Organizational Design: Culture & Incentives*

Willemien Kets†

April 27, 2021

Abstract

Why do some organizations perform so much better than others, and often persistently so? While the economics literature traditionally emphasizes the importance of incentives, other social sciences focus on the role of organizational culture. Yet very little is known about how an organization's incentive structure and its culture jointly shapes its performance. This paper therefore develops a formal model of how an organization's incentives and culture interact. The model offers a unified explanation for a variety of empirical phenomena, including why an organization's culture can be difficult to change and why there can be persistent performance differences across companies even when they have access to the same technology and resources.

*I am grateful to Colin Camerer, Alessandra Casella, David Cooper, Josse Delfgaauw, Robert Dur, Dan Friedman, Bob Gibbons, John Hannan, Elliot Lipnowski, Massimo Morelli, and numerous seminar audiences for helpful comments and stimulating discussions.

†Department of Economics, University of Oxford. E-mail: willemien.kets@economics.ox.ac.uk.

1 Introduction

A longstanding question in economics is why organizations that are fundamentally similar in all respects differ so much in their performance.[1] However, our understanding of the central determinants of organizational performance is still limited. While economics has a well-developed theory of how organizational features that are directly payoff-relevant – e.g., incentive systems, allocation of decision rights, information flows –,[2] there is mounting empirical evidence from across the social sciences that an organization's performance also depends on its culture. Despite this growing empirical literature, a unified theoretical framework is still lacking. This makes it difficult to address key questions such as: Why so some incentive structures work well for some organizational cultures but not others? How to design incentives when culture matters? Why is it so difficult to improve performance by changing the organization's culture?

To answer these questions, this paper develops a simple behavioral model of economic incentives and organizational culture. The central building block is the finding from psychology that an organization's culture shapes people's beliefs about what others might do (DiMaggio, 1997). This simple model is able to provide a unified explanation for a variety of disparate evidence, including why an organization's culture can be difficult to change, why mergers and acquisitions often fail to live up to expectations, and why there can be persistent performance differences across companies even when they have access to the same technology and resources. As a side product, it delivers new tools that bridge the work on organizations in economics and other disciplines, including management, strategy, and organizational theory.

The model focuses on a central problem in organizational design: How to incentivize agents to work when they can free-ride on each others' efforts. As in the seminal model of Holmstrom (1982), the principal wants to incentivize agents work towards a common project. Because individual effort levels are not observable, payments can depend only on the project outcome. Working increases the likelihood that the project succeeds but is costly. Agents are protected by limited liability. That is, the principal cannot penalize the agents: all payments to the agents have to be non-negative. Thus, it is without loss of generality to restrict attention to the case where the agents are paid a bonus $b \geq 0$ when the project succeeds and get 0 otherwise.

The principal's problem is to induce agents to work at the lowest possible cost. To illustrate, suppose that working (choosing H) costs $c > 0$ and that the probability that the project is a success when m agents shirk (choose L) is ν^m, where $\nu \in (0, 1)$. Then, if there are two agents, payoffs are given by:

[1] See Gibbons and Henderson (2013) for an excellent discussion and literature review.

[2] See Gib (2013) for an overview.

	H	L
H	$b-c, b-c$	$\nu\, b-c, \nu\, b$
L	$\nu\, b, \nu\, b-c$	$\nu^2 b, \nu^2 b$

There are obvious externalities: If an agent works, he increases the payoff to the other agent.[3] Of course, agents do not internalize these externalities so there is scope for free-riding. But this is not the only problem that the principal faces; there are also strategic complementarities. That is, an agent has a greater incentive to work if he other agent works.[4] This means there is scope for *coordination failure*: Depending on how the bonus is set, there can be two pure Nash equilibria: one in which both agents work, and one in which both agents shirk. So, even if it is incentive compatible for both agents to work (i.e., (H, H) is a Nash equilibrium), the agents may shirk if they are unsure whether the other will work. In the words of Palfrey (1990), "incentive compatibility is only half of the implementation problem. The other half is the multiple equilibrium problem."

The economics literature has used two opposing approaches to the multiple equilibrium problem: The partial implementation literature assumes that agents coordinate on the principal-preferred Nash equilibrium whereas the full implementation literature assumes that agents coordinate on the principal's least-preferred equilibrium. Thus, under partial implementation, the principal choose the minimum bonus $b_{partial}$ such that all agents work in *some* Nash equilibrium; and under full implementation, he selects the minimum bonus $b_{full} \geq b_{partial}$ such that all agents work in *all* Nash equilibria (Halac et al., 2021; Winter, 2004). But neither of these approaches can explain why some organizations seem to select better equilibria than others, and which interventions, if any, can lead them to coordinate on better outcomes (Gibbons and Henderson, 2013).

Building on my earlier work (Kets et al., 2019; Kets and Sandroni, 2019, 2021), I therefore explicitly model how an organization's culture affects behavior. The starting point is the observation is that many organizations rely on broad principles to guide workers' actions (Akerlof and Kranton, 2005; Akerlof et al., 2020; Camerer and Vepsalainen, 1988; Gibbons et al., 2021; Kreps, 1990). These cultural rules are generally left implicit and are often communicated through stories or slogans. For example, the organization may tell stories about how a difficult project was brought to a successful conclusion through exceptional teamwork. This is of course not how an organization would operate if its employees were all like the proverbial homo economicus. In such an idealized world, the organization would be able to costlessly communicate what it would

[3]For example, if agent j works, then, by working, agent $i \neq j$ increases j's payoff from $\nu\, b - c$ to $b - c$.

[4]The incentive to work for agent j (for a given action of agent $i \neq j$) is the difference in payoff from working and shirking. If agent i works, the incentive for j to work is $b - c - \nu\, b = (1-\nu)\, b - c$ whereas if i shirks, j's incentive to work is only ν, $b - c - \nu^2 b = \nu(1-\nu)b - c < (1-\nu)\, b - c$.

expect its employees to do in every possible contingency. But in real life, there are limits to what can be communicated. Telling stories can then be an effective way to communicate to employees what is expected of them. While this is obviously a noisy way to communicate, it has the advantage that stories often have a visceral impact, i.e., they directly influence people's instinctive responses (DiMaggio, 1997). For example, employees who have been exposed to stories about how teams operate within their organization may be inclined to approach teamwork in a similar way. Of course, talk is cheap: Telling stories, even highly effective ones, will not be sufficient to induce agents to work. However, an effective organizational culture can reduce incentive cost by helping to coordinate expectations.

This simple approach can shed light on why some organizations perform so much better than others. Because the organization's culture influences the reasoning process, some organizations are able to incentivize agents at a lower cost than others. An organization whose culture makes it very costly to induce agents to work may opt not to implement the first best (i.e., all work). Such an organization would thus suffer from low performance relative to an organization whose culture makes it possible to implement the first best at low cost.

The model can also shed light on how an organization's culture influences its incentive costs (and thus its ability to implement the first best). For example, the model predicts that it is costlier to incentivize agents if they do not identify with the organization's mission. That can help better understand why incentive payments are lower in some sectors. The model also predicts that having a weak culture can be optimal if the culture is not very effective. This is consistent with evidence from the management literature (Kotter and Heskett, 1992) but cannot be obtained with other models (see Section 4 for a literature review). The model can also helps better understand why culture change is so difficult: The model predicts that changing an organization's culture generally tends to increase incentive costs, even if adjusting incentives is costless. This can help explain why organizations whose culture is no longer effective may wish to start fresh at a greenfield site (Brynjolfsson and Milgrom, 2013).

The remainder of this paper is organized as follows. Section 2 introduces the model and Section 3 presents the main results. Section 4 concludes by discussing the related literature.

2 Model

2.1 Moral hazard

The baseline model builds on the classic model of moral hazard in teams of Holmstrom (1982). The principal offers a contract to a set $N = \{1, \ldots, n\}$, $n \geq 2$, of agents to incentivize them to work towards a common project. Each agent $j \in N$ can either work (denoted $e_j = H$) or

shirks ($e_j = L$). The cost of working is $c > 0$. Depending on whether agents work or fail, the project succeeds or fails: If m agents shirk, the probability that the project is a success is ν^m for $\nu \in (0, 1)$.

Individual effort is not verifiable. Hence, only the project outcome (success or failure) is contractible. Agents are protected by limited liability: Any payment from the principal to an agent is required to be nonnegative. Hence, it is without loss of generality to restrict to success-contingent bonuses: Each agent receives a bonus $b \geq 0$ if the project succeeds and 0 otherwise.[5] This induces an n-player game that we denote by $G_n(\nu, b, c)$

2.2 Equilibrium

I assume that agents form a belief about what others might do by taking their perspective, using their own experience as a guide. In psychology, the (cognitive) ability to take another person's perspective is termed *theory of mind*.[6] It has two phases. To form a belief about how other agents might respond to a situation, an agent first observes his own instinctive reaction and projects it onto others. This is a rapid and instinctive process referred to as first-person simulation (Goldman, 2006). It is followed by a slower, more deliberative process whereby the agent reasons about the other agents based on a naive understanding of psychology (Gopnik and Wellman, 1994). This may lead him to adjust his initial belief.

Following Kets and Sandroni (2021), I model this as follows. Each agent $j \in N$ has an *impulse* to work (denoted $I_j = H$) or to shirk ($I_j = L$). Impulses are drawn from a distribution $\mu((I_j)_j)$ that depends on the organization's culture, as discussed in Section 2.3 below. Impulses are privately observed and do not directly affect payoffs. Agents' instinctive reaction is to follow their initial impulse. That is, if an agent's impulse is $I_j = H$, then his pre-reflective inclination is to work, and if $I_j = L$, he is inclined to shirk. This defines the agent's level-0 strategy σ_j^0 (i.e., $\sigma_j^0(I_j) = I_j$). Since agents are introspective, each agent realizes that other agents likewise have an impulse. So, each agent j forms a posterior belief $\mu(I_{-j} \mid I_j)$ about the other agents' impulses. This allows the agent to formulate a best response to the other agents' level-0 strategy.[7] This defines his level-1 strategy σ^1. The reasoning does not stop here: For any level $k > 0$, agents' level-k strategy σ_j^k is a best-response to the level-$(k-1)$ strategy σ_{-j}^{k-1} of the other agents. Agents continue to reason in this way until they no longer wish to revise their choice. Accordingly, the

[5] The assumption that agents receive identical bonuses follows Winter (2004). Halac et al. (2021) consider the case where agents can receive different bonuses.

[6] Thus, theory of mind needs to be distinguished from empathy which is the emotional ability to take another person's perspective (Singer and Fehr, 2005).

[7] If there are multiple best responses, an action is chosen using a fixed tie-breaking rule. The choice of tie-breaking rule does not affect results.

behavior of agent $j \in N$ is described by the limit $\sigma_j := \lim_{k\to\infty} \sigma_j^k$ of the reasoning process (if it exists). The profile $\sigma = (\sigma_j)_j$ of limiting strategies is an *introspective equilibrium*.

Introspective equilibrium can be related to one of the classic solution concepts in game theory:

Proposition 1. (Kets and Sandroni, 2021) *Any introspective equilibrium is a correlated equilibrium.*

By the epistemic characterization of correlated equilibrium by Aumann (1987), Proposition 1 implies that agents' behavior is consistent with common knowledge of rationality. So, while agents choose their action by introspection rather than a deductive equilibrium analysis, they behave as if they are rational, believe that others are rational, believe that others believe that others are rational, and so on.[8]

2.3 Organizational culture

An organization's culture induces *common-*(p, q) *belief*. That is, with probability $p \in (0, 1)$, working is culturally salient in the sense that with probability $q \in (\frac{1}{2}, 1)$, each agent has an impulse to work (independently across agents). Thus, with probability p, in expectation, most agents have an impulse to work and think it is likely that most have an impulse to work and think it is likely that most... Likewise, with probability $1-p$, shirking is culturally salient in the sense that with probability q each agent has an impulse to shirk (again, independently across agents). So, if shirking is culturally salient, in expectation, most agents have an impulse to shirk and think it is likely that most have an impulse to shirk, etc.

The parameter $p \in (0, 1)$ can be viewed as a measure of how *effective* an organization is at aligning agents' impulses with its goals. If p is close to 1, then, ex ante, we expect most agents to have an impulse to work, while if p is close to 0, we expect most agents to have an impulse to shirk (ex ante). The parameter $q \in (\frac{1}{2}, 1)$ measures the *consistency* of an organization's culture, that is, the extent to which agents agree on what the cultural rule prescribes. If q is close to 1 then the organization's culture is highly consistent in the sense that it is highly likely that agents share the same impulse. If q is close to $\frac{1}{2}$, on the other hand, the culture is not very consistent in that, regardless of whether working is culturally salient, about half of the agents work (in expectation) while the other half shirks.

[8] A formal proof can be found in Kets and Sandroni (2021). The intuition is straightforward: At level 1, agents play a best response to their belief and hence cannot choose an action that is strictly dominated. By induction, at level k, agents choose a best response that other agents choose actions that survive k rounds of the iterated elimination of strictly dominated strategies. So, in introspective equilibrium, agents choose actions that survive the iterated elimination of strictly dominated strategies. Since agents' beliefs are consistent with a common prior (the impulse distribution), we have a correlated equilibrium.

An *organization* can then be viewed as a combination of a game $G_n(\nu, b, c)$ that specifies the strategic environment of the team together with its culture (p, q). We denote an organization by $\mathcal{O} = (G_n(\nu, b, c); p, q)$.

The following preliminary result shows that an introspective equilibrium exists and is unique except in knife-edge cases.[9]

Proposition 2. *Any organization has an introspective equilibrium and it is generically unique (i.e., it is unique for a set of organizations of Lebesgue measure 1).*

While Proposition 2 shows that organizations generically have a unique introspective equilibrium, an organization's introspective equilibrium may depend on its culture. That is, two organizations $\mathcal{O} = (G; p, q)$, $\mathcal{O}' = (G; p', q')$ that face the same economic environment G but have a different culture (i.e., $(p, q) \neq (p', q')$) may have different introspective equilibria. This captures the idea that behavior is shaped by both the economic environment and the organization's culture.

2.4 Organizational design

To gain insight into how an organization's culture affects performance, suppose that the first best outcome is that all agents work. That is, in the absence of incentive problems (individual efforts are observable), the organization's profit is maximized when all agents work (i.e., the profit $\Pi(m)$ for the organization when m agents work satisfies $\Pi(m) - \Pi(m-1) > c$ for all $m \leq n$). Then, given an organization $\mathcal{O}$, say that a bonus b *induces agents to work* if all agents work in every introspective equilibrium of $\mathcal{O}$. A bonus b^* is *optimal* if it minimizes the incentive costs (total expected bonus payment) among all bonuses that induce agents to work.[10] Notice that by Proposition 2 this problem is well-defined.

Characterizing the optimal bonus b^* can help better understand persistence differences across organizations. If an organization $\mathcal{O}$ has high incentive costs (i.e., b^* is high), then it may be prohibitively costly to implement the first best; such an organization may choose a bonus that induces only some agents to work. As a result, its performance will suffer relative to an organization $\mathcal{O}'$ whose incentive costs are lower and that chooses to implement the first best. A possibility that is of particular interest is when the organizations face the same economic environment but their incentive costs differ due to differences in their culture (i.e., $\mathcal{O} = (G; p, q)$,

[9] Kets and Sandroni (2021) prove a similar result for a continuum of agents.

[10] As in Winter (2004, footnore 3), the optimal bonus itself need not induce agents to work. Formally, b^* is the optimal bonus if (1) there is no bonus $b < b^*$ that induces agents to work; and (2) for any $\varepsilon > 0$, $b^* + \varepsilon$ induces agents to work.

$\mathcal{O}' = (G; p', q')$, $(p, q) \neq (p', q')$). The next section studies how an organization's culture affects its incentive costs in this case.

3 Organizational Culture and Incentives

This section studies how the effectiveness and consistency of an organization's culture affect its incentive costs. The first result shows that *effective cultures economize on incentive costs*. To state the result, denote the bonuses under partial and full implementation by $b_{partial}$ and b_{full}, respectively. Throughout the remainder of the paper, effort cost $c > 0$ and technology parameter $\nu \in (0, 1)$ are fixed so as to focus on how the optimal bonus varies with organizational culture (i.e., $b^* = b^*(p, q)$).

Proposition 3. [Effectiveness] *More effective cultures have lower bonuses: As the effectiveness p increases, the optimal bonus b^* falls. Moreover, $b^* \in (b_{partial}, b_{full})$.*

Proposition 3 shows that an organization's culture helps solve both the free-rider and the multiple equilibrium problem, and at a lower cost than under full implementation (i.e., $b^* < b$). Economic incentives and organizational culture are substitutes: More effective cultures have lower incentive costs.[11]

The intuition behind Proposition 3 is straightforward: If the culture is effective (i.e., p high), agents think it is likely that the others have an impulse to work. Because the game has strategic complementarities, this means that agents have a strong incentive to work at level 1 even if the bonus is not very high. So, the bonus that induces agents to work at level 1 is small. A similar argument applies at higher levels. Hence, if the culture is effective, the bonus that induces agents to work can be small.

An intuitive implication of Proposition 3 is that it is easier to incentivize agents if their impulses are more aligned with the organization. Thus, the model predicts that incentives costs will be lower in sectors where employees identify with the organization's mission. Conversely, if there is a change that causes impulses to be misaligned with with organization, incentive costs may increase.

The next result considers which level of consistency minimizes incentive costs. For given ν, b, n and p, the *optimal level of consistency* $q^*(p)$ minimizes b^*.

Proposition 4. [Consistency] *Assume $n = 2$. The optimal level of consistency $q^*(p)$ increases with the effectiveness p of a culture, where the increase is strict whenever $p > \frac{1}{2}$.*

[11] This need not be true in the long run, when the organization's culture and its incentive systems co-evolve (e.g., Bisin and Verdier, 2017).

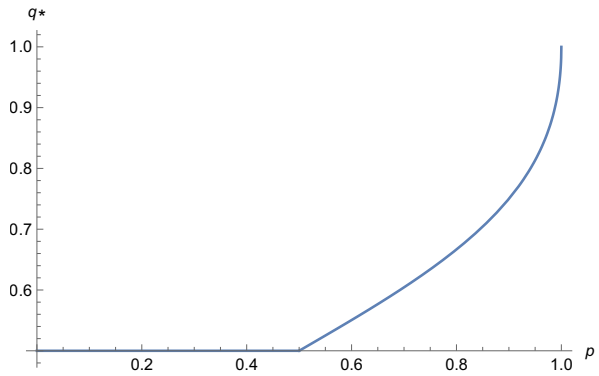

Figure 1: The optimal level of consistency $q^*(p)$ as a function of the culture's effectiveness p (for $\nu = 0.4$ and $n = 2$).

Proposition 4 is illustrated in Figure 1. The result shows that while having an effective culture is always beneficial (Proposition 3, consistency is more like a double-edged sword: it lowers incentive costs if the culture is effective (i.e., p close to 1) but raises them otherwise. This result is consistent with the management literature (Kotter and Heskett, 1992). This literature shows that in times of change when impulses are not aligned with the organization's goals (i.e., p small), it is optimal to have a weak culture. An important implication of Proposition 4 is that how a team responds to incentives may depend on the team's characteristics: A bonus that is sufficient to improve the performance of a team with a strong effective culture may not have a positive effect on a team with a weaker or less effective culture, in line with evidence (Delfgaauw et al., 2021).

The intuition behind Proposition 4 is that an inconsistent culture disrupts self-enforcing expectations. If the culture is highly consistent (q close to 1), then agents think it likely that others have the same impulse; but if the culture is highly inconsistent (q close to $\frac{1}{2}$), then agents' impulses are not very informative of other agents' impulses. Whether a consistent culture raises or lower incentive costs thus depends on whether the self-enforcing expectations favor the agents working or shirking. If the organization's culture is highly effective (i.e., p close to 1), it is likely that working is culturally salient; hence, expectations favor the agents working and having a consistent culture (i.e., q close to 1) is beneficial. If the culture is ineffective, on the other hand, expectations favor shirking and having an inconsistent culture (i.e., q close to $\frac{1}{2}$) minimizes incentive costs.

While Proposition 4 is intuitive (and in line with empirical evidence), it is difficult to obtain with other models. For example, if an organization's culture influences agents' preferences (Akerlof and Kranton, 2005, 2008; Akerlof et al., 2020), it is difficult to see why an inconsistent culture may be beneficial. The result is also difficult to obtain with a model that views an organization's culture purely as an equilibrium selection device (Kreps, 1990), as this approach is silent on how a culture's consistency matters for performance.

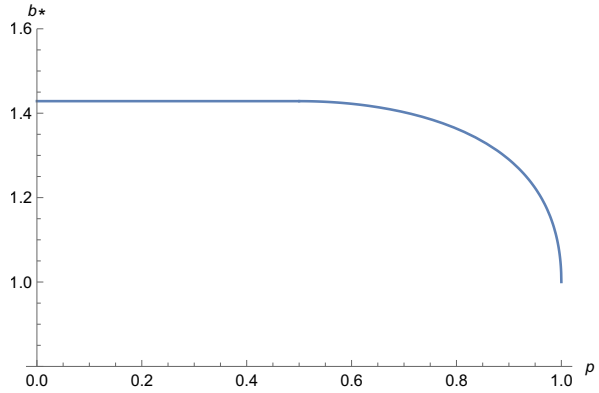

Figure 2: The optimal bonus $b^* = b^*(p, q)$ as a function of effectiveness p when the level of consistency is chosen optimally, i.e., $q = q^*(p)$ (for $\nu = 0.4$, $c = 1$, and $n = 2$).

An immediate corollary of Propositions 3–4 is that it is optimal for organizations to have a culture that is both highly effective and highly consistent (i.e., p, q close to 1):

Corollary 1. [Ideal Culture] *Assume $n = 2$. An organization's incentive costs are minimized when its culture is highly effective and consistent, i.e., $\lim_{p\to 1} b^*(p, q^*(p)) \leq b^*(p', q^*(p'))$ for any $p' \in (0, 1)$.*

The result follows directly from the proofs of Proposition 3–4 (and the envelope theorem) and is thus omitted. Corollary 1 is illustrated in Figure 2: If the level of consistency is chosen optimally to match the culture's effectiveness (i.e., $q = q^*(p)$), then the optimal bonus decreases with the culture's effectiveness. The next section considers in the context why organizations may not be able to realize this ideal culture.

3.1 Culture change

This section studies under which conditions an organization can benefit from changing its culture, in particular improving its effectiveness. Consider an organization whose culture is perhaps not very effective (i.e., $p < 1$) but whose consistency is well-matched with its effectiveness (i.e., $q = q^*(p)$). Then, the organization benefits from changing its culture to (p', q') if this reduces its incentive costs, i.e., $b^*(p', q') < b^*(p, q^*(p))$. The following result shows that increasing the effectiveness of the culture (i.e., $p' > p$) reduces incentive costs only under fairly stringent conditions.

Proposition 5. [Culture Change] *Assume $n = 2$ and let $p \in (0, 1)$. Then, changing the organization's culture from $(p, q^*(p))$ to (p', q') reduces incentive costs only if (i) its initial effectiveness is low ($p \leq \frac{1}{2}$) and its consistency strictly increases ($q' > q^*(p)$); or (ii) its initial level of effectiveness is not too low ($p > \frac{1}{2}$), the new level of effectiveness p' sufficiently large, and its new level of consistency q' is sufficiently close to its initial level $q^*(p)$. That is, for $p < \frac{1}{2}$,*

$b^*(p', q') > b^*(p, q^*(p))$ *only if* $p' > \frac{1}{2}$ *and* $q' > q^*(p)$*; and for* $p > \frac{1}{2}$*,* $b^*(p', q') > b^*(p, q^*(p))$ *only if* $q' \in (\underline{q}(p, p'), \overline{q}(p, p'))$.

The proof of Proposition 5 completely characterizes the bounds $\underline{q}(p, p'), \overline{q}(p, p')$. Proposition 5 implies that increasing the effectiveness of a culture may increase incentive costs. A particularly salient case is when the consistency of an organization's culture falls following a culture change. This could be, for instance, when the organization starts telling new stories about itself that conflict with the old ones; in such a case, employees' impulses are less likely to be aligned with each other. Proposition 5 says that in such a case, the culture change lowers incentive costs only if the new culture is sufficiently more effective (i.e., $p' - p$ sufficiently large) or the reduction in consistency is not too severe. So, except under fairly restrictive conditions, increasing the effectiveness of an organization's culture may actually increase incentive costs rather than reduce them. This means that it may be prohibitively costly for the organization to implement the first best following the culture change; and as a result, performance may decline.

An important implication of Proposition 5 is that if an organization's culture is no longer effective – e.g., because it wants to enter a new line of business –, then it may be best the start with a blank slate (at a so-called green-field site) if that allows it to create an effective and consistent culture, even if that means not fully utilizing human and physical capital (Brynjolfsson and Milgrom, 2013).

This result also suggests that changing an organization's culture can be costly even if it is costless to adjust incentives. So, the potential costs of culture change identified in Proposition 5 go above and beyond the costs of adjusting incentive systems as part of the change effort emphasized by organizational scholars (Kaplan and Henderson, 2005). This means that even if it is costless for an organization to adjust its incentives following a culture change, culture change may not improve organizational performance.

4 Related literature

This paper contributes to the large and growing literature on organizational culture in economics (Akerlof and Kranton, 2005, 2008, 2010; Akerlof et al., 2020; Besley and Persson, 2017; Carrillo and Gromb, 1999, 2007; Dessein and Prat, 2019; **?**; Gibbons et al., 2021; Gibbons and Henderson, 2013; Guiso et al., 2015; Hermalin, 2001; Lazear, 1999a,b; Van den Steen, 2010; Weber and Camerer, 2003). Within this literature, it is most closely related to the work of Kreps (1990) in that it assumes, like Kreps (1990), assumes that an organization's culture helps coordinate expectations. In contrast to Kreps (1990), an organization's culture is more than merely an equilibrium selection device. This paper also contributes to the literature on behavioral mech-

anism design (Koszegi, 2014). Unlike much of the existing literature on behavioral mechanism design, this paper focuses on how non-economic factors influences players' beliefs (rather than focusing on nonstandard preferences or bounded rationality). Finally, this paper contributes to the literature on coordination and organizations (e.g., Crémer, 1993; Crémer et al., 2007; Dessein and Santos, 2006). However, this literature does not focus on the question of design; moreover, it assumes that agents select the organization's preferred equilibrium.

Appendix A Proofs

A.1 Proof of Proposition 2

I prove the result for a more general class of games, so as to highlight the key drivers behind the result. The probability that the project is a success when $m \leq n$ of agents work (and the other $n-m$ agents shirk) is $P(m)$, where $P : \{0, \ldots, n\} \to [0,1]$ is strictly increasing and has strictly increasing differences. That is,

1. $P(m') > P(m)$ whenever $m' > m$;
2. if $m < m' < n$, $P(m'+1) - P(m') > P(m+1) - P(m)$.

The model in Section 2.1 is the special case where $P(m) = \nu^{n-m}$. Denote by $\mathbb{E}^k_{p,q}[P(m)]$ the expectation of P if agents follow their level-k strategies; other (conditional) expectations can be defined similarly. For concreteness, assume that ties are broken in favor of H: If an agent is indifferent between H and L at some level k, then he chooses H.

The following preliminary result will be useful:

Lemma 1. *For any culture* (p,q),

$$\mathbb{E}^0_{p,q}[P(m+1) - P(m) \mid I_j = H, m \leq n-1] \geq \mathbb{E}^0_{p,q}[P(m+1) - P(m) \mid I_j = L, m \leq n-1]$$

Proof. Since P has strictly increasing differences, the result follows if the distribution over the impulses of agents $j' \neq j$ given $I_j = H$ first-order stochastically dominates the distribution over the impulses of agents $j' \neq j$ given $I_j = L$. Let $\theta = s$ denote the event that action $s \in \{H, L\}$ is culturally salient. Then, if agent has impulse $I_j = s$, he assigns posterior probability $\mathbb{P}(\theta = H \mid I_j = s)$ to the impulses of the other agents $j' \neq j$ having a binomial distribution with parameters $(n-1, q)$ (where an agent j' having impulse $I_{j'}$ is a "success"); and he assigns posterior probability $\mathbb{P}(\theta = L \mid I_j = s)$ to the impulses of the other agents $j' \neq j$ having a

binomial distribution with parameters $(n-1, 1-q)$. We have

$$\mathbb{P}(\theta = H \mid I_j = H) = \frac{p\,q}{p\,q + (1-p)\,(1-q)} =: \pi_H;$$
$$\mathbb{P}(\theta = H \mid I_j = L) = \frac{p\,(1-q)}{p\,(1-q) + (1-p)\,q} =: \pi_L.$$

It is easy to check that $\pi_H > \pi_L$ and that the binomial distribution with parameters $(n-1, q)$ first-order stochastically dominates the binomial distribution with parameters $(n-1, 1-q)$ (as $q > \frac{1}{2}$). The result then follows from a standard coupling argument. □

At level 0, all agents follow their impulse. At level 1, action H is a best response for agent j with impulse $I_j = s$ if and only if

$$\mathbb{E}^0_{p,q}[P(m) \mid I_j = s, j \text{ chooses } H]\, b - c \geq \mathbb{E}[P(m) \mid I_j = s, j \text{ chooses } L]\, b.$$

By assumption, $P(m+1) - P(m) > 0$ for all $m \leq n-1$, so this is equivalent to

$$b \geq \frac{c}{\mathbb{E}^0_{p,q}\big[P(m+1) - P(m) \mid I_j = s, m \leq n-1\big]} =: B_{p,q}(s).$$

By Lemma 1, there are three cases. First, if $b \geq B_{p,q}(L)$, then choosing H is a best response for agent j at level 1 regardless of his impulse. Second, if $b \leq B_{p,q}(H)$, then choosing L is a best response for j at level 1 regardless of his impulse. Third, if $b \in (B_{p,q}(H), B_{p,q}(L))$, then it is a strict best response for an agent to follow his impulse, i.e., to choose H if his impulse is $I_j = s$ and to choose L if his impulse is $I_j = L$. Note that best responses are unique unless $b = B_{p,q}(L)$ or $b = B_{p,q}(H)$. So, given our tie-breaking assumption, at level 1, agents choose H (regardless of their impulse) whenever $b \geq B_{p,q}(L)$; they choose L if $b < B_{p,q}(H)$; and they follow their impulse otherwise.

I next show that the level-2 strategies are identical to the level-1 strategies. It then follows from a simple inductive argument that $\sigma^k = \sigma^1$ for all $k \geq 1$; hence, an introspective equilibrium exists and is given by $\sigma = \sigma^1$. To see this, first suppose that agents follow their impulse at level 1, i.e., $b \in [B_{p,q}(H), B_{p,q}(L))$. By construction, this is a best response to the belief that agents follow their impulse (at level 0). So, at level 2, it is a best response for agents to follow their impulse. Next suppose that agents choose H at level 1 (regardless of their impulse), i.e., $b \geq B_{p,q}(L)$. Again, by construction, this is a best response to the belief that others follow their impulse. Choosing H is a best response at level 2 for an agent with impulse $I_j = s$ if and only if $\mathbb{E}^1_{p,q}[P(m+1) \mid I_j = s, m \leq n-1]\, b - c \geq \mathbb{E}[P(m) \mid I_j = s, m \leq n-1]\, b$. Given the level-1 strategies, this reduces to $(P(n) - P(n-1))\, b \geq c$. The result then follows by noting that

$P(n) - P(n-1) \geq \mathbb{E}^0_{p,q}[P(m+1) - P(m) \mid I_j = s, m \leq n-1]$. The proof for the case where $b < B_{p,q}(H)$ is analogous and is therefore omitted.

To see that the introspective equilibrium is generically unique, note that $b \neq B_{p,q}(s)$ for $s \in \{H, L\}$ for all but a measure-0 set of payoff parameters. □

A.2 Proof of Proposition 3

We first prove the first claim. Let $P : \{0, \ldots, n\} \to [0, 1]$ be as defined in the proof of Proposition 2. Fix a culture (p, q). By Lemma 1, the minimum bonus that induces all agents to work at level 1 is

$$b^1 = \frac{c}{\mathbb{E}^0_{p,q}\big[P(m+1) - P(m) \mid I_j = L, m \leq n-1\big]} = B_{p,q}(L).$$

By the proof of Proposition 2, the bonus that induces all agents to work at level $k \geq 1$ is b^1. Hence, for any given culture (p, q), we have $b^* = B_{p,q}(L)$. Recall that, for an agent with impulse $I_j = L$, the impulses of other agents are distributed according to a mixture of two binomial distributions, one with parameters $(n-1, q)$ and one with parameters $(n-1, 1-q)$, with the former having weight

$$\pi_L = \frac{p\,(1-q)}{p\,(1-q) + (1-p)\,q}.$$

As π_L increases in p, the conditional distribution of the impulses of the other $n-1$ agents given $I_j = L$ increases with p (in the sense of first-order stochastic dominance). Hence, as P has increasing differences, $\mathbb{E}^0_{p,q}\big[P(m+1) - P(m) \mid I_j = L, m \leq n-1\big]$ increases with p, and the result follows.

We next relate b^* to the bonuses b_{full} and $b_{partial}$. It is easy to check that

$$b_{partial} = \frac{c}{P(n) - P(n-1)}; \qquad b_{full} = \frac{c}{P(1) - P(0)};$$

since P has strictly increasing differences, we thus have $b_{full} > b_{partial}$. The result then follows from the fact that P has strictly increasing differences together with the observation that the binomial distribution with parameters $(n-1, x)$ for some $x \in (0, 1)$ has full support on $\{0, \ldots, n-1\}$. □

A.3 Proof of Proposition 4

Define $Z(q, p, \nu, n) := \mathbb{E}_{p,q}[P(m+1) - P(m) \mid m \leq n-1, I_j = L]$; note that $Z(p, q, \nu, n) > 0$ for all organizations $\mathcal{O} = (G; p, q)$. Then, by the Proof of Proposition 2, the optimal bonus $b^* = b^*(p, q)$ when the organization's culture is (p, q) equals $b^* = c/Z(q, p, \nu, n)$. So, to identify

the optimal level of consistency $q^*(p)$ for given p, it suffices to maximize $Z(q, p, \nu, n)$. By the Binomial theorem,

$$Z(q, p, \nu, n) = \pi_L(q + \nu(1-q))^{n-1} + (1-\pi_L)(1-q+\nu\, q)^{n-1}.$$

Let $n = 2$ and fix $p \in (0,1)$ and $\nu \in (0,1)$. By the first- and second-order conditions, Z is maximized at $q_{max} = \frac{1}{2p-1}\big(p - \sqrt{p(1-p)}\big) \in (0,1)$. It is easy to check that q_{max} is increasing in p. Moreover, $q_{max} > \frac{1}{2}$ if and only if $p > \frac{1}{2}$. So, for $p > \frac{1}{2}$, $q^* = q_{max}$. For $p \leq \frac{1}{2}$, we get the corner solution $q^* = \frac{1}{2}$. □

A.4 Proof of Proposition 5

First consider the case $p \leq \frac{1}{2}$. Then, by the proof of Proposition 4, $q^*(p) = \frac{1}{2}$ and $Z(p, q, \nu, n) = \frac{1}{2} + \frac{1}{2}\nu$, independent of p. Clearly, $b^*(p', q') \leq b^*(p, q^*(p))$ whenever $p' \leq \frac{1}{2}$. So suppose $p' > \frac{1}{2}$. Then, it is easy to check that $b^*(p', q') > b^*(p, q^*(p))$ if and only if $q' > q^*(p)$.

Next suppose $p > \frac{1}{2}$ and let $p' \in (p, 1)$. The conditions on q' such that $b^*(p', q') > b^*(p, q^*(p))$ can be characterized using the following quartic functions (i.e., fourth-degree polynomial):

$$Q(\tilde{q}; p, p') := (p')^2 - 4(p')^2\tilde{q} + (-4p + 4p^2 + 2p' + 4(p')^2)\tilde{q}^2 + \ldots \\ (8p - 8p^2 - 4p')\tilde{q}^3 + (2p-1)^2\tilde{q}^4.$$

Using standard results on quartic functions, it is straightforward to show that all four roots of $Q(\cdot; \tilde{q})$ are real and distinct. Let $\underline{q}(p, p'), \overline{q}(p, p')$ be its smallest and second-smallest roots, respectively. Then, it can be checked that $q^*(p) \in (\underline{q}(p, p'), \overline{q}(p, p'))$; and, after some algebraic manipulation, $b^*(p', q') > b^*(p, q^*(p))$ only if $q' \in (\underline{q}(p, p'), \overline{q}(p, p')$. □

References

(2013). The handbook of organizational economics.

Akerlof, G. A. and R. E. Kranton (2005). Identity and the economics of organizations. *Journal of Economic Perspectives 19*, 9–32.

Akerlof, G. A. and R. E. Kranton (2008). Identity, supervision, and work groups. *American Economic Review 98*, 212–217.

Akerlof, G. A. and R. E. Kranton (2010). *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-being*. Princeton University Press.

Akerlof, R., N. Matouschek, and L. Rayo (2020). Stories at work.

Aumann, R. J. (1987). Correlated equilibria as an expression of Bayesian rationality. *Econometrica 55*, 1–18.

Besley, T. and T. Persson (2017). The joint dynamics of organizational culture, design, and performance. Working paper.

Bisin, A. and T. Verdier (2017). On the joint evolution of culture and institutions. NBER working paper.

Brynjolfsson, E. and P. Milgrom (2013). Complementarity in organizations. In R. Gibbons and J. Roberts (Eds.), *Handbook of Organizational Economics*. Princeton University Press.

Camerer, C. and A. Vepsalainen (1988). The economic efficiency of corporate culture. *Strategic Management Journal 9*, 115–126.

Carrillo, J. D. and D. Gromb (1999). On the strength of corporate cultures. *European Economic Review 43*, 1021–1037.

Carrillo, J. D. and D. Gromb (2007). Cultural inertia and uniformity in organizations. *The Journal of Law, Economics, & Organization 23*(3), 743–771.

Crémer, J. (1993). Corporate culture and shared knowledge. *Industrial and Corporate Change 2*, 351–386.

Crémer, J., L. Garicano, and A. Prat (2007). Language and the theory of the firm. *Quarterly Journal of Economics 122*, 373–407.

Delfgaauw, J., R. Dur, O. A. Onemu, and J. Sol (2021). Team incentives, social cohesion, and performance: A natural field experiment. *Management Science*. Forthcoming.

Dessein, W. and A. Prat (2019). Organizational capital, corporate leadership, and firm dynamics. Columbia business school research paper.

Dessein, W. and T. Santos (2006). Adaptive organizations. *Journal of Political Economy 114*, 956–995.

DiMaggio, P. (1997). Culture and cognition. *Annual Review of Sociology 23*, 263–287.

Gibbons, R. and R. Henderson (2013). What do managers do? Exploring persistent performance differences among seemingly similar enterprises. In R. Gibbons and J. Roberts (Eds.), *Handbook of Organizational Economics*. Princeton University Press.

Gibbons, R., M. LiCalzi, and M. Warglien (2021). What situation is this? Shared frames and collective performance. *Strategy Science*. Forthcoming.

Goldman, A. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.

Gopnik, A. and H. Wellman (1994). The "theory theory". In L. Hirschfield and S. Gelman (Eds.), *Mapping the mind: Domain specificity in culture and cognition*, pp. 257–293. Cambridge University Press.

Guiso, L., P. Sapienza, and L. Zingales (2015). Corporate culture, societal culture, and institutions. *American Economic Review 105*(5), 336–39.

Halac, M., E. Lipnowski, and D. Rappoport (2021). Rank uncertainty in organizations. *American Economic Review*. Forthcoming.

Hermalin, B. E. (2001). Economics and corporate culture. In C. L. Cooper, S. Cartwright, and P. C. Earley (Eds.), *The International Handbook of Organizational Culture and Climate*, pp. 217–261. John Wiley and Sons.

Holmstrom, B. (1982). Moral hazard in teams. *The Bell Journal of Economics*, 324–340.

Kaplan, S. and R. Henderson (2005). Inertia and incentives: Bridging organizational economics and organizational theory. *Organization Science 16*(5), 509–521.

Kets, W., W. Kager, and A. Sandroni (2019). The value of a coordination game. Working paper.

Kets, W. and A. Sandroni (2019). A belief-based theory of homophily. *Games and Economic Behavior 115*, 410–435.

Kets, W. and A. Sandroni (2021). A theory of strategic uncertainty and cultural diversity. *Review of Economic Studies 88*, 287–333.

Koszegi, B. (2014). Behavioral contract theory. *Journal of Economic Literature 52*(4), 1075–1118.

Kotter, J. P. and J. L. Heskett (1992). *CorporateCulture and Performance*. Free Press.

Kreps, D. M. (1990). Corporate culture and economic theory. In J. Alt and K. Shepsle (Eds.), *Perspectives on Positive Political Economy*, pp. 90–143. Cambridge University Press.

Lazear, E. P. (1999a). Culture and language. *Journal of Political Economy 107*(S6), S95–S126.

Lazear, E. P. (1999b). Globalisation and the market for team-mates. *Economic Journal 109*, C15–C40.

Palfrey, T. R. (1990). Implementation in Bayesian equilibrium: The multiple equilibrium problem in mechanism design.

Singer, T. and E. Fehr (2005). The neuroeconomics of mind reading and empathy. *American Economic Review 95*, 340–345.

Van den Steen, E. (2010). Culture clash: The costs and benefits of homogeneity. *Management Science 56*, 1718–1738.

Weber, R. A. and C. F. Camerer (2003). Cultural conflict and merger failure: An experimental approach. *Management science 49*(4), 400–415.

Winter, E. (2004). Incentives and discrimination. *American Economic Review 94*(3), 764–773.