From the second period on, the principal transfers all the net surplus to the agent, while the first period transfer is such that both parties get a positive share of the present value of expected net surpluses. If a deviation occurs, the deviator is punished by keeping him/her at the maxmin. By construction, these strategies give the principal no incentive to deviate. Condition (IC) ensures that also the agent has no incentive to deviate. Therefore we obtain the following:

Lemma 2. *Suppose condition (S) holds, and $\delta \geq \delta^*$ (where δ^* is defined as in Lemma 1). Then a task partition is efficient if and only if it solves the following problem*

$$\min_{\mathbf{I},\mathbf{C},\mathbf{E},\mathbf{DE},\mathbf{DA}} \widehat{c}_1(\mathbf{C},\mathbf{E},\mathbf{DE},\mathbf{DA}) \qquad (P)$$
$$s.t.\ (IC)$$

The solution to problem (P) in some cases may not be unique, for the following reason. Since the set of task partitions is finite, a solution will typically not satisfy the (IC) constraint as an equality. It may therefore happen that switching the rules of some tasks from formal to informal and *viceversa* does not violate (IC) and has no impact on the objective function. This is the case, for example, if tasks j and k have the same writing costs, and at a solution of (P), j is regulated by informal rule and k by contingent rule. This example, however, holds for nongeneric values of the cost parameters (the contingent clauses C_j and C_k must have the same writing cost). Indeed, it can be shown that the solution to (P) is *generically* unique.[25] From now on we take for granted that the assumptions of Lemma 2 hold and that the solution to (P) is unique.

The next proposition states conditions under which the optimum is fully informal, or fully formal, or a mix of formal and informal contracting. Suppose without loss of generality that $d_1 = \min_k d_k$.

[25] More precisely, it can be shown that the set of parameter vectors for which there are multiple solutions is *nowhere dense* in the parameter space (i.e., its closure has empty interior).

Proposition 3. *At the solution of problem (P):*
(i) If $d(\mathbf{N}) \leq \delta\pi(\mathbf{N})$*, all tasks are regulated informally.*
(ii) If $d(\mathbf{N}) > \delta\pi(\mathbf{N})$*,* $d_1 < \frac{\delta}{1-\delta}[\pi(\mathbf{N}) - d(\mathbf{N})]$*, and writing costs are sufficiently small, some tasks are regulated formally and some are regulated informally.*
(iii) If $d_1 > \frac{\delta}{1-\delta}[\pi(\mathbf{N}) - d(\mathbf{N})]$*, all tasks are regulated formally.*

This result states that formal and informal contracting coexist if the total disutility incurred by the agent over all tasks is relatively large ($d(\mathbf{N}) > \delta\pi(\mathbf{N})$) and there is at least one task characterized by a relatively low disutility ($d_1 < \frac{\delta}{1-\delta}[\pi(\mathbf{N}) - d(\mathbf{N})]$). If the first condition is not satisfied, then a fully informal contract is incentive compatible, and hence efficient; if the second condition is not satisfied, then introducing any of the tasks in the informal contract violates (IC), therefore the optimum is fully formal.[26]

An important aspect of this result is that, as long as d_1 is relatively low, the optimal contract is partly or fully informal, and this is true even if writing costs are very small. This contrasts sharply with Baker *et al* (1994): in their model, if the imperfection in formal contracting is small, it is impossible to sustain any informal contracting.

We can say something more on the impact of writing costs on the optimal mix of formal and informal contracting. As the following remark states, a reduction in writing costs may even *facilitate* informal contracting:

Remark 1. *As writing costs decrease, the fraction of tasks regulated informally may increase.*

If writing costs are heterogenous across tasks, this comparative-statics exercise should be interpreted as decreasing (weakly) all writing costs. The intuition for this result can be gained by inspecting the incentive constraints (IC): decreasing writing costs reduces $\widehat{c}_{t+1}$ and hence increases the net surplus available in the relationship, thus relaxing (IC). This in turn makes it easier to sustain an informal contract.

[26] The reason we need sufficiently small costs at point (ii) is that if noncontingent formal rules are used, then the present value of future writing costs enters the incentive constraint. See the proof of Proposition 3 in the appendix.

An interesting question is: which tasks are regulated formally and which are regulated informally? Intuitively, informal contracting should be harder to sustain for tasks characterized (other things equal) by higher disutility, because the agent has a stronger incentive to shirk on these tasks. The next proposition confirms this intuition, under a *ceteris paribus* assumption on other task characteristics:

Proposition 4. *Suppose that writing costs and transition probabilities are the same for all* $k \in \mathbf{N}$. *Then, at the solution to (P), there exists a critical level* $\bar{d}$ *such that tasks with* $d_k < \bar{d}$ *are regulated informally and tasks with* $d_k > \bar{d}$ *are regulated formally.*

The basic argument behind this result is the following. Consider two tasks characterized by different disutilities, and suppose that one task must be regulated formally and the other informally. Which one will be handled informally? The surplus is the same independently of which task is chosen, but the low-d_k task implies a lower incentive to cheat, hence it is better to regulate this task informally.[27]

A natural question is how other task characteristics, and in particular the degree of uncertainty and the incremental benefit π_k, affect the choice between formal and informal contracting. Consider first the role of uncertainty. Even if tasks differ only in the degree of uncertainty, there is no clear-cut result: comparing two tasks it is possible that the higher-uncertainty task is regulated informally and the lower-uncertainty task is regulated formally, but also the viceversa is possible. This is clear if one considers the case in which all formal clauses are contingent: then uncertainty is irrelevant for the choice between formal and informal rules.

Also the task-specific benefit π_k has no role in the optimal allocation of tasks between formal and informal contracting: if tasks differ only with respect to π_k, there is no way of saying which ones are regulated formally and which ones are regulated informally. This is because the π_k parameters enter problem (P) only through their sum, $\pi(\mathbf{N})$.[28]

[27] Recall that we are assuming that the solution to (P) is unique. While this holds generically in the unrestricted parameter space, it does not hold generically in the subspace where writing costs and transition probabilities are the same. However we can show that, if there is a continuum of tasks, the optimal allocation between formal and informal contracting is unique even in this subspace and conforms to the proposition above.

[28] As the next section makes clear, this is no longer true if writing costs are large.

The model does have something to say about the effect of changing the total benefit $\pi(\mathbf{N})$ on the optimal mix between formal and informal contracting. Intuitively, an increase in $\pi(\mathbf{N})$ relaxes the incentive constraints (IC) and hence will tend to decrease the fraction of tasks that are regulated formally. The reason we speak of a *tendency* is that, due to the heterogeneity of tasks and the discrete nature of problem (P), one cannot rule out "perverse" cases in which the number of formal tasks increases even though (IC) gets relaxed.[29] We conjecture that these pathological cases would be ruled out if we had a continuum of tasks rather than a discrete number of tasks.[30]

As an alternative measure of the importance of formal contracting, one could consider the (expected present value of) writing costs at an efficient task partition. This can be interpreted as the amount of resources expended in writing formal contracts. Mathematically, this is the value of program (P), which we denote $\hat{c}$. This alternative measure does not have the shortcoming described in the previous paragraph, and allows us to state an unambiguous result:

Remark 2. *The importance of formal contracting, as captured by $\hat{c}$, is (weakly) decreasing in $\pi(\mathbf{N})$.*

This result follows immediately from the fact that an increase in $\pi(\mathbf{N})$ relaxes the incentive constraint without affecting the objective function, and hence it (weakly) decreases the value of program (P). Note that, as $\pi(\mathbf{N})$ increases, the gross potential surplus increases, and hence the gains from contracting increase. Thus, broadly speaking, the model predicts that informal contracting should be used more intensively in long-term relationships where gains from contracting are larger.

A legitimate question is whether the results of the previous section on the choice between

[29] Consider the following example. There are three tasks, tasks 1 and 2 characterized by small disutility and writing costs, task 3 by relatively high disutility and writing costs. Suppose that in the initial situation it is possible to sustain informally task 1 and 2 but not task 3, because of this task's high disutility. If $\pi(\mathbf{N})$ is increased slightly, (IC) gets relaxed, and it is possible that in the new situation it is incentive compatible to regulate informally task 3 alone, but not in conjunction with any other task. Then it is optimal to switch from regulating tasks 1 and 2 informally to regulating only task 3 informally.

[30] Extending this model to a continuum of tasks is very hard because tasks are heterogenous with respect to several dimensions, but we can prove this conjecture if writing costs and transition probabilities are the same for all tasks, as assumed in Proposition 4.

spot and contingent formal contracts still hold in the presence of informal contracting. The answer is that the qualitative insights are still valid, but we no longer have sharp results because the presence of informal contracting may act as a confounding factor. To illustrate the issue, consider an increase in δ: this tends to favor contingent contracts over spot contracts, but at the same time it relaxes the incentive constraint, and this may generate counterintuitive results. For example, if there are two tasks it is possible that for lower values of δ one task is regulated by contingent formal contracting and the other task by spot formal contracting, and for higher values of δ one task is regulated informally and the other by spot formal contracting.[31]

A distinct but related question is: how does the availability of informal contracting affect the choice between contingent and spot contracts? More precisely, is the relative importance of contingent contracting higher at the Markov perfect equilibrium or at an efficient SPE of the game? Recall that the latter differs from the former in that it allows for informal rules, and the cost-minimization problem is subject to the (IC) constraint. Intuitively, with a spot approach writing costs must be incurred repeatedly over time, and this imposes a strain on the incentive constraint; whereas with a contingent approach writing costs are incurred only at $t = 1$, and this does not create incentive problems. As a consequence, the presence of informal contracting should tend to favor contingent contracting. The following remark confirms this intuition.

Remark 3. *A given task k may be regulated by a spot rule at the Markov perfect equilibrium and by a contingent rule at an efficient SPE, but not viceversa.*

This result suggests that formal contracts should tend to be more contingent in relationships where formal and informal contracting coexist, relative to relationships that are governed exclusively by formal contracting.

Before concluding the section, we want to discuss our assumptions about timing, the possibility to pay "bonuses" (i.e. payments in excess of what is specified by the agreed-upon formal contract) and the possibility to write multi-period contracts.

[31] Something less ambiguous can be said about the impact of uncertainty on the choice between spot and contingent contracting. If tasks differ only by the degree of uncertainty, then a set of low-uncertainty tasks will be regulated by formal spot contracting, and a set of high-uncertainty tasks will be regulated by informal or contingent formal contracting.

According to the rules of the game, the principal pays the agent before he acts and the payment is exactly the one specified by the offered contract (if accepted). It is clear that each of these assumptions is without loss of generality given the other. If bonuses are not allowed, it does not matter when the payment m occurs, because the principal has to pay m independently of whether the agent has shirked on the informally regulated tasks or not. On the other hand, if any payment has to be made before the agent acts, bonuses are redundant, because the best incentive to keep the agent from shirking is still to hold him down to his maxmin from period $t+1$ if he shirks in period t. We now argue that even the *joint* assumption about timing and bonunes is without loss of generality. Suppose that the principal is allowed to pay an informal bonus immediately after the agent acts. In this case, the agent has a stronger incentive not to shirk, because shirking will prevent him from enjoying an immediate reward. On the other hand, with the modified assumption, the principal has an incentive to renege on the informally promised bonus, whereas with our current assumption he can only omit to offer the formal contract specified by the equilibrium. It can be shown that these two effects cancel out.

Similarly, it can be shown that there is no gain from writing formal multi-period contracts. To gain intuition, suppose there is only one task. Even though committing to future wages in the current contract would remove the principal's incentive to renege on payments, this would ruin the agent's incentives: if future wages are assured, the agent will surely cheat if the task is regulated informally. And if the task is regulated formally, then there is no gain relative to one-period formal contracts.

5. Large writing costs

If writing costs are large, the main change in results is that it may not be optimal to implement the first best for some (or all) of the tasks. In particular, two additional possibilities emerge: it might be optimal to regulate a task by *rigid rule*, that is by writing a noncontingent R_k clause once and for all, or to leave a task to the agent's *discretion* with no informal agreement to take the efficient action. Given our assumptions on payoffs, a rigid rule yields a gross incremental surplus that is positive but lower than a first-best rule, and leaving a task discretionary yields zero gross incremental surplus.

We interpret rigidity and discretion as two forms of contractual incompleteness, while we think of a contract implementing the first-best outcome as a "complete" contract. If one adopts this terminology, then the main implication of large writing costs is that they lead to contractual incompleteness. Next we show how the analysis can be extended to this more general case.

We drop assumption (S). The only condition we assume on writing costs is that they are non-prohibitive in the sense that for every $\delta \in [0,1)$ the optimal formal contracting policy yields a strictly positive present value of the net surplus.[32] A contracting plan can now be represented by a (possibly) non exhaustive task partition $(\mathbf{I}, \mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA}, \mathbf{R})$ and a transfer process $(m_t)_{t\geq 1}$, where $\mathbf{R}$ is the set of tasks regulated by rigid rules. The tasks that are not covered by the partition are discretionary.

Let $\hat{\pi}_{t+1}(\mathbf{R}|h_t)$ and $\hat{d}_{t+1}(\mathbf{R}|h_t)$ denote respectively the expected present value of future benefits and disutilities associated with the set of rigid rules conditional on h_t, and let $\mathbf{FB} \equiv \mathbf{I} \cup \mathbf{C} \cup \mathbf{E} \cup \mathbf{DE} \cup \mathbf{DA}$ denote the set of tasks regulated in a first best way. In this more general case the expected present value of future net surpluses is $\frac{\delta}{1-\delta}[\pi(\mathbf{FB}) - d(\mathbf{FB})] + \delta[\hat{\pi}_{t+1}(\mathbf{R}|h_t) - \hat{d}_{t+1}(\mathbf{R}|h_t)] - \delta\widehat{c}_{t+1}(\mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA}|h_t)$. By the argument of the previous section, a task partition is incentive compatible only if

$$\forall t \geq 1, \forall h_t \in H_t,$$

$$\frac{\delta}{1-\delta}[\pi(\mathbf{FB}) - d(\mathbf{FB})] + \delta[\hat{\pi}_{t+1}(\mathbf{R}|h_t) - \hat{d}_{t+1}(\mathbf{R}|h_t)] - \delta\widehat{c}_{t+1}(\mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA}|h_t) \geq d(\mathbf{I}) \quad \text{(IC')}$$

This is a generalization of the incentive constraints in the previous section. The only difference is that now the surplus terms in the left hand side must be calculated taking into account that the tasks regulated by rigid rules yield a stochastic stream of (gross) surpluses, and that discretionary tasks yield no surplus at all.

Relying on Lemma 1, we can show that (IC') is also sufficient for incentive compatibility if the discount factor is high enough. Therefore we obtain the following characterization result:

[32] This is equivalent to the following simple condition: there is at least one task j such that $c(a_j) < \pi_j - d_j$. The condition is obviously sufficient, because it allows to obtain a postitive net surplus for task j by using clause R_j or $\overline{R}_j$, as appropriate, in every period. To see that the condition is necessary, suppose that $c(a_k) \geq \pi_k - d_k$ for all k and let $\delta = 0$. Then introducing any (nonempty) clause in the contract would yield a (weakly) negative net surplus in the first period that could not be compensated by positive surpluses in future periods.

Lemma 3. *Suppose $\delta \geq \delta^*$ (where δ^* is as in Lemma 1). Then a task partition is efficient if and only if it solves the following problem*

$$\max_{\mathbf{I},\mathbf{C},\mathbf{E},\mathbf{DE},\mathbf{DA},\mathbf{R}} \left[\frac{1}{1-\delta} \left[\pi(\mathbf{FB}) - d(\mathbf{FB}) + \hat{\pi}_1(\mathbf{R}) - \hat{d}_1(\mathbf{R}) \right] - \widehat{c}_1(\mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA}, \mathbf{R}) \right] \quad (P')$$

s.t. (IC').

Note that the presence of large writing costs modifies not only the incentive constraints but also the objective function of the problem: since it may not be optimal to implement the first best outcome, we need to consider explicitly the surplus implications of different formal rules, therefore we no longer have a cost-minimization problem.

In general this is a fairly complex problem, but we can say something about the case in which tasks differ only by their disutility:

Proposition 5. *Suppose tasks differ only with respect to d_k. Then, at the solution to (P'), there exist three critical levels, $d' < d'' < d'''$, such that tasks with $d_k < d'$ are regulated by informal rule, tasks with $d' < d_k < d''$ are regulated by formal first-best rule, tasks with $d'' < d_k < d'''$ are regulated by formal rigid rule, and tasks with $d_k > d'''$ are left to the agent's discretion.*

The logic that drives this result is simple. Putting aside informal contracting for a moment, let us ask: which tasks should be regulated by first-best rule, which by rigid rule, and which should be discretionary? Note that, since the benefit π_k is assumed constant across tasks, lower d_k implies higher gross surplus $(\pi_k - d_k)$. Intuitively, high-surplus tasks are regulated by a first best rule, intermediate-surplus tasks are handled with a rigid rule, and low-surplus tasks are left to the agent's discretion. This part of the result is analogous to Proposition 1 in Battigalli and Maggi (2002), and is a consequence of the fact that the writing cost is higher for a first-best rule than for a rigid rule, and is zero for a discretionary task. Now consider also the possibility of informal contracting. As we saw in the previous section, with small writing costs, lower disutility favors informal contracting over formal contracting. This continues to be true with large writing costs, hence the ranking described in Proposition 5 follows.

Thus the analysis shows that the main consequence of large writing costs is that the optimal contracting plan may include rigidity and/or discretion. Within the set of tasks regulated by a first-best rule, the qualitative insights developed in the previous sections on the choice between formal and informal contracting and on the optimal structure of formal contracts should still be broadly valid.

Another interesting point that is brought out in the presence of large writing costs is the following. Formal and informal contracting not only tend to be used jointly, but are *complementary* in a stricter sense: when some tasks are discretionary, increasing the number of tasks regulated formally (for example as a consequence of a reduction in writing costs) increases the surplus from the relationship, thus relaxing the incentive constraint and making it easier to regulate more tasks informally.[33]

We conclude with a note on the nature of writing costs in our model. We considered only "variable" writing costs, that is costs that increase with the number of modifications to the contract. One may also consider the impact of two types of "fixed costs": costs that are incurred the first time a contract is offered, and "quasi-fixed" costs, that are incurred every time the contract is modified. Fixed costs of the first kind would not affect our results insofar as they are not prohibitive, i.e. they do not prevent formal contracting. Quasi-fixed writing costs would affect the mode of formal contracting in two ways: first, they would tilt the balance in favor of contingent contracting as opposed to spot contracting, since the latter involves contract modifications over time and the former does not; second, if spot contracting is nevertheless efficient, there will be a tendency to postpone some contract modifications so as to "cluster" several modifications in a single period. This would complicate the analysis without adding much insight.

6. Conclusion

In this paper we have implicitly assumed away an alternative mode of governance that could avoid the costs of writing detailed contracts, namely giving *authority* to the principal. If the

[33] A similar complementarity between formal and informal contracting is highlighted by Baker *et al.* (1994).

principal could instruct the agent on what actions to take in each period, there would be no need to specify contingencies or actions in a contract. In this concluding section we discuss how results would change if we allowed for authority as a governance mode.

As a premise, it is useful to distinguish between *formal* and *informal* authority. We speak of formal authority when the principal's authority is specified in the formal contract, and the agent can be punished by courts for disobeying the principal. We speak of informal authority when the principal's orders are not enforced by courts, but by credible punishment mechanisms.

Let us focus on formal authority first. It is critical to note that formal authority is enforceable only if two conditions are met: (i) the principal can send verifiable messages to the agent; this requires that messages be written, or at least recorded; and (ii) messages must be expressed in a language understood by the courts. In other words, messages must be *formal*. For this reason, even if a system of formal messages is feasible, it is not clear that its costs would be significantly lower than a system of formal contracts. This might also explain why formal-authority relationships are not observed very often in reality.

Informal authority is a more common mode of governance in real organizations. In our model, however, there is no role for such an arrangement, due to the assumption of symmetric information. Given that the principal and the agent have the same information, informal authority cannot improve on an informal contract as we defined it, because in the latter arrangement the agent knows what actions to take under any contingency, hence there is no need for further instructions from the principal. A role for informal authority would probably arise if the principal had private information. An extension of the model in this direction is left for future research.

7. Appendix

Proof of Proposition 1

First note that $\mathcal{DE}$ dominates $\mathcal{DA}$ if $q < 1/2$, and viceversa if $q > 1/2$. So we can partition the set of tasks into two groups, those for which $q < 1/2$ and those for which $q > 1/2$, and look at each subset separately (we ignore the knife-edge case).

(i) Case $q < 1/2$.

The candidate plans are $\mathcal{C}$, $\mathcal{E}$ and $\mathcal{DE}$. Since all three plans implement the first best, we only need to compare the present value of writing costs:

Present value of cost under plan $\mathcal{C}$:

$$\widehat{c}(\mathcal{C}) = c(a) + c(e)$$

Present value of expected costs under plan $\mathcal{E}$:

$$\widehat{c}(\mathcal{E}) = c(a) + [c^r(a) + c(e)] \sum_{t=2}^{\infty} \delta^{t-1} (p)^{t-2} (1-p) = c(a) + [c^r(a) + c(e)] \frac{\delta(1-p)}{1-\delta p}$$

Present value of expected costs under plan $\mathcal{DE}$:

$$\widehat{c}(\mathcal{DE}) = c(a) + c^r(a) \sum_{t=2}^{\infty} \delta^{t-1} \Pr[s_t(e) = 0] = c(a) + c^r(a) \frac{\delta(1-p)}{(1-\delta)[1-\delta(p+q-1)]}$$

Note that (1) $\mathcal{C}$ is preferable to $\mathcal{DE}$ iff $\gamma < \gamma^*_{\mathcal{C}/\mathcal{DE}} = \frac{\delta(1-p)}{(1-\delta)[1-\delta(p+q-1)]}$; (2) $\mathcal{E}$ is preferable to $\mathcal{C}$ iff $\gamma > \gamma^*_{\mathcal{C}/\mathcal{E}} = \frac{\delta(1-p)}{1-\delta}$; (3) $\mathcal{E}$ is preferable to $\mathcal{DE}$ iff $\gamma < \gamma^*_{\mathcal{E}/\mathcal{DE}} = \frac{(1-\delta p)}{(1-\delta)[1-\delta(p+q-1)]} - 1$. This proves the first part of the claim. Rule $\mathcal{E}$ may be optimal iff $\gamma^*_{\mathcal{C}/\mathcal{E}} \leq \gamma^*_{\mathcal{E}/\mathcal{DE}}$, which is true iff $q + p \geq 1$.

(ii) Case $q > 1/2$.

The candidate plans are $\mathcal{C}$, $\mathcal{E}$ and $\mathcal{DA}$. Again, since all three plans implement the first best, we only need to compare the present value of expected costs under the different plans: $\widehat{c}(\mathcal{C})$, $\widehat{c}(\mathcal{E})$ and

$$\widehat{c}(\mathcal{DA}) = c(a) + c^r(a) \sum_{t=2}^{\infty} \delta^{t-1} \Pr[s_t(e) \neq s_{t-1}(e)] = c(a) + c^r(a) \frac{\delta(1-p)[1-\delta(2q-1)]}{(1-\delta)[1-\delta(p+q-1)]}$$

Note that $\mathcal{C}$ is preferable to $\mathcal{DA}$ iff $\gamma < \gamma^*_{\mathcal{C}/\mathcal{DA}} = \frac{\delta(1-p)[1-\delta(2q-1)]}{(1-\delta)[1-\delta(p+q-1)]}$; (ii) $\mathcal{E}$ is preferable to $\mathcal{C}$ iff $\gamma > \gamma^*_{\mathcal{C}/\mathcal{E}}$; (iii) $\mathcal{E}$ is preferable to $\mathcal{DA}$ iff $\gamma < \gamma^*_{\mathcal{E}/\mathcal{DA}} = \frac{\delta(1-q)[1-\delta(2p-1)]}{(1-\delta)[1-\delta(p+q-1)]}$. Rule$\mathcal{E}$ may be optimal iff $\gamma^*_{\mathcal{C}/\mathcal{E}} \leq \gamma^*_{\mathcal{E}/\mathcal{DA}}$, or equivalently $p \geq q$, which holds by assumption.

Proof of Proposition 2

(i) Both in the *i.i.d.* case and in the symmetric-persistence case the relevant thresholds for the parameter $\gamma = \frac{c(e)}{c^r(a)}$ (defined in the proof of Proposition **1)** collapse to the same value $\gamma^*_{\mathcal{C}/\mathcal{DE}} = \gamma^*_{\mathcal{E}/\mathcal{DE}} = \gamma^*_{\mathcal{E}/\mathcal{DA}} = \frac{\delta(1-p)}{(1-\delta)}$, which is decreasing in p. By Proposition 1, the number of tasks regulated in a "spot" way is therefore decreasing as uncertainty increases (p decreases).

[Using the general expressions for $\gamma^*_{\mathcal{C}/\mathcal{DE}}$, $\gamma^*_{\mathcal{E}/\mathcal{DE}}$ and $\gamma^*_{\mathcal{E}/\mathcal{DA}}$ it is also easy to check that these thresholds are decreasing in p keeping q fixed. This proves the claim made in footnote that the number of tasks regulated in a "spot" way is increasing if, for all tasks k, p_k decreases with q_k fixed, which is another way to increase uncertainty.]

(ii) We verify that the thresholds $\gamma^*_{\mathcal{C}/\mathcal{DE}}$, $\gamma^*_{\mathcal{E}/\mathcal{DE}}$ and $\gamma^*_{\mathcal{E}/\mathcal{DA}}$ are increasing in δ:

$$\frac{\partial \gamma^*_{\mathcal{C}/\mathcal{DE}}}{\partial \delta} = (1-p)\frac{1+\delta^2[1-p-q]}{(1-\delta)^2[1+\delta(1-p-q)]^2} > 0$$

$$\frac{\partial \gamma^*_{\mathcal{E}/\mathcal{DE}}}{\partial \delta} = \frac{[1-\delta(p+q-1)](1-p)+(p+q-1)(1-\delta)(1-\delta p)}{(1-\delta)^2[1-\delta(p+q-1)]^2} > 0.$$

$$\frac{\partial \gamma^*_{\mathcal{E}/\mathcal{DA}}}{\partial \delta} = (1-q)\frac{[1-\delta(2p-1)]\{(1-\delta)+\delta(1-q)[1-\delta(p+q-1)]\}}{(1-\delta)^2[1-\delta(p+q-1)]^2} > 0$$

By Proposition 1, it follows that the number of tasks regulated in a "spot" way is decreasing in δ.

Proof of Lemma 1

Let $\overline{c}_{DE}$ be the average per period cost of using rule $\mathcal{DE}_k$ for every task k with $c(a_k) < \pi_k - d_k$ in the worst case scenario where the current state is $(s(e_1), ..., s(e_N)) = (0, ..., 0)$ and the cost of recalling the description of an action is the same as the cost of describing it for the first time, that is

$$\overline{c}_{DE} = \sum_{k:c(a_k)<\pi_k-d_k} \frac{(1-\delta p_k)}{[1-\delta(p_k+q_k-1)]} c(a_k)$$

[see the proof of Proposition 1 and let $c^r(a_k) = c(a_k)$]. Since $\frac{(1-\delta p_k)}{[1-\delta(p_k+q_k-1)]} < 1$ and writing costs are non-prohibitive, we have $\overline{c}_{DE}<\pi(\mathbf{N}) - d(\mathbf{N})$. Let δ^* be the largest discount factor satisfying the following inequality:

$$\pi(\mathbf{N})-d(\mathbf{N}) \geq \frac{\delta}{1-\delta}\left[\pi(\mathbf{N}) - d(\mathbf{N}) - \overline{c}_{DE}\right].$$

Note that $\delta^* > \frac{1}{2}$, but $\delta^* \to \frac{1}{2}$ as $c(a_k) \to 0$ for each k.

We exhibit a SPE keeping the principal at his maxmin under the parameter restriction $\delta \geq \delta^*$. Consider the following strategies: for all $(\widetilde{\mathbf{g}}, \mathbf{M}, s)$ (where $\widetilde{\mathbf{g}}$ is the set of default clauses inherited from the previous period and $\mathbf{M} = (\mathbf{M}^a, \mathbf{M}^e)$ is the pair of memory sets inherited from the previous period) the principal makes amendments and exceptions as in the MPE. Wages are determined according to a punishment phase.[34] There are two punishment phases $\mathcal{P}_P$ and $\mathcal{P}_A$. The system starts in phase $\mathcal{P}_P$. When the system is in phase $\mathcal{P}_P$ the (offered) wage is the net profit generated by the offered contract. When the system is in phase $\mathcal{P}_A$ the (offered) wage is the disutility generated by the offered contract. As soon as player i deviates from his strategy the system switches immediately to phase $\mathcal{P}_i$. If the system is in phase $\mathcal{P}_A$ the agent accepts the offered contract (and chooses the one-shot best response). Thus, in phase $\mathcal{P}_A$ the MPE is played. If the system is in phase $\mathcal{P}_P$, the new set of default clauses and memory is $(\widetilde{\mathbf{g}}', \mathbf{M}')$, the offered contract is $(\mathbf{g}', m)$ and the state of nature is s, then the agent accepts if and only if

$$m - d(\mathbf{g}', s) > \delta v(\widetilde{\mathbf{g}}', \mathbf{M}'|s), \tag{7.1}$$

where $\delta v(\widetilde{\mathbf{g}}', \mathbf{M}'|s)$ is the expected present value of the of the net surpluses generated by the MPE starting from next period at $(\widetilde{\mathbf{g}}', \mathbf{M}')$ given the current shocks s, and $d(\mathbf{g}', s)$ is the disutility induced by the set of clauses $\mathbf{g}'$ given s.

By construction, the agent has no incentive to deviate. In particular, suppose that the state of nature is s, the principal moves the default and memory to $(\widetilde{\mathbf{g}}', \mathbf{M}')$ and offers $(\mathbf{g}', m)$ so that, as a consequence, the system enters (or stays in) phase $\mathcal{P}_P$.

If $m - d(\mathbf{g}', s) \leq \delta v(\widetilde{\mathbf{g}}', \mathbf{M}'|s)$, the agent is supposed to reject. The expected payoff if the agent conforms is $\delta v(\widetilde{\mathbf{g}}', \mathbf{M}'|s)$. The expected payoff of a one-shot deviation is $m - d(\mathbf{g}', s)$,

[34] Implicitly, we describe strategies as finite automata.

because after the deviation the system enters phase $\mathcal{P}_A$ where the agent gets his maxmin (zero). Therefore rejection is indeed a best response.

If $m - d(\mathbf{g}', s) > \delta v(\widetilde{\mathbf{g}}', \mathbf{M}'|s)$, the agent is supposed to accept and this is obviously a best response.

We now check that the principal has no incentive to deviate in phase $\mathcal{P}_P$. The only way for the principal to make a profitable one-shot deviation is to offer a contract $(\mathbf{g}', m)$ (by way of appropriate amendments and exceptions) satisfying the acceptance condition (7.1). Observe that $v(\widetilde{\mathbf{g}}', \mathbf{M}'|s) \geq \frac{1}{1-\delta}\left[\pi(\mathbf{N}) - d(\mathbf{N}) - \overline{c}_{DE}\right]$ because the parties have always the option to follow the $\mathcal{DE}_k$ rule for every task k such that $c(a_k) < \pi_k - d_k$. Therefore the net payoff the principal can get by "tempting" the agent is bounded above by

$$\max_{\widetilde{\mathbf{g}}', \mathbf{g}'}[\pi(\mathbf{g}', s) - d(\mathbf{g}', s) - \delta v(\widetilde{\mathbf{g}}', \mathbf{M}'|s)] \leq \pi(\mathbf{N}) - d(\mathbf{N}) - \frac{\delta}{1-\delta}\left(\pi(\mathbf{N}) - d(\mathbf{N}) - \overline{c}_{DE}\right).$$

But $\pi(\mathbf{N}) - d(\mathbf{N}) - \frac{\delta}{1-\delta}\left(\pi(\mathbf{N}) - d(\mathbf{N}) - \overline{c}_{DE}\right) \leq 0$ because $\delta \geq \delta^*$. Therefore the principal has no profitable one-shot deviation.

Proof of Lemma 3

We first prove that the set of aggregate incentive constraint (IC') is necessary for incentive compatibility. Suppose that the tuple $(\mathbf{I}, \mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA}, \mathbf{R}, (m_t)_{t\geq 1})$ is part of a SPE. Then it must be the case that the present value of the principal's expected profits is always (weakly) positive and that the agent's expected utility from following $(\mathbf{I}, \mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA}, \mathbf{R}, (m_t)_{t\geq 1})$ is (weakly) larger than what the agent can get by accepting the formal contract offered by the principal, "shirking" on the tasks in $\mathbf{I}$ and rejecting all future offers. To write these incentive constraints in a relatively simple form, define

$$\widehat{m}_{t+1}(h_t) := \sum_{k=1}^{\infty} \delta^{k-1} \mathrm{E}(m_{t+k}|h_t)$$

and let $\mathbf{FB} = \mathbf{I} \cup \mathbf{C} \cup \mathbf{E} \cup \mathbf{DE} \cup \mathbf{DA}$, $\mathbf{FFB} = \mathbf{C} \cup \mathbf{E} \cup \mathbf{DE} \cup \mathbf{DA}$ denote, respectively, the set of tasks governed by first best rules and the set of tasks governed by *formal* first best rules. Then for all $t \geq 1$, all $h_{t+1} = (h_t, s_{t+1})$

$$\frac{\pi(\mathbf{FB})}{1-\delta} + \widehat{\pi}_{t+1}(\mathbf{R}|h_t, s_{t+1}) - \widehat{c}_{t+1}(\mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA}|h_t, s_{t+1}) \geq m_{t+1}(h_t, s_{t+1}) + \delta \widehat{m}_{t+2}(h_t, s_{t+1}), \tag{IC$_P^{t+1}$}$$

and

$$m_t(h_t) + \delta \widehat{m}_{t+1}(h_t) - \frac{d(\mathbf{FB})}{1-\delta} - \widehat{d}_t(\mathbf{R}|h_t) \geq m_t(h_t) - d(\mathbf{FFB}) - d(\mathbf{R}, s_t), \qquad (\mathrm{IC}_A^t)$$

where $\widehat{\pi}_{t+1}(\mathbf{R}|h_{t+1})$ is the expected present value of gross profits for tasks in $\mathbf{R}$ at date $t+1$ conditional on h_{t+1}, $\widehat{d}_t(\mathbf{R}|h_t)$ is the expected present value of disutilities from tasks in $\mathbf{R}$ at date t conditional on h_t, and $d(\mathbf{R}, s_t)$ is the disutility of the tasks regulated by rigid rules at s_t, the last element of h_t. Incentive constraint (IC_A^t) can be written as

$$\widehat{m}_{t+1}(h_t) \geq \frac{1}{1-\delta}\left[d(\mathbf{FFB}) + \frac{d(\mathbf{I})}{\delta}\right] + \widehat{d}_{t+1}(\mathbf{R}|h_t)$$

[note that $\widehat{d}_t(\mathbf{R}|h_t) - d(\mathbf{R}, s_t) = \delta \widehat{d}_{t+1}(\mathbf{R}|h_t)$]. Taking the expected value of both sides of (IC_P^{t+1}) w.r.t. s_{t+1} (conditional on h_t) and combining with the above inequality we obtain

$$\frac{\pi(\mathbf{FB})}{1-\delta} + \widehat{\pi}_{t+1}(\mathbf{R}|h_t) - \widehat{c}_{t+1}(\mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA}|h_t) \geq \widehat{m}_{t+1}(h_t) \geq \frac{1}{1-\delta}\left[d(\mathbf{FFB}) + \frac{d(\mathbf{I})}{\delta}\right] + \widehat{d}_{t+1}(\mathbf{R}|h_t),$$

which yields (IC'):

$$\frac{\delta}{1-\delta}[\pi(\mathbf{FB}) - d(\mathbf{FB})] + \delta[\hat{\pi}_{t+1}(\mathbf{R}|h_t) - \hat{d}_{t+1}(\mathbf{R}|h_t)] - \delta \widehat{c}_{t+1}(\mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA}|h_t) \geq d(\mathbf{I}).$$

An efficient task partition maximizes the present value of net surpluses subject to the constraint of being implementable by a SPE (incentive compatibility). We just proved that (IC') is necessary for incentive compatibility. Therefore we only have to show that a task partition that maximizes the present value of net surpluses subject to (IC') can be implemented by a SPE.

Suppose that $(\mathbf{I}, \mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA}, \mathbf{R})$ solves the maximum problem subject to (IC'). Consider a strategy profile based on four phases, $\mathcal{N}_1$, $\mathcal{N}_2$, $\mathcal{P}_A$, $\mathcal{P}_P$. In the normal phases $\mathcal{N}_j$ each task covered by $(\mathbf{I}, \mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA}, \mathbf{R})$ (which need not be exhaustive) is regulated informally or by the corresponding formal rule and the other tasks are left to the agent's complete discretion. The system starts in phase $\mathcal{N}_1$ and then from the following period moves to phase $\mathcal{N}_2$ if no deviation occurs. In phase $\mathcal{N}_1$ (period 1) the transfer m_1 is such that the participation constraints of principal and agent are satisfied (such m_1 must exist because if the task partition solves the maximum problem it yields a nonnegative net surplus). In phase $\mathcal{N}_2$ and period t ($t \geq 2$) the transfer is such that all the net surplus goes to the agent:

$$m_t(h_t) = \pi(\mathbf{FB}) - d(\mathbf{FB}) + \pi(\mathbf{R}, s_t) - d(\mathbf{R}, s_t) - c_t(\mathbf{E}, \mathbf{DE}, \mathbf{DA}, h_t).$$

In normal phases the agent accepts the proposed contract and chooses the efficient action for all the tasks $k \in \mathbf{I}$. As soon as player i deviates the system switches immediately from the current phase to the punishment phase $\mathcal{P}_i$ which is defined as in the proof of Lemma 1. Hence, there are no incentives to deviate in the punishment phases.

The proof that the principal has no incentive to deviate in a normal phase is essentially the same as in the proof of Lemma 1. To see that the agent has no incentive to cheat in a normal phase simply note that the LHS of (IC') is the expected present value of future net benefits to the agent (which he forgoes if he cheats), whereas the RHS is the temptation to cheat, i.e. the disutility the agent avoids by cheating on informal tasks.

We finally show by contradiction that the agent has no incentive to reject the contract in a normal phase. Suppose he does, then it must be the case that for some task $k \in \mathbf{E} \cup \mathbf{DE} \cup \mathbf{DA}$ the present value of net surpluses in the unfavourable state (i.e., when an enrichment, or exception, or amendment has to be made) is negative. But this implies that it is possible to increase the present value of the net surplus at the outset by regulating task k with a rigid rule. This modification relaxes the incentive constraint (IC'). Therefore the hypothesis that the original task partition solves the maximum problem is contradicted.

Proof of Lemma 2.

It can be shown that under assumption (S) (small writing costs) an efficient task partition must cover all tasks with first best (formal or informal) rules. Then the result follows as a special case of Lemma **3**.

Proof of Proposition 3

Parts (i) and (iii) are obvious. To see that (ii) holds, first note that $d(\mathbf{N}) > \delta\pi(\mathbf{N})$ implies that a fully informal contract violates (IC). Next consider the task partition corresponding to the MPE modified by replacing the formal rule for task 1 with the informal rule. By definition of $\overline{c}_{DE}$ (see the proof of Lemma 1), the expected present value of writing costs conditional on any history is bounded above by $\overline{c}_{DE}/(1-\delta)$ (otherwise it would be possible to decrease costs by using rule $\mathcal{DE}_k$ for some task k, contradicting the Markov equilibrium assumption). If

$d_1 \leq \frac{\delta}{1-\delta}[\pi(\mathbf{N}) - d(\mathbf{N}) - \overline{c}_{DE}]$, then the modified partition satisfies (IC), which means that a fully formal partition cannot be a solution to problem (P). The latter inequality is equivalent to $d_1 < \frac{\delta}{1-\delta}[\pi(\mathbf{N}) - d(\mathbf{N})]$ and $\overline{c}_{DE} \leq \pi(\mathbf{N}) - d(\mathbf{N}) - \frac{1-\delta}{\delta}d_1$.

Proof of Proposition 4

Suppose that a (complete) task partition $(\mathbf{I}, \mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA})$ satisfies (IC) and there are tasks j and k such that $j \in \mathbf{I}$, $k \in \mathbf{C} \cup \mathbf{E} \cup \mathbf{DE} \cup \mathbf{DA}$ and $d_j > d_k$. Since by assumption, j and k have the same writing costs and transition probabilities, swapping the respective rules does not affect costs and relaxes the constraints (IC). Then $(\mathbf{I}, \mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA})$ cannot be the unique solution to (P). This shows that the (unique) solution to (P) must satisfy the property that for some $\overline{d}$ all tasks k with $d_k < \overline{d}$ are informal and all tasks k with $d_k \geq \overline{d}$ are formal.

Proof of Proposition 5

Suppose that tasks differ only with respect to d_k and a task partition $(\mathbf{I}, \mathbf{C}, \mathbf{E}, \mathbf{DE}, \mathbf{DA}, \mathbf{R})$ satisfying (IC') violates the property stated in the proposition. If $d_j < d_k$, but $j \notin \mathbf{FB}$ and $k \in \mathbf{FB}$, then it is possible to increase the net surplus and relax constraint (IC') by swapping the rules for j and k. Therefore the original partition cannot be the solution to (P'). Similarly, if $d_j < d_k$, but $j \notin \mathbf{FB} \cup \mathbf{R}$ and and $k \in \mathbf{FB} \cup \mathbf{R}$, it is possible to increase the net surplus and relax constraint (IC') by swapping the the rules for j and k. Again, this means that the original partition cannot be the solution to (P'). This shows that high surplus (low disutility) tasks are governed by first best rules, intermediate surplus (disutility) tasks are governed by rigid rules, and low surplus (high disutility) tasks are left to the agent's discretion. Relying on the uniqueness of the solution to (P') one can show as in the proof of Proposition 4 that, among the first best tasks, those with lower disutility are governed informally.

References

[1] AL NAJJAR, N., L. ANDERLINI and L. FELLI (2002): "Unforeseen Contingencies," Theoretical Economics Discussion Paper TE/02/431, STICERD, London School of Economics.

[2] ANDERLINI, L. and L. FELLI (1994): "Incomplete Written Contracts: Undescribable States of Nature," *Quarterly Journal of Economics,* **109**, 1085-1124.

[3] ANDERLINI, L. and L. FELLI (1999): "Incomplete Contracts and Complexity Costs," *Theory and Decision,* **46** (1), 23-50.

[4] BAKER, G., R. GIBBONS and K. MURPHY (1994): "Subjective Performance Measures in Optimal Incentive Contracts," *Quarterly Journal of Economics,* **109**, 1125-1156.

[5] BATTIGALLI, P. and G. MAGGI (2002): "Rigidity, Discretion, and the Cost of Writing Contracts," *American Economic Review*, **92**, 798-817.

[6] BULL, C. (1987): "The Existence of Self-Enforcing Implicit Contracts," *Quarterly Journal of Economics*, **102**, 147-159.

[7] DYE, R. A. (1985a): "Costly Contract Contingencies," *International Economic Review*, **26**, 233-50.

[8] DYE, R. A. (1985b): "Optimal Length of Labour Contracts," *International Economic Review*, **26**, 251-270.

[9] FUDENBERG, D. and J. TIROLE (1991): *Game Theory.* Cambridge MA: MIT Press.

[10] GRAY, J. (1976): "Wage Indexation: A Macroeconomic Approach," *Journal of Monetary Economics,* **2**, 221-236.

[11] GRAY, J. (1978): "On Indexation and Contract Length," *Journal of Political Economy*, 86, 1-18.

[12] HART, O. and B. HOLMSTROM (1987): "The Theory of Contracts," in T. Bewley (Ed.) *Advances in Economic Theory, Fifth World Congress.* Cambridge UK, C.U.P.

[13] KRASA, S. and S.R. WILLIAMS (2001): "Incompleteness as a Constraint in Contract Design," mimeo, University of Illinois.

[14] MacLEOD, W. B. (2000): "Complexity and Contract," *Revue d'Economie Industrielle*, **92**, 149-178.

[15] MacLEOD, W. B. and J.M. MALCOMSON (1989): "Implicit Contracts, Incentive Compatibility, and Involuntary Unemployment," *Econometrica,* **57**, 447-480.

[16] MASKIN, E. and J. TIROLE (1999): "Unforeseen Contingencies and Incomplete Contracts," *Review of Economic Studies*, **66**, 83-114.

[17] MEIHUIZEN, H.E. and S.N.WIGGINS (2000), "Information Cascades and Contractual Incompleteness in Natural Gas Contracting," mimeo.

[18] PEARCE, D. and E. STACCHETTI (1998): "The Interaction of Implicit and Explicit Contracts in Repeated Agency," *Games and Economic Behavior,* **23**, 75-96.

[19] WILLIAMSON, O. (1985): *The Economic Institutions of Capitalism.* New York: Free Press.