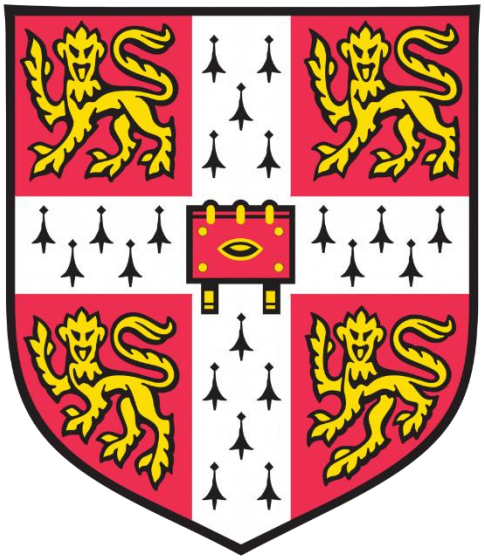
Firms and the Evolution of Culture and Cooperation

FRANCISCO JOSE BRAHM MORALES

Clare Hall

University of Cambridge



August 2018

The dissertation is submitted for the degree of

Doctor of Philosophy

Para Camila, el amor de mi vida

Para Diego y Rocío, la alegría de nuestras vidas!

Para mis Padres, por su amor incondicional

Para el Coco, seguiremos transmitiendo tu llama

"Nothing is more fundamental in setting our research agenda and informing our research methods than our view of the nature of human beings whose behaviors we are studying. . . It makes a difference to research, but it also makes a difference for the proper design of . . . institutions"

Herbert Simon

"We are much better at learning from others than animals are, and equally important, we are motivated to learn from others even when we do not understand why our models are doing what they are doing. This psychology allows human populations to accumulate pools of adaptive information that greatly exceed the inventive capacities of individuals. Cumulative cultural evolution is crucial for human adaptation."

Robert Boyd

"Cooperation can be seen as the master architect of evolution, as the third fundamental principle of evolution beside mutation and selection."

Martin Nowak

"It must not be forgotten that although a high standard of morality gives but a slight or no advantage to each individual man and his children over the other men of the same tribe, yet that an increase in the number of well-endowed men and an advancement in the standard of morality will certainly give an immense advantage to one tribe over another … At all times throughout the world tribes have supplanted other tribes; and as morality is one important element in their success, the standard of morality and the number of well-endowed men will thus everywhere tend to rise and increase"

Charles Darwin

Declaration of Originality

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Abstract. At least two thirds of the original contribution of each of the co-authored papers is my own work. The dissertation has not previously been submitted to any university for any degree or other qualification and does not exceed the maximum length stipulated by Cambridge Judge Business School. All sources of information are acknowledged and referenced in the text and bibliography.

Acknowledgements

First, I would like to thank Professor Christoph Loch for being a wonderful supervisor and supporter of my work and research. In the midst of his busy schedule, he constantly devoted quality time to meet and talk about my different research projects. He encouraged me and provided the necessary freedom to pursue the topics I deeply cared about. I have learned a lot from Christoph, not only from his breadth and depth of knowledge, but also from his character. I am privileged to have been his student.

I would also like to thank Professor Stelios Kavadias for all his support and guidance. Our many conversations have been very valuable for my development as a scholar. To the faculty and staff at the Cambridge Judge Business School – the support I have received from you has been wonderful. As a PhD student, everything has been in place so that I can focus entirely on our coursework and research. In particular, I would like to thank Professor Jennifer Howard-Grenville, Professor Stefan Scholtes, and Ms. Joanna Blakeman for a great job running the PhD program. I also thank the OTM faculty, Dr. Vincent Mak, Dr. Paul Kattuman, Dr. Dominique Lauga, Dr. Andreas Richter, Professor Yasemin Kor, Professor Sucheta Nadkarni, and Professor David De Cremer for having supported me, each one in its own unique way. To all of the members of Clare Hall College, particularly all the staff – you made Cambridge my family's home and I will forever be in your debt. To Brian Silverman of the Rotman School of Management – thanks for providing valuable feedback on this dissertation. In addition, I would like to express my sincere gratitude to the Cambridge Trust and CONICYT. This research would not have been possible without their generous financial support.

This journey would have not been possible without the support of my family. To my wife Camila – I will be forever grateful for your selfless support in the pursuit of my dream of becoming an Academic. Moving abroad is challenging, and I deeply value all the sacrifices that you have made these years. These years have laid a solid foundation for our family's future. To our children, Diego and Rocio, the joy you bring to our life is unmeasurable. Any simple gesture –a look, a laugh, a call– provided all the energy I needed. To my parents, whose unconditional love and support have been an irreplaceable foundation in my life. You made this journey easier in so many ways. To my siblings, who have enthusiastically followed our progress abroad, particularly Jorge,

whose many visits brought joy to his godson. To my parents-in-law for their support, especially for constantly connecting with Camila and Diego, and visiting us at the time of Rocio's birth.

I would also like to thank the Economics and Management Department at the Pontifical Catholic University of Chile (PUC), which supported the continuation of pre-PhD research projects while in Cambridge and during my visits to Chile. Special thanks to my long time mentor and collaborator, Jorge Tarzijan. Without your generous support and guidance, my transition towards academia would have been much harder. Our collaboration certainly counts as a "PhD in the shadows"; it prepared me incredibly well for a PhD and a career in academia. Marcos Singer was also extremely generous, supportive and resourceful in facilitating that transition stage of my career.

Last, but not least, I would like to thank my fellow PhD students. Thanks Antoine and KC for many conversations around our research, our projects, our challenges and successes. I deeply cherish our friendship. To Corinna, for being a caring and sharing peer, particularly at the job-market stage. To Armando, Katie, Systke, Geoffroy, Isabel, Jan, Shi, Niklas, Jan, Stavros, and Charlie for being great fellow travellers in this journey. Best of luck to you all in continuing the knowledge expansion adventure!

Preface

As a researcher, I am deeply interested in organizations, broadly understood as the relatively stable collection of individuals that share a common goal. I focus my attention on the business firm, including its variants and subdivisions. My aim is to understand two issues about organizations. First, what is their nature, that is, what essential forces explain their origins as well as their functions for agents and economies at large. Second, how can we better understand and design their structures –both formal and informal– in order to improve their performance.

Let me describe how this research identity unfolded over time (and how this PhD dissertation fits in).

My interest for academic research began with my undergraduate thesis to obtain a Bachelor of Science in Management. At that time, given that in microeconomics courses I was told that competition would wipe abnormal rents out, and that reality wasn't exactly matching the theory, I became puzzled by the drivers of firm performance. After wasting several months in order to realize that a quest for generic drivers of high performance –that is, invest in or execute 'A' and performance will ensue – was akin to the search of a philosophers' stone in medieval alchemy, I discovered and settled on the topic of firm performance variance decomposition and persistence analysis. That is, I discarded a search for general drivers, to focus on their "locus" and "sustainability". In that thesis, we decomposed the return over assets of Chilean listed companies into several components, studied their persistence and compared them to the US[1]. A peer-reviewed publication came out of this effort (Tarzijan, Brahm and Daiber, 2008), which immediately planted a seed: I realized early on that I could contribute to scientific knowledge and, best of all, that I could share it with the world.

Blossoming would wait. After graduation, I embarked in (senior) managerial positions for a span of five years. Over time, however, I kept struggling with trying to understand the object that I was trying to manage: what is a firm? why not only markets? It happened that my Bachelor of Science

[1] We showed that, consistent with research in the United States (US), "firm effects" was the largest component of the variance in ROA, but that "corporate" and "industry effects" carried a higher relative weight in Chile; regarding persistence, we found that in Chile it was explained more evenly by the business-, industry- and corporate- effects in comparison to the US.

was heavy on economics, so I become somewhat knowledgeable on the nature of markets. However, as a manager a few years later, I realized that I had a poor understanding of what actually a firm is. The question of the nature of the firm nicely complemented my previous interests on the roots of firm performance. My curiosity about academia grew.

Circumstances led to the opportunity of completing a Master of Science in Management, and after that, a dual career as a consultant and as an associate researcher at PUC. I gave the seed a good chance. Of course, the theory of the firm became the topic of the MSc's dissertation. To ground this topic, I adopted the typical approach in the literature, which is to study the boundary between the market and the firm, that is, the make-or-buy problem. By understanding the frontier, so the argument goes, we would understand the nature of firms. I did this using a fine-grained dataset of construction projects. Crucially, several academic publications emanated from the dissertation[2]. I loved the process of going from dissertation to papers: refining the arguments, engaging with reviewers, presenting you work at seminar and conferences were all quite fun. Moreover, it was not all that difficult. They say you know you have a talent for something when you do it with relative ease, and you enjoy it! I was on track; the seed had become a plant.

Over those four years as a research associate, I naturally expanded out from the theory of the firm, by discovering and studying a broader field, organizational economics. This field combines contract theory, transaction cost economics and property rights theory in order to understand how organizations, typically the business firm, are formally structured (and how this affects performance). Formal structure is a broad term that includes a myriad of (largely observable and enforceable) choices: the extent of delegation vis a vis centralization, the use of incentives and monitoring, make-or-buy choices, corporate diversification, contracting with external parties, among many others. I engaged with this literature executing several projects[3].

---

[2] In Brahm and Tarzijan (2014) we show that pre-existing capabilities not only increase vertical integration, they also mitigate the positive impact of transaction hazards on integration. In Brahm and Tarzijan (2013, 2012) we show that different integration choices are complementary to each other.

[3] In a project that ended published in Brahm and Poblete (2017), we studied the dynamics of incentive schemes by way of a field experiment in a Chilean salesforce. In Brahm and Tarzijan (2016) we show how vertical integration of activity 'A' promotes centralization in the interacting activity 'B'. In Brahm, Tarzijan, and Singer (2017) we analyse how increases in product diversification reduce the efficiency of operational routines.

However, two nuisances kept surfacing. First, by drawing from organizational economics, I used self-interested maximization as the main behavioural assumption in my research. And, of course, by simple introspection we know that this is only half of the action, and within organizations, most likely less. Trust, cooperation and altruism, and the idea of social norms as the guide for action (instead of calculativeness), are surely a large part of what constitutes organizations and what makes them perform.

The second nuisance was that, largely by engaging with real world organizations as a consultant, I noticed that the "informal organization" of companies carries a disproportionate weight in explaining firm behaviour and performance. In addition, this aspect was absent from the main extant theories of the firm (Baker et al, 2002 is a notable exception). Informal organization can be (broadly) referred to as "firm culture", that difficult-to-define concept encompassing, social norms, tacit knowledge, routines, values and relationships (in the lingo of contract theory, informal organization encompasses all that is difficult to observe/enforce and therefore, displays limited contractibility). Although I had already contributed to informal organization research (mostly by studying its interaction with formal organization)[4], it felt that I had only scratched the surface of the informal aspects of organizing.

By now, as you might imagine, I was well on my way to a PhD, having decided to pursue it at CJBS. Quite naturally, during my PhD I decided to dive into informal organization and culture, both to understand organization and performance better and to probe into the nature of firms. I consider this a "big" topic that, at the same time it naturally flows out of my previous interests, it is novel and rich enough to provide the steam for a PhD and for the next decade of research.

A "big" topic requires focus. At the start of the PhD I scanned different theoretical approaches that would suit my topic and complement my quantitative research style. I quickly settled on two related evolutionary approaches, rooted in basic disciplines: "Cultural Evolution", developed by evolutionary anthropologists Robert Boyd and Peter Richerson (and their collaborators) (Boyd and

[4] In Brahm and Tarzijan (2016b) we show that relational contracts stemming from expected future interaction decrease vertical integration, especially if prior interactions and asset specificity are present. We also demonstrate that prior interactions increase the use of fixed-price contracts in mega-projects (Brahm and Tarzijan, 2015). In Brahm and Singer (2013) we show that the most effective safety training methods are those whose structure promote worker engagement.

Richerson, 1985 and 2005; Henrich, 2004 and 2015), and "Evolution of Cooperation", championed by mathematical biologist, Martin Nowak (and his collaborators) (Nowak, 2006; Rand and Nowak, 2013). The former is the foundation for the second section of this dissertation; the later for the third section. Below I introduce each section in the abstract section, and then, in the introduction, I connect these approaches with the sections.

The scholarly journey so far has been extremely rewarding. I am very lucky to be well embarked on a career of knowledge discovery and dissemination. I am hopeful, as well as confident, that I will be able to gradually grow a tree that casts a long shadow.

References

Baker, G., Gibbons, R. and Murphy, K.J., 2002. Relational Contracts and the Theory of the Firm. The Quarterly Journal of Economics, 117(1), pp.39-84.

Boyd, R., Richerson, P.J. 1985. Culture and the Evolutionary Process. Chicago: University of Chicago Press.

Boyd, R., Richerson, P.J. 2005. The Origin and Evolution of Cultures. Oxford: Oxford University Press.

Brahm, F., Poblete, J. 2017. Incentives and ratcheting in a multiproduct firm: A field experiment. Management Science, forthcoming.

Brahm, F., Singer, M. 2013. Is more engaging safety training always better? Evidence from Chilean panel data". Journal of Safety Research, (47): 85-92.

Brahm, F., Singer. M 2017. Do individuals and teams respond differently to co-located competitors under different contract choices? Evidence from supermarkets, working paper.

Brahm, F., Tarziján, J. 2012. The impact of complexity and managerial diseconomies on hierarchical Governance, Journal of Economic Behavior and Organization, 84(2): 586-599.

Brahm, F., Tarziján, J. 2013. Boundary choice interdependency: Evidence from construction firms. Industrial and Corporate Change, Vol 22(5): 1229-1271.

Brahm, F., Tarziján, J. 2014. Transactional hazards, capabilities, and institutional change: Integrating the theories of the firm. Strategic Management Journal, 35(2): 224-245.

Brahm, F.,Tarziján, J. 2015. "Does Complexity and Prior Interactions Affect Project Procurement? Evidence from Mining Mega-Projects". International Journal of Project Management, Vol 33, Nº 8, p. 1851-1862

Brahm, F., Tarziján, J. 2016a. Toward an integrated theory of the firm: The interplay between internal organization and vertical integration. Strategic Management Journal. 37(12): 2481-2502.

Brahm, F., Tarziján, J. 2016b. Relational Contracts and Collaboration in the Supply Chain: Impact of Expected Future Business Volume on the Make-or-Buy Decision. Journal of Supply Chain Management, Vol 52, Nº 3

Brahm, F., Tarziján, J., Singer. M 2017. The impact of frictions in routine execution on economies of scope. Strategic Management Journal, forthcoming.

Henrich, J. 2004. Cultural group selection, co-evolutionary processes and large-scale cooperation. Journal of Economic Behavior & Organization, 53(1): 3-35.

Henrich, J. 2015. The Secrets of our Success. Princeton University Press.

Nowak, M. A. 2006. Five rules for the evolution of cooperation. Science,314(5805): 1560-1563.

Rand, D. G., Nowak, M. A. 2013. "Human Cooperation." Trends in Cognitive Sciences 17 (8): 413-425.

Abstract

The second section, titled "The Evolution of Productive Organizations", attempts to break new ground in our explanations of the nature of guilds, partnerships and other pre-modern firm-like organizations (executed in collaboration with Joaquin Poblete[5]). I use the theory of Cultural Evolution to develop a foundation for the evolutionary origins of firms. In extant theory, a historically rooted explanation for the evolution of firms is largely absent. I argue that Cultural Evolution theory can address this challenge, putting knowledge, learning, and cooperation at centre-stage. This theory, developed in Evolutionary Anthropology, studies culture as an evolutionary system. It specifies micro-foundational mechanisms for inheritance, selection and variation. Culture is defined as information that is acquired from other individuals via social learning mechanisms, such as imitation and teaching. Information includes beliefs, norms, knowledge, skills and artefacts/technology.

In this second section, we develop a cultural evolution model that illuminates the evolution of pre-modern productive organizations, such as, partnerships and guilds. Specifically, we introduce productive organizations in a workhorse cultural evolution model, widely used to explore the conditions that make social learning fitness enhancing. If organizations are exclusive and facilitate social learning, they stop the negative externality generated by the replication of social learners. The basic insight provided by the model is that productive organizations evolved because they favoured the conditions that sustain the process of cumulative culture. Productive organizations make social learning –and therefore culture– useful to society, playing an important role on the adaptive success of the human species.

Our model has predictions regarding the benefits of organizations for society that are at odds with standard models of firms in economics and management based on transaction costs. For example, while in transaction costs theories the firm-like organizations is more valuable when uncertainty

[5] Joaquín Poblete contributed by translating our many conversations into a theoretical model. Proofs of propositions were executed by Joaquín. The solution of the model, as well as the simulations, were executed by myself. Most of the writing in the introduction, plus half of the model and extensions sections, was executed by myself. The empirical section of the section is my exclusive contribution.

is high, in our model the firm-like organization is more valuable when it is low. These differences allow for empirical comparison of the theories.

We test our theory using data from the Ethnographic Atlas and the Standard Cross Cultural Sample. We measure the presence of technologies in pre-modern societies (e.g., weaving, metalworking, pottery) and whether they were used throughout the society or mainly by a small group of people, that is, within a productive organization. Across several tests and robustness checks, we find consistent evidence for the propositions and comparative statics of our model.

The third section, titled "The evolution and Impact of Cooperation in Large Groups: Evidence from Administrative Data and a Field Experiment" (executed in collaboration with Christoph Loch and Cristina Riquelme[6]), I zoom in on the informal structure of firms (or "culture") by studying the drivers of cooperation in large groups. As groups grow large, it is increasingly hard for workers to accept to pay a cost in order to provide a benefit to colleagues or the group at large. There is a vast theoretical literature in the fields of evolutionary biology and evolutionary anthropology regarding the conditions and mechanisms that favour the evolution of cooperation in large populations (i.e., increase in frequency). For cooperation to evolve, a mechanism is required that allows favouring cooperators over defectors. This mechanism is an interaction structure that specifies *who interacts with whom* in the population (i.e., random v/s structured v/s flexible) and *how* the agents interact in order to receive payoffs (e.g., what is known by whom, degree of repetition, order of play, details of payoff functions, enforcement technology). On the former element of 'how', the main mechanisms are spatial/network selection and group selection; on the latter, direct reciprocity and indirect reciprocity.

Using this theory, we collaborate with three organizations to study a workplace safety practice that is based on voluntary cooperation by workers. In this practice, an initial core of cooperators strives to expand cooperation within the implemented site (e.g., plant or store). The methodology leverages cooperation: training and counselling is costly to observers while the benefits of

[6] Christoph Loch contributed by guiding the research effort, particularly in the treatment design stage. He also contributed in thoroughly editing the text. Cristina contributed with the execution of the experiments on the ground, particularly in the execution of the randomization, the entry survey, the exit interviews, the monitoring of treatment progress, the measurement of treatment take-up, and the compilation and preliminary analysis of accident data. All these activities were performed under my guidance.

improved safety flow mostly to the observed workers. Moreover, the strive for expansion provides a unique setting to study the evolution of cooperation (i.e., its increase in frequency). Using a detailed administrative dataset, we first show that the methodology reduces accidents and improves culture, documenting the power of cooperation. However, the dataset also demonstrates that, in line with theories of cooperation breakdown in large groups, this positive impact decreases very quickly as the number of observers expands.

Then, we examine the idea of interaction structures, by analysing the impact of direct reciprocity in two ways. First, using the administrative dataset, we document traces of the positive impact of direct reciprocity in the adoption and impact of the practice. Second, we executed a field experiment in four sites where we intervened the established safety methodology with a baseline "direct reciprocity" treatment, plus two additional interaction treatments, aimed at solving the breakdown problem. We show that: i) the effort of the additional observers is restored when the expansion of observers is structured around small groups (1st treatment – "Direct Reciprocity"), ii) lifting the anonymity of the observed workers is detrimental to observers' effort, eliminating the benefits of direct reciprocity (2nd treatment – "Identity"), and iii) public display of effort is mute (3rd treatment – "Indirect Reciprocity"), but interacts with the 'private enforcement' –measured with administrative data– in subtle ways. Further, we find that these treatment effects on effort translate into the speed of diffusion (i.e., the likelihood of becoming observer) and into safety outcomes (i.e., safe behaviour and accidents of the workforce): both increase with treatment 1 but decrease with treatment 2.

Overall, the third section provides unique field evidence of cooperation breakdown when groups grow large, as well as of "structured growth" (sustained by direct reciprocity) as a crucial mechanism that allows for its recovery and evolution.

TABLE OF CONTENTS

1. Introduction

The human species is unique. We have conquered every habitat on earth --from dry deserts, to the cold at the poles, and everything in between. We cooperate with unrelated individuals across time and space at rates unmatched in nature, even by eusocial species (which mostly cooperate locally among related individuals). What explains these two outstanding feats? The answer to this question will allow me to place both studies of this dissertation in context.

The field of Cultural Evolution provides a cogent answer based on a few very simple but precise (micro-founded) theoretical models, and a burgeoning empirical agenda (Boyd and Richerson, 1985 and 2005; Richerson et al, 2013; Boyd; 2018; Boyd and McElreath, 2007; Henrich, 2015; Laland, 2017). The unique (compared to the animal kingdom) capacity of human beings to learn from one another –labelled "social or cultural learning"– allows innovations to accumulate gradually over time. This happens even if social learners do not understand why a trait or behaviour might be superior or why a model is performing it: natural selection, but more importantly, social learning biased towards the majority, the skilful, or the successful can do the trick. This process of cultural adaptation operates fast: it has been estimated to be at least 50 times quicker than genetic adaptation (Boyd, 2018).[7]

The nature of biases in social learning generates stable heterogeneity between groups, that is, groups come to develop behavioural and technological traditions (e.g., norms and tools) that are sturdy and often different due to drift. In contrast to genetic evolution, where little migration dissolves genetic differences, this stable heterogeneity allows group selection (or multilevel selection) to be a powerful evolutionary force (Wilson, 2015; Turchin, 2015): more capable groups outcompete and replace others. "Capable" here refers to groups that have a deeper cultural and knowledge baggage, and, as the Darwin quote on the opening pages indicates, that cooperate successfully.

[7] Explaining the evolution of social learning itself –the basic primitive in the theory– is however much tougher because it is frequency dependent. A few theoretical explanations have been proposed (see Boyd and Richerson, 2005, Boyd, 2018 or Laland, 2017).

In Figure I we classify the evolutionary process of cultural evolution in a two-by two matrix. In the horizontal dimension, we divide the elements that evolve in culture into cooperation and other cultural traits (especially knowledge-based "technologies", or ways of doing things). The vertical dimension distinguishes two scales (in size and time) of cultural evolution, Micro- and Macro-evolution. We analyse first the horizontal dimension.

Figure I. Matrix of "what evolves when"

		Aspect that evolves:	
		Knowledge / Technology / Norms / Institutions	Cooperation
Timescale and extent of evolution:	Micro-evolution	I	III SECTION 3: "The evolution and impact of large scale cooperation"
	Macro-evolution	II SECTION 2: "The evolution of productive organizations"	IV

Cooperation –individuals bearing a cost to provide a benefit to a third party or the group at large– is a fundamental explanandum in cultural evolution. Natural or cultural selection favours individual fitness/utility, and therefore it should get rid of cooperation. However, cooperation is ubiquitous. A substantial contribution to understanding the puzzle of the evolution of cooperation comes from mathematical biology, specifically from the contribution made by Martin Nowak and his collaborators in the last two decades (Nowak, 2011; Rand and Nowak, 2013)[8]. The approach by Nowak et al is broad, generating insight valid for non-volitional entities (cells, animals) as well as for humans. As discussed in section 3, the crucial concept is that of "interaction structures", a short hand for the details of 'who interacts with whom and how' in a population. These structures

[8] The contributions by W. D. Hamilton and Robert Axelrod were fundamental in launching this research stream (Axelrod and Hamilton, 1981).

are mechanisms that favour co-operators against defectors. These are 'direct reciprocity', 'indirect reciprocity', 'spatial/network selection' and 'kin-selection'. Contributions by scholars more directly identified with Cultural Evolution focus mostly on the mechanisms of 'group selection': cooperative groups would survive and expand (Turchin, 2015; Henrich, 2005; Bowles, 2009)[9].

The second aspect that evolves includes, (very) broadly speaking, all the remaining cultural traits other than cooperation. A broad brush would classify these as knowledge and technology on one side (e.g., tools, blueprints, productive routines) and norms and institutions on the other (e.g., "eye for an eye", encoded law, centralized state). The standard biased social learning story is the main mechanism driving the evolution of knowledge, technology and norms of behaviour (Henrich, 2015). Here information and traits are assumed to be transmitted across individuals. In contrast, given their group-level nature, group selection is the main mechanism argued for the evolution of institutions such as religions or centralized states (Turchin, 2015).

In sum, social learning generates cumulative knowledge and culture, while interaction structures allow for cooperation to spread; group selection sits on top, further propelling both.

The story is also self-referent, becoming more subtle (and more insightful!). Given the adaptive value of cooperation and cumulative culture, natural selection pressures will generate the evolution --that is, the increase in frequency-- of biological or cultural adaptations that favour the operation of social learning or the interaction structures. An example in biology is the capacity of human language which probably evolved because it boosted the mechanism of indirect reciprocity (Nowak, 2011) or the capacity for social learning, particularly teaching (Laland, 2017) (the same could be said about writing, a cultural trait). Example in culture are political complexity and big-god religions, which may have evolved because they fostered large scale cooperation, which in turn provided advantages to tribes and nations in the group-selection process (Norenzayan et al, 2016; Turchin, 2015).

The second dimension of Figure I displays the timescale and the extent of evolution. Microevolution studies whether a trait can evolve, and how it does, in a population of agents. It

[9] There is a big focus on punishment, both peer-to-peer or centralized, as a key mechanism propelling cooperation. However, as nicely noted by Nowak, punishment is a cooperative trait that requires an evolutionary explanation in the first place.

uses evolutionary game theory where strategies and payoff are specified, and replication dynamics are assumed – that is, in every period a small proportion of agents copies the strategy that is more advantageous. These models are used to study the whether traits can invade, under which circumstances they do so, and study the properties of the evolutionary equilibrium, such as the long-run percentage of different strategies in the population (McElreath and Boyd, 2007). As such, these models are well suited to study what happens within a single population, exploring how strategies evolve. Therefore, they can, and usually are used to, explore shorter time periods to explain local change. The models by Boyd and Richerson or Nowak largely qualify here. Some empirical examples are the studies by Alex Mesoudi where he analyses how different bows and arrows evolve within in the lab (so to inform and explain the recorded archaeological record in North America) (Mesoudi, 2011); the social norms shifts studied by Peyton Young, such as contracting norms in farming (Young and Burke, 2001); and the lab experiments around cooperation surveyed by Rand and Nowak (2013). *This is where the third section of this dissertation operates (quadrant III of Figure I)*: *we study how interaction structures – following the propositions of extant micro-evolutionary models – can favour the evolution of cooperation over the span of one to three years in a specific population (in our case, the sites that implement the safety technology).*

We argue in section 3 that this micro-founded, but nonetheless population level approach, is exactly what is needed for studying cooperation in firms. Cooperation in large groups is a root cause of firm performance, as attested by the ever-present calls for collaboration and team spirit in values statements of companies. Micro-evolutionary models can unveil the mechanics of cooperation and of other cultural traits that manifest themselves at the group level. In comparison to cultural microevolution models, current approaches in management and economics deal poorly with these. In organization theory, cultural traits such as cooperation are simply described and categorized without addressing the underlying mechanics (e.g., Giorgi et al, 2015). In economics, modelling approaches using rational maximization tend to focus on dyadic relations, such as models of relational contracting (Gibbons and Henderson, 2012), and therefore, they eschew the large group interactions –or simply extend the conclusion of a dyadic model to the group level.

(This is probably done because, if N rational agents are assumed, models might become intractable; however, promising strides are being made in network economics)[10].

In contrast, macroevolution[11] studies differences across populations over a much longer time span, such as ethnic groups, nations, states or countries (Gray and Watts, 2017). What is typically compared are traits that have largely invaded a population and have become established (fixed) in it[12]. A typical approach is to study how traits have co-evolved across the different branches of the evolutionary tree of populations across the globe. For example, Watts et al (2016) show, by way of a phylogenetic comparison of Austronesian societies, that ritual human sacrifice promoted and sustained the evolution of stratified societies. With the advent of long run longitudinal datasets covering regions, group selection hypotheses have started to be rigorously tested and confirmed (Turchin et al, 2013)[13].

However, macroevolution runs the risks of offering unfounded "functional explanations", that is, arguing that a trait 'X' exist in a population because it provides some benefits to the population, with too slim a description of the bi-directional connections with benefits or the micro-mechanisms that generated them (Elster, 1983)[14]. To avoid this, macroevolutionary explanations require to be complemented by micro-evolutionary models. These models provide a way to show the mechanics of the benefits generation process, and then test all the predictions and comparative statics that they generate. *This is what section 2 accomplishes: we provide a long run macro-evolutionary explanation for a cultural "technology" that has invaded human societies –productive*

[10] The "easy" case of market-mediated cooperation is an exception rather than the rule: it focuses on the case where individuals do not have externalities (i.e., they don't play a game) and thus individual maximization generates maximum collective welfare (Frank, 2012).

[11] A mapping can be done between the Macro- and Micro-evolution and the concepts of ultimate and proximate causation and the derived "four questions of Tinbergen" regarding ontogeny, control, function and phylogenies (Bateson and Laland, 2013). However, a full description of this mapping is beyond the scope of this introduction. Its absence is not detrimental to the contextualization and framing of the sections of this dissertation.

[12] If all populations have the trait, then the analysis becomes extremely difficult. This is the case of human language, where all known populations display it. Comparison with primates can shed some but not definitive light on the issue (Laland, 2017).

[13] The Seshat database is promising. An example is Turchin et al (2017) that shows that several traits of cultural complexity (e.g., social scale, economy, features of governance, and information systems) display strong evolutionary relationships with each other in a complementary way, and thus, can be summarized in one principal component. This suggest the presence of an underlying general principle in historical evolution of cultural complexity (instead of many "equally valuable" configurations).

[14] Of course, this does not considers intentional/purposeful explanations –which indeed can generate traits simply by invoking to the intended / expected benefits of agents. This is the explanatory currency in most of economics.

organizations such as guilds, roman societae, and partnerships– by arguing from a micro-evolutionary model; this allow us to empirically test not only the benefits of the mere presence of productive organizations, but also the myriad of comparative statics that emanate from the model.

References

Bateson, P. and Laland, K.N., 2013. Tinbergen's four questions: an appreciation and an update. Trends in ecology & evolution, 28(12), pp.712-718.

Boyd, R. 2018. A Different Kind of Animal: How Culture Transformed Our Species. Princeton University Press.

Boyd, R., Richerson, P.J. 1985. Culture and the Evolutionary Process. Chicago: University of Chicago Press.

Boyd, R., Richerson, P.J. 2005. The Origin and Evolution of Cultures. Oxford: Oxford University Press.

Boyd, R., Richerson, P.J. and Henrich, J., 2013. The cultural evolution of facts: Facts and Theories. Cultural evolution: society, technology, language, and religion, 12, p.119.

Bowles, S. (2009). Did warfare among ancestral hunter-gatherers affect the evolution of human social behaviors?. Science, 324(5932), 1293-1298.

Elster, J., 1983. Explaining technical change: A case study in the philosophy of science. CUP Archive.

Frank, R. H. 2012. The Darwin Economy: Liberty, Competition, and the Common Good. Princeton University Press.

Gibbons, Robert, and Rebecca Henderson (2012). "Relational contracts and organizational capabilities." Organization Science 23, no. 5: 1350-1364.

Giorgi, S., Lockwood, C. and Glynn, M.A., 2015. The many faces of culture: Making sense of 30 years of research on culture in organization studies. The academy of management annals, 9(1), pp.1-54.

Gray, R.D., Watts, J., 2017. Cultural macroevolution matters. Proceedings of the National Academy of Sciences, 114(30), pp.7846-7852.

Henrich, J. 2004. Cultural group selection, co-evolutionary processes and large-scale cooperation. Journal of Economic Behavior & Organization, 53(1): 3-35.

Henrich, J. 2015. The Secrets of our Success. Princeton University Press.

Laland, K. 2017. Darwin's Unfinished Symphony: How Culture Made the HumanMind. Princeton University Press

McElreath, R., & Boyd, R. (2007). *Mathematical Models of Social Evolution: A Guide for the Perplexed*. Chicago: Univ of Chicago Press.

Mesoudi, A. (2011) Cultural Evolution: How Darwinian theory can explain human culture and synthesize the social sciences. Chicago, IL: University of Chicago Press.

Norenzayan, A., Shariff, A.F., Gervais, W.M., Willard, A.K., McNamara, R.A., Slingerland, E. and Henrich, J., 2016. The cultural evolution of prosocial religions. Behavioral and brain sciences, 39.

Nowak, M. A. 2006. Five rules for the evolution of cooperation. Science,314(5805): 1560-1563.

Nowak, M. A. 2011. Supercooperators: Altruism, Evolution, and Why We Need Each Other to Succeed. Simon and Schuster.

Rand, D. G., Nowak, M. A. 2013. "Human Cooperation." Trends in Cognitive Sciences 17 (8): 413-425.

Turchin, P. 2015. Ultrasociety: How 10,000 Years of War Made Humans the Greatest Cooperators on Earth. Berest books.

Turchin, P., Currie, T., Turner, E., Gavrilets, S. 2013. "War, Space, and the Evolution of Old World Complex Societies." *Proceedings of the National Academy of Sciences* 110 (31): 16384–89.

Turchin, P. et al, 2017. Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization, PNAS, early internet view.

Watts J, Sheehan O, Atkinson QD, Bulbulia J, Gray RD (2016) Ritual human sacrifice promoted and sustained the evolution of stratified societies. Nature 532:228–231

Wilson, D.S. 2015 Does Altruism Exist? Culture, Genes and the Welfare of Others. Yale University Press.

Young, H.P. and Burke, M.A., 2001. Competition and custom in economic contracts: a case study of Illinois agriculture. American Economic Review, 91(3), pp.559-573.

2. The Evolution of Productive Organizations

Abstract

We develop a cultural evolution model that illuminates the evolution of pre-modern productive organizations, such as, partnerships, roman societae and guilds. The basic insight provided by the model is that productive organizations evolved because they favoured the difficult-to-propel process of cumulative culture by virtue of being exclusive and facilitating social learning. Productive organizations make social learning and culture useful to society, playing a fundamental role on the adaptive success of the human species. The model also illuminates issues regarding adaptation and rigidity, the locus of innovation, secrecy and the origins of specialization. We test the model using a sample of pre-modern societies drawn from the Ethnographic Atlas. The empirical analysis provides supportive evidence for our predictions.

KEYWORDS: Organizations, Theory of the Firm, Knowledge, Social Learning, Cultural Evolution, Specialization.

2.1. Introduction

Organizations, defined as a stable collection of individuals with a common goal, have played a crucial role throughout human history. Hunting bands, armies, academies and churches are some examples. One of such organizations is the "productive organization" (PO) whose goal is to produce goods and services that satisfy the material needs of human populations (e.g., food, shelter, tools). Whether it is the societas in roman times (Hansmann et al., 2006), the guilds in medieval times (Ogilvie, 2014), or the partnerships in early renaissance (Padgett et al., 2006), goods and services of POs have consistently sustained towns, cities and states. In the last century and a half, the influence of modern firms such as large corporations cannot be overstated: their scale and reach dominates modern economic life (Chandler, 1990)[15].

[15] Although the theory we develop in this paper lends itself to a broad application, organizations are varied. We focus on POs as we judge to be it the most direct and informative application of our model.

Organizations are varied, so let us contextualize POs at the outset. There is a vast literature in sociology describing and cataloguing organizations (Blau and Scott, 1963; Scott, 2003). First, organizations exclude social aggregates --collection of people in the same place (e.g., public in stadium)-- and networks --fluid collection of connected individuals sharing an interest (e.g., collectors). Organizations are groups --a stable and interacting collection of individuals. But not all groups are organizations. "Primary groups" such as family, household, and friends do not have a specific goal; instead they serve an intimate and emotional supporting role. Organizations, in contrast, are a "secondary group" that exhibit a common and specific goal, often practical or utilitarian. Most organizations also display a high degree of formalization (e.g., roles, rules governing behavior, clear boundaries). POs are located here: they are a large subset of these formal organizations[16]. POs differentiate from another important type of formal organization, the "voluntary organizations", where members can freely join or leave (e.g., some charitable organizations, clubs, churches); POs, as we define them, have instead the crucial distinction of restricted access. Although the most common type of POs are for profit companies producing goods and services, our definition does not preclude other POs providing more specific services such as defence (e.g., police, army), healthcare and education (e.g., schools, universities). However, we restrict our attention to companies, especially its pre-modern predecessors such as guilds and partnerships, to focus the discussion. This allows contrasting with theories in economics, and it avoids dealing with the nuances that arise when discussing some specific services such as defence or education.

Extensive research in economics and business provides explanations for the existence of POs, focusing on incentives and governance. Under the umbrella of the "theory of the firm", several theories propose, in a nutshell, that firms are a way to avoid the potential hazards involved in the market exchange of goods, labor and assets (Coase, 1937). Under conditions of uncertainty and specificity, transactions among self-motivated agents become costly, favouring the use of a hierarchical organization to govern them. Different theories emphasize different costs of market exchange, and thus different rationales for firms. For example, firms allow for ex-ante investment incentives when parties could behave opportunistically (Hart and Moore, 1990); firms allow for the use

[16] In contrast, "informal organization" refers to the tacit consensus and norms that guide the goal-directed behavior of the group. All "formal organizations" contain informal organization within them; in contrast, informal organizations can exist with minimal, or even none, formalization (however, this is rarer).

of authority and fiat when transaction complexity requires constant coordination (Williamson, 1991); firms solve moral hazard problems that stem from diversity of tasks in a transaction (Holmstrom and Milgrom, 1991), among others. Empirical evidence is supportive of these governance functions of firms (Lafontaine and Slade, 2007).

Recent research indicates, however, that the nature of productive organizations is not only about protection from hazards in exchange or investment, but also about being carriers and transmitters of culture, knowledge and intangible capital. Plenty of evidence is consistent with this view. Recent evidence on a comprehensive sample of US firms shows that vertically integrated companies display a surprisingly small flow of physical goods and a significant flow of knowledge and intangible capital (Atalay et al., 2014); the literature on organizational learning shows that knowledge diffusion is enhanced within firms (Argote and Miron-Spektor, 2011); the existence of persistent performance differentials across firms has been related to firm specific know-how based on relationships and culture (Gibbons and Henderson, 2012); the central function of guilds in medieval Europe was the efficient transmission of skills and tacit knowledge (De La Croix et al., 2017); modern partnerships (e.g., a law firm) essentially provides know-how to its members through effective training and mentoring (Morrison and Wilhelm, 2008). Common to these examples is the idea of knowledge and information transmitted among individuals via social learning (e.g., copying, teaching, apprenticeship). Thus, a first challenge is to incorporate social learning -the basic building block of culture- into the theory of productive organizations.

A second challenge to mainstream theories of POs is that their evolutionary origin is not addressed and, consequently, we lack an understanding of their historical role in the development of our civilization. Current theories of POs focus on modern firms such as corporations, without addressing their relationship with their historical "predecessors" (e.g., partnerships, guilds, roman societas). In current theories, POs are a de novo phenomenon, without explaining how they went from non-existence to current universality. In other words, we lack an understanding of the evolutionary and historically grounded mechanisms that made the POs to be selected, to gradually increase in frequency and to come to dominate the modern economic landscape[17]. Clarifying the

[17] As the following quote attest, this is also true for evolutionary theories of POs: "One kind of glaring omission in our 1982 book was the failure to think about evolution, and industry evolution in particular, in a historical context.

historical role and evolutionary roots of POs is necessary to fully understand their nature and their contribution to our species success.

In this article, we develop and test a theory of the origin of POs that tackles these challenges. First, our theory puts social learning at the centre of the role played by POs. Using a cultural evolution model (Boyd and Richerson, 1995), we show that POs can improve the conditions that are necessary for knowledge, technologies and other cultural traits to accumulate over time, a key driver of cultural complexity and the adaptive success of our species. In particular, POs need to restrict access and facilitate social learning in order to favour cultural ratcheting. This produces an account of POs based on knowledge and culture that does not require incentive and governance considerations. Further, by focusing on knowledge and cultural accumulation we go beyond extant theories that consider knowledge but that focus on the optimal organization of knowledge activities (e.g., hierarchies as a natural response to problems of varying complexity and agents of varying capacity; see Garicano, 2000).

Second, our theory provides a logic for the rise and invasion of POs. In our evolutionary past, the first exclusive and social-learning-enhancing POs favoured the conditions for cumulative culture, and consequently, several mechanisms -natural selection, biased social learning, and group selection- would have operated in selecting the POs, which then gradually invaded the landscape. This provides an additional layer to our understanding of POs. In current incentive-based theories the set of exchanges or investments requiring governance originate from an already-in-place and exogenous pool of knowledge and technologies. Instead, in our theory POs arose and expanded because they favoured the expansion of this pool. Given that knowledge and cultural accumulation is at the base of our species' success, it is natural to look there for the origins of POs.

The theory of the origin of POs applies straightforwardly to pre-modern POs such as guilds and long-standing POs such as partnerships. Consistently, we test our theory in the context of pre-modern societies using the Ethnographic Atlas. To the extent that modern firms (e.g., corporations, limited

---

This is a real head-thumper kind of realization after the fact." (stated by Sidney Winter in Murmann et al, 2003, p. 28)

liability companies) descend from these earlier organizations, our theory also provides insights into to the nature of modern firms[18].

2.1.1. Our argument and findings

Our model is based on standard theories on cultural evolution (Boyd and Richerson, 1985; Boyd and Richerson, 2005; Henrich, 2015). These ideas, developed in Evolutionary Anthropology, study culture as an evolutionary system (which can also feed back into genetic evolution, hence its tag of "dual-inheritance theory"). The theory specifies micro-founded mechanisms for the inheritance, selection and variation of cultural traits. Culture is defined as information that is acquired from other individuals via social learning mechanisms, such as imitation and teaching from parents (vertical transmission) or non-related peers (horizontal and oblique transmission). Information includes beliefs, norms, knowledge, skills, artefacts and technology. A central insight of this theory is that social learning –a fundamental behavioural trait of the theory– is able to generate cumulative culture. By way of diffusing innovations in a population, social learning can make small innovations to accumulate over time in a society. This has allowed the human species to adapt and conquer, in a very short span of evolutionary time, every environment in the globe.

[18] Based on the accounts of Dari-Mattiacci et al (2017), Guinnane et al (2007) and De la Croix et al (2018), among many others, it is possible to establish guilds, and particularly partnerships, as ancestors of modern firms. Modern firms such as corporations and private limited liabilities companies display incorporation, limited liability, asset shielding, capital lock-in and tradeable shares (either privately or publicly, in stock markets). Guilds were incorporated entities that produced the bulk of goods and services consumed in pre-modern economies; modern companies fulfil that role nowadays. The partnership ancestry of modern firms is much clearer. The limitations of partnerships to lock-in capital in the wake of long distance maritime trade in early 17th century, drove the partnerships' owners to push for the creation of two historical corporations whose charter included asset lock-in, the Dutch India Company and its British counterpart, the East India Company (Dari-Mattiacci et al, 2017). Over time, these corporations slowly added the other complementary features of modern firms such as tradeable shares and limited liability. Corporations had to wait for early- and mid-1800 to be fully established themselves through legislation as a "free access" organizational form, without the need for case-by-case charters. However, their incidence remained low as they were costly in terms of capital and informational requirements; partnerships remained the dominant organizational form for small and medium sized business and new ventures. In the late 19th century and early the 20th century, a new type of modern firm arose, the Private Limited Liability Companies (PLLC). The PLLC combined the best of corporations, such as limited liability and asset lock-in, with the best of partnership, flexibility and low cost. Once established, the PLLC quickly replaced the partnerships as the preferred organizational form for new firms, with a share of 60% to 80% in Europe (Guinnane et al, 2007). In the US, PLLC was introduced in 1980s. Interestingly, over time partnerships had already evolved to partially mimic the PLLC and continued to do so. For example, LLPs (Limited Liability Partnership) include limited liability for the "general" partner, typically the one providing capital. Again, as in other countries, the PLLC has been successful in outplacing the partnership as the preferred organizational vehicle.

Oftentimes, this cultural selection process occurs independent of people's awareness or immediate understanding (Henrich, 2015).

However, social learning is not automatically conducive to adaptive cumulative culture (i.e., a culture that increases fitness). An important class of cultural evolution models show that, although social learning is favoured by selection, it does not directly lead to an increase in Fitness (Rogers, 1988; Boyd and Richerson, 1995). The problem is that, under changing environmental conditions (which call for novel technologies or cultural traits after a change), if part of the society invest in understanding the current state of nature and developing the required technology (i.e., individual learning) and others simply copy someone else in the society at a lower cost (i.e., social learners), then social learners enjoy higher fitness and, over time, expand until their Fitness equalizes to that o§ individual learners. The outcome is a society with culture (i.e., social learning diffuses the technology and the cultural traits) but whose fitness hasn't increased (as compared to the starting condition of only individual learners). Simply put, social learning generates culture, but it is non-adaptive. This result, known as Roger's paradox, has generated an important literature exploring the conditions that makes social learning a source of adaptive cumulative culture (Laland, 2017; Boyd and Richerson, 1995; Boyd et al., 2013). Some of these conditions are payoff biased social learning, selective learning (e.g., using social or individual learning depending on the circumstances), and traits/technologies that are not independent across environmental states (see Boyd et al., 2013 for more details).

In this paper, we show that POs allows societies to overcome Roger's paradox and thus facilitate adaptive cultural accumulation. The exclusiveness of POs allows society to benefit from the improvements in social learning that happen within the POs. We contend that this benefit is the key to understanding the origins and subsequent increase in frequency of POs. Once a "right" PO arises (exclusive and better at social learning), then cultural group selection will exploit its capacity to generate adaptive culture, leading to an increase in its frequency over time. There are several group selection processes at play: first, POs will be imitated by other groups or societies; second, the group or society with POs will attract migration from worse-o§ better societies; third, the group or society with POs produce more offspring and therefore will expand geographically more rapidly; fourth, POs help in generating technological and numerical advantage in armed conflicts with other societies. Note that these selection processes have been shown to select and diffuse

other group-level or society-level traits, such as routines, religion, or political institutions (Henrich, 2015; Turchin, 2016).

To be clear, we do not model the long-run cultural selection processes. We do not model different groups or societies with different institutions, for example with POs of different types, competing against one another over time. In this paper we determine the conditions of POs that make social learning useful as a source of adaptive cultural accumulation. Then we rely on the four mechanisms of the cultural selection processes described above to do the "long-run lifting". This type of explanation is standard in evolution. In biology, an evolutionary explanation consists of clarifying how the trait favours reproduction and survival (e.g., colour of flowers attract bees which then help polonization), and then simply pointing that natural selection will gradually select that trait in the organism's population (e.g., flowers with no colours will be selected out)[19]. Similarly, given that it is well documented that adaptive cumulative culture and group selection are the crucial processes that generated the success and expansion of the human species (Boyd, 2017; Henrich, 2015; Turchin, 2016; Laland, 2017), it only su¢ ces to demonstrate that the trait being explored favours the conditions for these processes to operate (or alternatively, that the trait will be harnessed by these processes) in order to claim that the trait is going to invade the social landscape. In our case, we show that POs directly support adaptive cumulative culture, which in turn, supports group selection.

The modelling approach is as follow. We use is a workhorse cultural evolution model, widely used to explore the conditions that make social learning adaptive (Rogers, 1988, Boyd and Richerson, 1995). On every period, a population of agents adopts an activity (or technology) that confers benefits but whose value is lost if the environment changes. These agents can have one of two strategies: individual learning, where agents always adapt their activity to the environment of the current period with a cost C, or social learning, where the agent bear cost c (lower than C) in order to copy a random agent from the previous period. Consistent with prior literature, we show that, although favoured by

[19] Paleontology and phylogenetic trees complement this by building the "tree of life", that is, the documentation of the branching of the different species over time, including their timing, sequence and differential traits. In cultural evolution, the equivalent is: i) the tracing the evolution of artifacts (or institutions such as kinship structures) across time and place by archeologists, anthropologists and historians, and ii) the study of the branching of the different languages in the human species by linguists.

selection, social learning does not enhance the fitness of the population because social learning expands until its benefit is the same as that of individual learning. As social learners expand, they place a negative externality on the rest of social learners (i.e., social learners become increasingly liable to a change in environment). However, we prove that by placing a portion of the population into a small group that has a lower cost of social learning equal to c̃and whose access is restricted -that is, a PO is introduced-, the fitness of this group, and therefore of the average fitness of the population, increases. As a corollary, POs end up populated exclusively by social learners and thus, POs adapt slowly to environmental changes (an interior equilibrium is reached outside of the firm). We explore several comparative statics in order to explore conditions that make POs more or less beneficial. Among these, we find that "secrecy" –the fact the social learners located outside the PO learn imperfectly from the agents within the PO– decreases the fitness of POs.

Intuitively, POs make social learning adaptive by using two mechanisms. First, POs improve the efficiency and the fidelity of knowledge transmission ($\tilde{c} < c$) , a crucial driver of cumulative culture (Lewis and Laland, 2012). Knowledge transmission is a cooperative act (Laland, 2017; Fogarty et al., 2011) and POs naturally generate conditions that favour cooperation among their members. This comes from many sources, such as population structure and assortment of cooperative types (Nowak, 2016; Rand and Nowak, 2013) or triggering deep seated tribal instincts (Bowles, 2009). The second mechanism is that POs, by way of being exclusive, put a halt to the expansion of social learners, limiting the negative externality they generate. This restriction of access enables that any improvements in the cost of social learning within POs can be adaptive and useful for society. Thus, instead of preventing hazards when agents interacts (as in current theories of POs), POs mitigate the negative externality that the expansion of social learners generates.

An important contribution of our theory is that it provides a novel solution to Roger's paradox. Current models suggest that in order to increase fitness, social learning has to allow individual learning to perform better (Boyd and Richerson, 1995). We are the first to show that social learning can be beneficial to society without resorting to a positive impact on individual learning. Within exclusive POs, improvements in social learning are sufficient. Equally important, our theory also provides an explanation for the origin of specialization within POs, a condition that has been prevalent in these organizations. Contrary to extant explanations based on comparative advantages and economies of scale, we show that even in societies in which trade is absent, specialization

within POs -that is, social learners of a PO share the same activity- will be beneficial to the society. In a world with multiple productive activities, specialization within POs maximizes the social learning advantage of POs, which in turn maximizes the fitness benefits that POs bring to society.

We test our theory using data from the Ethnographic Atlas and the Standard Cross Cultural Sample, as provided by the D-PLACE dataset (Kirby et al., 2016). We measure the presence of technologies in pre-modern societies (e.g., weaving, metal working, pottery) and whether they were executed throughout the society or mainly by a small group of people, that is, within a PO. Using several measures of population and cultural complexity as our dependent variable, we find consistent evidence for the main propositions and comparative statics of our model. The results are robust to alternative explanations and endogeneity corrections, and we successfully test a secondary prediction of the model following Giuliano and Nunn (2017). We see this empirical exercise as a first attempt to bring the theory to data. It grounds the theoretical arguments and provides compelling correlations in line with the predictions.

2.1.2. Other related literature

Our paper is related to other strands in the literature on the nature of firms. Knowledge based accounts of POs have been put forward in the economics and management literature. These theories point at the importance of knowledge as a basis for POs focusing mostly on the integration of different bits of specialized knowledge in order to apply it to current productive challenges (Grant, 1996) or in order to solve novel problems (Nickerson and Zenger, 2004). Also, they explore the benefit that hierarchies provide in matching problems of varied difficulty to agents with different levels of skills (Garicano, 2000), the substitution of knowledge transfer by authority and directives (Demsetz, 1988) or the organization of the firm aims to reduce asymmetric information between the firm and it's clients (Levin and Tadelis, 2015, Poblete, 2015) (for a discussion see Spulber, 2009) . We focus on a more basic and general issue -the transmission of knowledge among individuals. In this sense we are closer to work on organizational learning (Argote and Miron-Spektor, 2011), knowledge replication (Winter and Szulanski, 2001), idiosyncratic culture as the root of performance in POs (Gibbons and Henderson, 2012), and knowledge transmission costs as the source of boundaries in knowledge intensive firms (Espinosa, 2017). Our theory is also explicitly dynamic, formally modelling the transmission of knowledge and its evolutionary process.

An evolutionary take on POs is also part of extant research in economics, sociology and management (Nelson and Winter, 1982; Aldrich (1999); Hannan and Freeman, 1997; Levinthal, 1997). Although these theories successfully inform questions of change and adaptation, they do not address the question of what are the evolutionary and historical origins of POs. Why is it that POs have gone from nonexistence to domination of the modern economic landscape is not explained. By studying how POs impact the process of cultural evolution, we provide a first step into this direction.

Our paper also relates to the literature in economics that explores the long term and historical determinants of economic development (Nunn, 2009; Spolaore and Wacziarg, 2013). Within this body of work we relate more directly to research that studies the origin and evolution of institutions (North, 1991), particularly the branches that high- light the role played by culture (Alesina and Giuliano, 2015; Nunn, 2012; Mokyr, 2016; Tabellini, 2008) and that use a cultural transmission (Bisin and Verdier, 2011) or a cultural evolution approach (Giuliano and Nunn, 2017; Giuliano, 2016; Enke, 2017). Our unique contribution lies in providing a model for the evolution of organizations, complementing the prevailing focus on institutions (e.g., state centralization). Our theoretical arguments on the origins of POs also complements the economic history literature that is tracing empirically the origin of guilds, corporations, and private limited liabilities companies (Dari et al., 2017; De La Croix et al., 2017; Lamoreaux, 1998; Guinnane et al, 2007; Hansmann et al., 2006). We believe that our theory (and its further developments) can inform the historical data.

Our theory also weighs in the long standing debate on the role and function of guilds. While one camp stresses the rent-seeking costs of Guilds based on their monopoly power (Ogilvie, 2014 and 2019), our theory supports the camp that stresses its benefits (Greif et al, 1994; Epstein, 1998 and 2008; Epstein and Praak, 2008; De La Croix et al., 2017), particularly those related to diffusion of knowledge. Guilds are beneficial because they can diffuse knowledge conditional on an exogenous stream of innovation. Our argument indicates that this is flawed in evolutionary terms; in the very long run different strategies –that is, social-learning/imitation and individual-learning/innovation– are endogenous. In that scenario, cheaper (or costlier) social learning (vis a vis individual learning) simply means that society gets more (less) of it, without necessarily benefiting from it.

Finally, our paper also related to the study of culture by organizational scholars (Schein, 2010; Giorgi et al, 2015; Weber and Dacin, 2011; Meek, 1988; O'Reilly and Chatman, 1996). In general, this literature studies the nature of organizational cultures: how to define culture, what are its constitutive elements, how it can be molded over time, how organizational members are influenced by culture as well as exert an influence on it, its relation to organizational structure and practices, among others. However, this literature has different strands; our paper is much closely related to the "structural-functional" perspective on culture (e.g., Schein, 2010; O'Reilly and Chatmant, 1996) with whom it shares the provision of a function for culture (a connection with performance) and a broader definition of culture. In contrast, the "symbolic" perspective defines culture as the repertoire of language, codes, metaphors, and other symbolic elements, without specifying a function that connects it with group performance (Weber and Dacin, 2011; Meek, 1988). Culture differentiates from these two perspectives by having a broader definition of culture ("anything that is transmitted via social learning"), by specifying endogenous origins for culture (via social learning; the other perspectives have exogenous origins such as founders or the external environment); and by specifying in a clear and precise way the connection of culture with fitness and collective payoffs.

2.2. Model and Predictions

We build on a workhorse model in the cultural evolution literature (Rogers, 1988; Boyd and Richerson, 1995). A mathematical proof of all the results is presented in the appendix. Models in this tradition use evolutionary game theory. The basic buildings blocks of this modelling approach are the following (see McElreath and Boyd, 2007 for an introduction): i) there is a large population of agents, ii) each agent has a strategy that is fixed (i.e., players are "dumb" and don't display strategic behaviour) (instead of strategy, the word "cultural trait" or "cultural variant" is frequently used), iii) there is a fitness function that relates the strategy to fitness and that depends on the relative frequency of strategies in the population and environment parameters (such as uncertainty), iv) fitness is an abstract measure of adaptation to the environment (fitness in cultural evolution is frequently labelled payoffs), v) after each period a small proportion of agents copies the strategy with highest fitness (in the case the periods occur within a generation, for example 5 years); this is known as "replicator dynamics" and allows the share of strategies in the population to change over time (in the case of biological evolution, fitness is equal to offspring and replicator dynamics is simply the expression that relates differential

fitness of different genotypes –the analog of strategies– to the change in their proportions in the population) (cultural evolution can also occur across generations with parents inheriting their strategies or traits to their children), vi) the solution concept that is utilized to solve the game is "evolutionary stable strategies"; this is the idea that there is one or more equilibriums in term of shares of strategies that, if disturbed, for example by the entry of a small share of one of the strategies, the system will return to the equilibrium (this equilibrium is the same as Nash equilibrium if rationality and foresight is assumed instead), vii) in the case of more than one equilibrium, there are threshold points that separate the different the basins of attractions of each equilibrium; these basins of attraction can be of different sizes (when an evolutionary game is stochastic, these basins are important).

On every period, a continuum of long-lived agents adopt a technology that confers fitness, but whose value is subject to changes in the environment. There are *N* environmental states. In each period, there is a probability that the state may change with probability *p*. For every state, there is a unique technology that provides fitness. By a normalization, we can assume without loss of generality that the fitness of a technology is 0 unless it is tuned to match the state. Agents adopt a technology by using one of two behavioural strategies. An Individual Learner studies and understands her environment and is able to develop each period a new technology tuned to the current state. This strategy has a cost *C*, which is bounded between 0 and 1. The second alternative is to learn socially. A Social Learner looks at what some other randomly chosen member of the population did on the previous period, and simply copies its technology incurring in a cost $c < C$. This strategy is less costly because the agent does not need to understand the underlying state, but simply copy what others do. In order for social learning to survive we assume $(C - c) > p$.[20]

Let r_I be the share of individual learners and r_S the share of social learners in the population and let q be the percentage of people with a tuned technology in the population. For any given pair of shares r_I and r_S the expected ratio of tuned agents is given by $q^e(r_I, r_S) = \frac{r_I}{1-(1-p)r_S}$.[21] The

[20] Given that we model traits that can be transmitted among individuals, our theory is better suited for POs that display this behaviour, such as guilds and partnerships. Modern corporations add to the mix the integration of different knowledge sets, which we do not formally model.

[21] This is achieved by noting that the percentage q is governed by the differences equation $q(t) = r_I + q(t+1) \cdot (1-p) \cdot r_S$. The expected value is calculated by computing the steady state.

fitness of an individual learner is $f_I = (1 - C) > 0$, because she is always tuned to the state bearing the cost C. The fitness of a social learner is $f_S = (1 - p)q^e - c$. Social learners, given that they copy their behaviour from others, sacrifice tuning if the state of the world changes or if inadvertently they copy from an untuned member. The fitness of social learners is increasing in the share of individual learners, because individual learners increases the chances of copying from tuned individuals.

We assume every period a small proportion of agents adopt the strategy of other agents with higher fitness levels. This type of evolution dynamics is known as quasi birth and death process and converges to evolutionary stable strategies (ESS).[22] Formally, a population with shares (r_I, r_S) plays an ESS if a small group of invaders using any alternative strategy achieves a strictly lower average fitness. Consistent with prior literature, we find that there exists a unique equilibrium where both strategies are present with shares that depend on C, c, and p,[23] and the fitness of both types is $1-C$. (See the appendix for a formal proof.) Intuitively, in equilibrium there cannot be only individual learners because in that case the ratio of tuned population q^e approaches one and the fitness of social learners would be higher than that of individual learners. In the same way, there cannot exist only social learners because then the ratio of tuned population q^e approaches zero, and the fitness of individual learners becomes larger than that of social learners.

Observe that because both behavioural types in equilibrium achieve the same fitness $1 - C$, this implies that society as a whole does not benefit from the existence of social learning, because the same average level of fitness can be achieved with individual learning only. The fact that social learning gets selected but does not affect the fit- ness of the population is known as Roger's paradox (Rogers, 1988). This result has demonstrated to be robust to different specifications and assumptions, leading Boyd and Richerson to state that "to improve the average fitness of the population, imitation must make individual learning cheaper or more accurate" (Boyd and Richerson, 2005; p. 39). In what

[22] We could also assume that agents are short lived and their reproduction rate depends on their fitness level. In either case, the equilibrium concept is ESS and our results are the same.

[23] The share of individual learners is given by $r_I = \frac{p \cdot [1-\{C-c)]}{(1-p) \cdot (C-c)}$. Thus, individual learning increases with uncertainty and decreases with the cost advantage of social learning.

follows, we show that adding a PO in the society solves this paradox in a way that, we argue, is fundamentally different from other solutions proposed in the literature, because it does not require improving the fitness of individual learning.

2.2.1. Productive organizations

We now introduce productive organizations in the model. Two characteristics describe a productive organization (PO). First, access to the PO is limited. A λ fraction of agents is located inside the organization and this fraction is fixed. This means that even though additional members might want to be a part of the PO, membership is limited by the value of λ. This does not mean that λ cannot shift or evolve; we address this issue in detail in the section 3.3.

This characteristic mirrors the condition that POs have had across history: roman societas, medieval guilds, renaissance partnerships and modern corporations, all limit the access of the population to become a member. Exclusiveness appears to be present from the first records of POs. For example, Apel (2008) describes the production daggers in Scandinavian society in the late Neolithic and explains that "the production is consciously organized to keep the recipes of the technology exclusive to certain segments of the society". More general reviews confirm that exclusivity is a salient characteristic of the first non-kin goal-oriented organizations (sodalities). For example, in the revision by Anderson (1971) of Lowie (1948), he states that in his description of early sodalities, "he could find no common characteristics beyond the fact that they all excluded non-members."

Second, agents that belong to the same PO, can learn from each other at a lower cost. When a social learner adopts its technology from another member of the PO it has a cost $\tilde{c} < c$ (if she copies outside, she bears a cost c). This entails that for a specific information to be transmitted, lower effort would be required; or alternatively, for a given amount of effort the fidelity of the information transmission is higher. Theoretically, this assumption for POs is sustained in the fact that population structure favours cooperative behaviour through assortment of cooperatives types or higher frequency of interactions (Nowak, 2016; Rand and Nowak, 2013). This, in turn, favours the emergence of teaching or mentoring which in essence is a cooperative act (Laland, 2017; Fogarty et al., 2011; Dean et al., 2012). For example, when social learners imitate technologies inside the organization, they can be favoured by the active transmission from the subject they are attending to and who might otherwise be passive. More generally, this assumption can also be

sustained by pointing at the deep-seated tribal tendencies of humans beings which make them prone to identify with their group and to trust and help fellow members (Bowles, 2009).

Empirical evidence supports cheaper social learning within POs. There is considerable evidence in the management and economics literature showing that learning from others is more efficacious when learning from other co-workers of the organization, as opposed from the outside (Argote and Miron-Spektor, 2011). Research on guilds and partnerships documents their role in improving knowledge transmission between its members (De La Croix et al., 2017; Morrison and Wilhelm, 2008). In archaeology, Coto-Sarmiento et al (2018) provides compelling evidence from three centuries of amphorae production in workshops in the Roman Empire. Their analysis suggests that the variability of amphorae between workshops is mostly consistent with a process of high-fidelity social learning within workshops (i.e., master to disciples) instead of horizontal transmission or mobility between workshops.

As before, let r_I be the share of individual learners outside the firm and $\widetilde{r_I}$ be the share of individual learners inside the PO. In the same fashion define r_I, $\widetilde{r_I}$, q, $\widetilde{q}$, q^e and $\tilde{q}^e$. The fitness of an individual learner is the same outside or inside the PO, $\widetilde{f_I} = f_I = (1 - C) > 0$. The fitness of a social learner outside the PO is,

$$f_S = (1 - p)[(1 - \lambda) \cdot q^e + \lambda \cdot \tilde{q}^e] - c \qquad (1)$$

while the social learner inside the PO enjoys a fitness of,

$$\widetilde{f_S} = (1 - p)[(1 - \lambda) \cdot q^e + \lambda \cdot \tilde{q}^e] - [(1 - \lambda)c + \lambda\widetilde{c}] = f_S + \lambda(c - \widetilde{c}) \qquad (2)$$

In equilibrium (or equivalently, in the long run due to evolution) the expected fitness of both behavioural strategies outside the firm equals. Provided there exist at least one social learner outside the PO (something that happens if the PO is not too big), there exists a unique equilibrium in which the average fitness of society is larger than that of individual learners. The result is stated formally in the next proposition.

Proposition 1: If λ is sufficiently small the existence of the PO increases average fitness in the population.

To understand the intuition behind this proposition it is useful to compare the model with POs to the basic model of the previous section. In the basic model, social learners reproduce and grow,

lowering the average fitness of the population q until the fitness decreases to the level of individual learners $1 - C$. With the introduction of a PO of limited size, this negative externality is put to a halt before all the benefits are diluted away, thus allowing society to benefit from social learning (figure 1A). A corollary of this result is that, inside the PO there are only social learners, as any equilibrium with individual learners inside the PO would be invaded by social learners (figure 1D).

The PO has several other interesting effects. First notice that as the PO is populated only by social learners, they bear a relatively larger risk of environmental change and thus the share of fitted population inside the firm $\tilde{q}$ is lower than the share of fitted population outside the firm q. This makes POs slower to adapt, and the mere existence of PO's decrease the average level of tuning in the population ($\lambda \cdot \tilde{q}^e + (1 - \lambda) \cdot q^e$).

To understand how PO's affects the ESS in the population, observe that as the PO change size (λ), two effects take place. On the one hand, social learners inside become relatively fitter, as the difference between inside and outside social learners is given by $\lambda(c - \tilde{c})$ (see equation 2). On the other hand, the PO reduces average tuning in society ($\lambda \cdot \tilde{q}^e + (1 - \lambda) \cdot q^e$) reducing the fitness of all social learners (both inside and outside). Eventually the second effect dominates making the benefit of POs to reduce with size (figure 1B).

Another mechanism at play is that the PO increases the share of individual learning outside the PO (figure 1C). The PO makes the social learners outside worse off because they now are "forced" to copy members of the PO whom, like themselves, are liable to environmental change. This generates an increase in the number of individual learners outside, benefiting the PO. As λ grows, social learners inside the PO gradually substitute social learners outside.

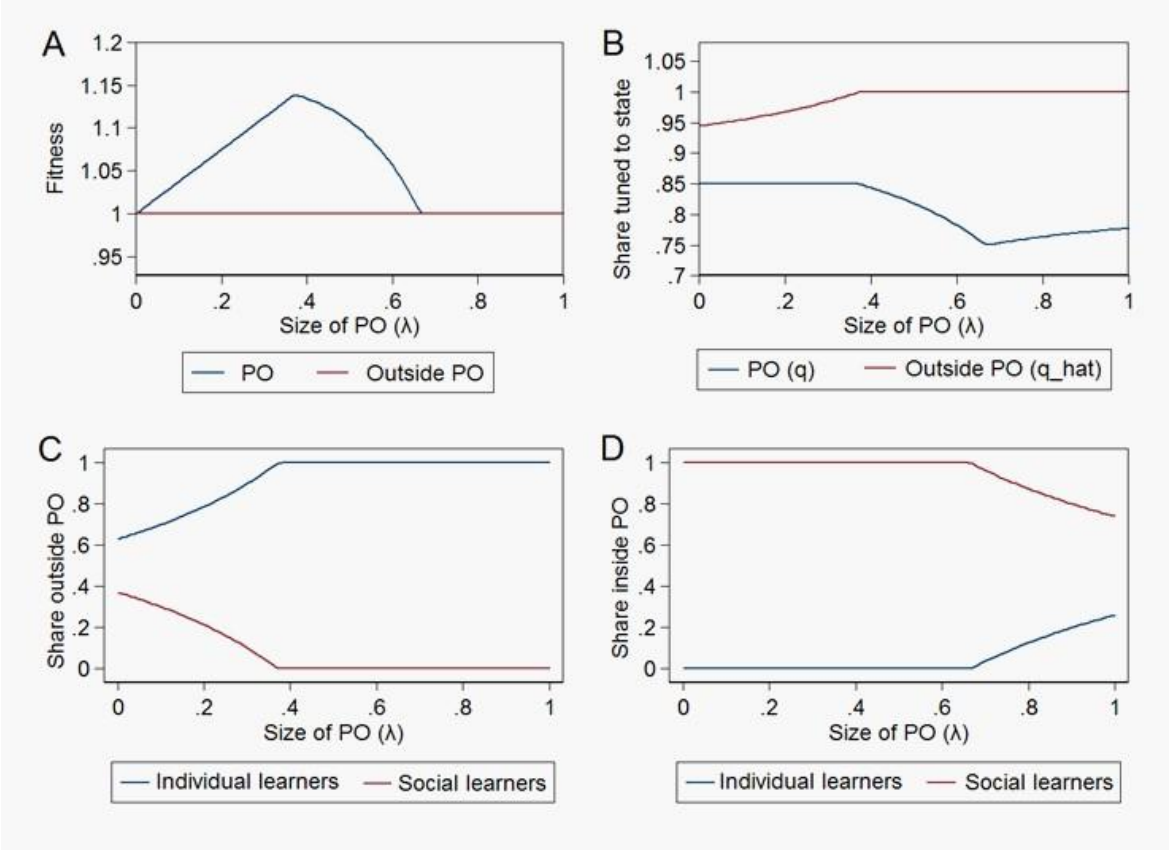
These results suggest that early POs -hunting bands, roman societas, medieval guilds, and renaissance partnerships- probably had a key historical role in promoting hard- to-propel cumulative culture by way of making social learning cheaper and exclusive. Evolutionary benefits of organizations are typically justified by multilevel selection (Turchin, 2016), where stable organizational heterogeneity plus competition leads to cultural progress. Our model points to a more basic advantage: through cooperation, identity and trust, exclusive POs facilitate social learning making culture useful for society. This account provides a clear evolutionary origin to PO's, one that puts knowledge and cooperation at centre stage. This contrasts to extant theories of

PO's, most of which are a-historical and focus on incentives and effort, assuming that a cultural tradition or pool of knowledge is already in place.

Our results also point at the issue of the locus of innovation. In our model, innovation occurs outside the PO. POs may still generate innovations that decrease the cost of social learning (e.g., unique language, enforcement devices, adjusting the technology for improved replicability), but they do not generate the radical innovations needed to track the state of nature. Over history, radical innovations have tended to happen outside POs, for example, in academia or by inventors and entrepreneurs (which then might set up a PO to exploit the innovation).

Finally, proposition 1 is robust to PO-biased social learning, that is, social learning inside (outside) the PO is preferentially executed inside (outside) the PO (i.e., the likelihood of social learning inside the firm is larger for members than for non-members of the organization).

Figure 1. Equilibrium values of the model for different values of lamda. We use $C = 0.6$, $c = 0.45$, $\tilde{c} = 0.3$, and $p = 0.1$. (A) Fitness inside and outside organization we multiply fitness by 2.5 to obtain fitness equal or superior to 1). (B) Percentage of the population that has a technology tuned to the state. (C) Share of social and individual learners outside the organization. (D) Share of social and individual learners inside the organization.



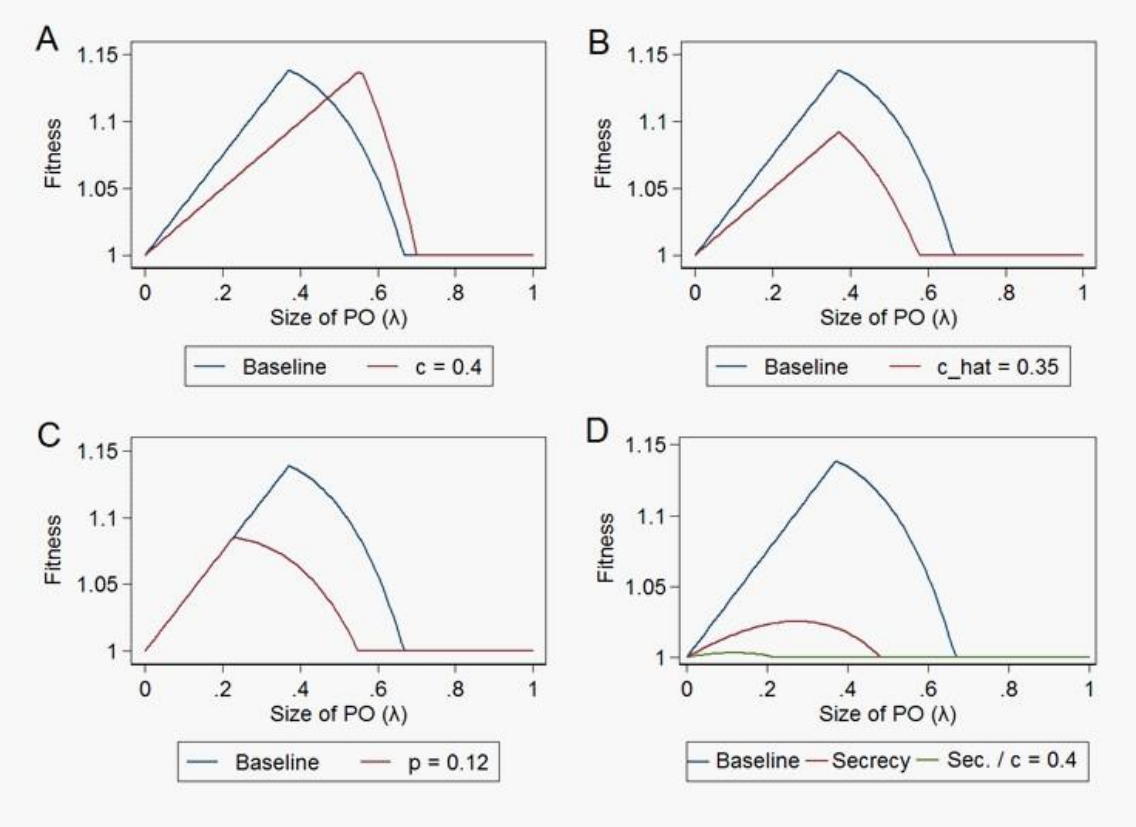
2.2.2. Comparative statics

The model presents several interesting comparative statics which are depicted in figure 2. Confirming the intuition that the fitness of the PO comes from their ability to facilitate social learning, we found that the fitness of the PO depends negatively on the cost of social learning inside the organization c (figure 2B). Empirically, this means that improvements in the ability of executing social learning inside organizations will impact positively their fitness. There is a large literature in organizational learning that provides evidence for this (Argote and Miron-Spektor, 2011). Less intuitively, we find that changes in the cost of social learning outside the PO (parameter c) doesn't translate to a monotonic change in the PO's fitness; instead, a decrease in c decreases the fitness of a small PO, but increases the fitness of the PO if it is sufficiently large (figure 2A).

Another interesting prediction is that increases in uncertainty generate a decrease in the fitness of POs (figure 2C). Given that POs are populated entirely by social learning, it is easy to see that as the parameter p increases, then the PO will be increasingly liable to a change in the environment, reducing its expected fitness.

This result is opposite to what is predicted by theories of PO based on transactional cost economics (TCE), where the value created by the firm is increasing in the uncertainty of the environment (Williamson, 1991; Tadelis and Williamson, 2013). Given this disagreement, it is interesting to assess available evidence. Despite well documented cases where uncertainty favours the use of hierarchies (Forbes and Lederman, 2009), two meta-analyses of TCE show that uncertainty is slightly related with a decrease in the use of hierarchies in favour the use of markets and hybrids, a sign of lower fitness of POs under higher uncertainty (Geyskens et al., 2006; Crook et al., 2013). The literature on industry and product life-cycles provides consistent evidence as well: the size of firms is smaller at earlier stages of the cycle, when uncertainty is higher and product or technological standards are not yet defined (Klepper, 1996). We expand on the comparison with TCE on the empirical section below.

Figure 2. Comparative statics. For the baseline case we set We use $C = 0.6$, $c = 0.45$, $\tilde{c} = 0.3$, and $p = 0.1$ (as in figure 1, fitness is multiplied by 2.5). (A) We set c = 0.4. (B) We set $\tilde{c} = 0.35$. (C) We set p = 0.12. (D) In the red line, we make the PO secretive, that is, social learners outside cannot imitate members of the PO. In addition, in the green line we reduce the value of social learning outside of the PO to $c = 0.4$.



2.3. Extensions

In this section we discuss four extensions to our basic setup. First we study the impact of secrecy in the fitness of POs. Historically many productive organizations have been reluctant to share their knowledge with people outside the organization, for example, guilds were protective of their knowledge and techniques (Ogilvie, 2014). We show that secrecy is detrimental to the fitness of POs, but for a different reason that those highlighted in the literature on guilds (Ogilvie, 2014). Second, we extend our basic model by allowing for several competing technologies. We show that in the presence of multiple technologies, the PO specialize so that on equilibrium all members of the PO adopt the same technology. Further, if many POs are allowed, then they specialize on different technologies. This result provides a novel explanation for specialization within and across organizations. Third, we briefly discuss the endogeneity of the size of the firm λ, and its impact in our analysis. Finally, by adding the idea of know-how and incremental innovation we provide a

broader interpretation of our model. This interpretation allows to expand the positive that POs bring to society.

2.3.1. Secrecy in productive organizations

In this subsection, we study the impact of secrecy on the fitness of POs. In terms of the model this means that social learners outside the PO can't imitate members of the PO. Consequently, outside the PO we are back to the original case without PO where the fitness of social learners outside the PO is

$$f_S = (1-p)q^e - c\ , \tag{3}$$

independent of the size of the PO. The fitness of social learners inside the PO is as before, given by

$$\widetilde{f_S} = (1-p)[(1-\lambda)\cdot q^e + \lambda\cdot\ \tilde{q}^e] - [(1-\lambda)c + \lambda\tilde{c}]. \tag{4}$$

In this case, it is not straightforward whether f_S or $\widetilde{f_S}$ is higher. This is so because although inside the PO individuals benefit from cheaper social learning ($c < \tilde{c}$), individuals outside the PO benefit from copying relatively more tuned population ($q^e > \tilde{q}^e$). In the case without secrecy, the PO imposes a negative externality on social learners outside the PO, so that the share of individual learners increases with λ (and social learners inside the PO gradually substitute the social learners outside the PO) (see figure 1C). In the case with secrecy, this negative externality is not present, and thus social learners outside the PO copy more tuned agents on average (as compared to social learners inside the PO).

It is relatively straightforward to show that in this setting, proposition 1 still holds, that is, provided the PO is sufficiently small, it increases the fitness of the population. However, society benefits less from secretive POs and in order to be productive, their size needs to be smaller (see figure 2D). This is stated formally in the next proposition.

Proposition 2: The positive impact of the PO on the fitness of the population decreases if the PO is secretive.

Intuitively, the benefit of POs are now smaller because under secrecy outsiders are not allowed to learn from the organization's members. Therefore to make sure that in equilibrium all members outside the PO have a fitness of $1 - C$, the share of social learners outside the PO increases, which

in turn implies a negative externality for the PO, because there are less individual learners to learn from. In other words, secrecy creates a relative disadvantage to social learners inside the PO because on average they learn from a less tuned pool than social learners outside the PO.

This also forces the size of POs to be smaller, because this disadvantage is smaller when λ is small because then level of tuning inside the PO $\tilde{q}$ is closer to the level of tuning outside q (or alternatively, the relative tuning advantage of social learners outside the PO is lower).

Notice that, just like the case of uncertainty discussed above, the prediction of lower fitness and size of PO under secrecy runs opposite to what TCE would indicate (Williamson, 1991). Secrecy, as a symptom of weak appropriability regimes and threats of leakage, would be correlated in TCE with higher benefits of using POs, moving transactions away from market and increasing the size of POs. In the empirical section we show which of these opposing predictions fares better in our data.

The finding that secretiveness decrease fitness seems to be at odds with the fact that historically several POs have been secretive. For example, guilds tended to be protective of their knowledge and techniques (Ogilvie, 2014). This suggest that, in order to evolve, secrecy in POs must have provided an additional benefit. We speculate that secrecy might generate a lower learning cost $\tilde{c}$ because secrecy can galvanize the notion of "us vs. them" that promotes within group identity, trust and cooperation (Bowles, 2009). It is also possible that secretive organization evolved first, with the benefits of non-secretiveness being discovered latter, perhaps in current times.

An interesting finding is that secrecy is more detrimental (compared to non-secrecy) when the cost of social learning outside the PO is reduced (figure 2D, green line). This result can explain the trend towards more transparent and open organizations that have occurred in the last thirty years (e.g., Wikipedia). This trend has largely been a result of better communication technologies that generate a higher reduction of social learning costs outside POs (Benkler, 2006).

2.3.2. Specialization

We now add multiple technologies to our model. Suppose there are M technologies and as before, nature can change with probability p, rendering the technologies less useful. For example, a society living in an island with dry climate, and thus non-fertile soil, would exploit a myriad of sea technologies, such as net fishing, spearfishing, angling, shellfish harvesting, boat building, among others. If the environment changes, then the optimal way of executing each of these activities would

change, making the techniques less beneficial. For example, a migration of larger fish might render current nets too weak, requiring stronger nets for effective net fishing.

Agents, like before, can be either social learners or individual learners but now their behavioural strategy also specifies an activity ($j \in J$) whose share in the population will depend on the replication dynamics. All activities are assumed ex-ante equally productive, with fitness levels equal to the previous section. The behavioural strategies and proportion of tuned agents in the population now have a superindex $j \in J$ that specify activity. Social learning occurs randomly but is restricted to the set of people that executed the same technology j in the previous period. This requires the assumption that a social learner can identify people that executes the same activity that they do. Following our example, if a social learner uses net fishing, she will be copying agents in the population that executed net fishing in the previous period, some of which will be tuned -using the optimal net fishing technique for the current environment- and some of which will be not tuned -using a less optimal but still identifiable net fishing technique.

Replication dynamics in this model drives the fitness among social and individual learners and among all surviving activities to be the same. Because in this model there could be multiple (but qualitatively identical) equilibria, in order to make comparative statics we focus our attention on the symmetric equilibria where individual learners are uniformly distributed among activities. (This could be the result, for example, of uniform natural preferences of individual learners, or some decreasing returns to scale in each activity).

The results of this model is straightforward, and mirrors the case with only one technology. Social learners are selected into the population but the overall fitness of the society does not increase. Given our assumptions, individual and social learners will be distributed evenly across technologies. The total share of social learners will decrease with c and p.

We now allow for the existence of one productive organization of size λ. A social learner with activity j will now bear a cost c when imitating agents on activity j whom are sharing her PO, and cost c when learning from agents with activity j located outside her PO. Thus, inside the PO

social learners only copy agents that have their same technology. This captures the idea that is more difficult to learn across than within technologies[24].

To show our main result, define $x^j = r_S^j + r_I^j$ as the share of agents that execute activity j outside the PO, and equivalently define $\tilde{x}^j = \tilde{r}_S^j + \tilde{r}_I^j$ as the share of people outside the PO that executes activity j. The fitness of the social learner outside the PO is given by

$$f_S^j = (1-p)\left[\frac{\lambda\,\tilde{x}^j}{\lambda\,\tilde{x}^j + (1-\lambda)x^j}\,\tilde{q}^e + \frac{(1-\lambda)x^j}{\lambda\,\tilde{x}^j + (1-\lambda)x^j}\,q^e\right] - c. \tag{5}$$

Inside the PO the fitness can be written as follows

$$\tilde{f}_S^j = f_S^j + (c-\tilde{c})\frac{\lambda\,\tilde{x}^j}{\lambda\,\tilde{x}^j + (1-\lambda)x^j}. \tag{6}$$

Observe that the advantage in fitness of social learners within the PO is increasing in the share of members in the PO that execute the same activity. Therefore, the evolutionary dynamics will drive social learners inside the PO to specialize in the same activity. This is stated formally in the next proposition.

Proposition 3: Given a sufficiently small λ, the PO specializes in a specific technology.

The intuition for this result is that the advantages of lower social learning costs increase when you have a larger group that can learn from each other and reap the benefit of cheaper social learning. The total costs of social learning will be reduced within the PO if everybody specialized in the same activity. If there are several PO's in the society, each PO will specialize in a different activity.

[24] The results of this section also hold if we assume that social learners learn disproportionately from within the PO.

The standard explanation for specialized productive organizations is that they al- low to take advantage of economies of scale and comparative advantage by trading with other specialized POs. Our results suggest an alternative and likely complementary explanation. Even in societies that haven't discovered trade (both internal or external) and where comparative advantages are absent (i.e., all agents bear the same opportunity cost of doing any technology, or in our model, a constant c), specialization within productive organization will be beneficial. This is so because it maximizes the social learning advantage of POs which in turn maximizes the fitness benefits that POs (of limited size) bring to society. In the empirical analysis, we provide evidence that specialized POs are beneficial even in the absence of trade.

It is interesting to point out that in standard economic theory, the benefits from firms driven by specialization and trade are increasing in the environmental uncertainty (Burgess and Donaldson, 2010), while the benefits from social learning we discuss in this paper are decreasing in environmental uncertainty (see figure 2C). Thus, it is possible to empirically test whether the benefit of having specialized POs is mostly obtained by social learning or trade by analysing the impact of POs on fitness at different levels of environmental uncertainty. We discuss this point with more detail in the empirical section.

2.3.3. The size of the firm

So far we have taken the size of the firm as exogenous. Although it is not difficult to endogenize size, we choose not to do it, in order to keep the message of the paper as simple as possible. Below we provide a brief discussion regarding the main issues surrounding the size of POs.

First, observe that regardless of the size of the PO, people will always weakly prefer to belong to a PO. Therefore, to endogenize the size it is sufficient to assume people will be admitted as long as doing it increases the fitness of the PO members. The model clarifies a crucial trade-off in this process: a small PO doesn't take full advantage of cheaper social learning but a large PO loses track of environmental changes. Given that our model is continuous it follows that there exists an interior size that maximizes the average fitness of its members (see figure 1A). Several mechanisms of modifying the number of members, will make the size of the PO to converge (via trial and error) to the size that maximizes the members' average fitness. These mechanisms can be purposeful (e.g., voting in a partnership) or exogenous (e.g., a deadly disease). In both cases,

especially the latter, it is necessary to assume that PO members track the resulting fitness of a change and update λ as a result. It is important to notice as well that the PO, as it changes its size, retains its exclusivity condition. Endogenizing size means adjusting the size while keeping access restricted; otherwise, the size of the firm would be expanded until its benefits are completely diluted.

Second, given the continuity of our environment it is possible to show that the size of the PO that maximizes fitness (which is the natural candidate for a model with endogenous ,\) is always small from a social perspective. To see this let $\widetilde{f}(\lambda) + (1 - C)$ be the fitness of agents inside the P O At the level that maximizes fitness it must be that $\widetilde{f}'(\lambda) = 0$. Overall fitness in the population is given by $1 - C + \lambda \widetilde{f}(\lambda)$ therefore the derivative with respect to the size λ at the point that maximizes member's fitness is $\widetilde{f}'(\lambda) > 0$, meaning that the society will benefit from increasing the size of the PO further.

Third, we deem likely that λ and $\widetilde{c}$ are connected. If POs can generate a lower cost of social learning through group identity, trust and cooperation, then it follows that a larger λ will probably increase $\widetilde{c}$, which in turn would reduce the size of the PO.

2.3.4. Extended interpretation

The interpretation of the model can be extended by acknowledging that in order to perform an activity (or apply a technology) both learning and production are required. One has to learn not only what to do, for example learn that net fishing generates a good catch in the current environment, but also develop an understanding and capacity of how to do it, for example, know which are the best raw materials to produce the net, how to assembly it properly, how to use it best, when and where it provides more benefits, and so on. An individual learner has to figure out these two challenges, learning what to do and then refine how to do it; similarly, a social learner has to copy both elements and then execute properly.

By adding the production stage, we can introduce the idea of incremental innovation, defined as the process through which the execution of an activity can be improved bit by bit over time. Evidence from the organizational learning literature indicates that these accumulated improvements can be substantial (Argote and Miron-Spektor, 2011; Levitt, List and Syverson, 2013). Incremental innovation contrasts with the innovation that is executed solely by individual

learners, which can be equated to radical or disruptive innovation: they spend a lot of effort to understand the new state of nature, and then generate a novel and well-adapted technology. In addition, incremental innovations are not protected from a change in environment: no matter how refined the technique has become, the environment can change in favour of a different technology.

By allowing agents to engage both in learning and production, the parameters C, c and $\tilde{c}$ would now include not only learning costs but also the costs of production. The difference between C and c would also capture any advantages in production costs by social learners. If we assume that in every instance of social learning the learner can introduce incremental improvements on execution (which can then be passed down to other social learners), then the aggregate cost advantage generated by social learners can be very large.

Within this framework, POs can further generate benefits to society by providing an increased capacity to generate incremental innovations (which would be reflected in a larger difference between c and $\tilde{c}$). This capacity can be positive and substantial if incremental innovation is complementary with the effectiveness of social learning. This is not unlikely, as it is easier to improve upon a well-replicated technology. This seems particularly true under specialization of POs. Although a full fledge development of this idea is beyond the scope of this paper, this interpretation allows for a larger impact of POs: any improvement in incremental innovation they generate would be beneficial for society.

2.4. Empirical analysis

Our empirical analysis has five parts. First, we introduce the Ethnographic Atlas data and the variables measurement, and we show that the fitness of pre-modern societies is correlated with the presence of technologies only when the technologies are performed within POs. Second, we make the argument that there iscausal relation between POs and fitness by using instrumental variables and by using bounds that show that our results are not likely to the reverted by omitted variables. Third, we do a comparative statics analysis to see how the benefit of POs depend on uncertainty and the cost of social learning. Here, we find that the results favour our theory of POs over standard TCE reasoning. Fourth, we execute several robustness checks that allow us to show consistency across different dependent variables and rule out alternative explanations, such as the presence of trade. Finally, we replicate the results of Giuliano and Nunn (2017) in order to do an out-of-sample

test of a separate prediction of the model: we show that use of POs in ancestors increases the reliance and persistence of tradition in modern populations.

The combination of these five empirical exercises provides compelling evidence that POs played an important role in making social learning adaptive as our theory predicts. Still, as indicated in the introduction, an important part of our empirical exercise is to provide a "proof of concept", that is, as a way to verify that the theory has practical potential and that it can be empirically productive.

2.4.1. Data

To test the predictions of the model, we use the Ethnographic Atlas (EA) and the Standard Cross Cultural Sample (SCCS) provided by the D-Place dataset (Kirby et al., 2016). The EA describes cultural practices for 1291 pre-modern societies, ranging from societies with complex agricultural economies and political systems to small hunter- gatherer groups. The societies are globally distributed with especially good coverage of Africa and western North America (see figure 3). The SCCS is a subsample of the EA where additional information about societies is provided. We use the SCCS to measure several variables that are needed to test the predictions of the model. These datasets were created by coding the available information about societies that is present in the extensive ethnographic accounts in the anthropology literature.

The EA provides information on the presence of eleven productive activities (or technologies) in the society: metalworking, pottery making, weaving, leather working, hunting, boat building, house construction, gathering, agriculture, fishing, and animal husbandry. The dataset identifies whether each activity was present in the society, and if so, whether it was "normally performed by many or most adult men, women, or both" or whether it was "largely performed by a small minority who possess specialized skills". Following our theoretical model, we identify the second condition as the addition of a PO to the execution of a specific activity. We measure PO in this

way because the types of “small minorities" covered by the EA fit the requirements of our theoretical model.[25]

The minorities in the EA can be of four types: senior age specialization (i.e., only men or women beyond the prime of their life), junior age specialization (i.e., only boys or girls before the age of puberty), craft specialization (which includes occupational castes where the rights to execute certain activity were inherited), and industrial specialization (i.e., specialization is removed from age or craft specialization and is executed using industrialized techniques). Aggregating across activities, craft specialization covers roughly 85 percent of the societies that have a minority executing the activity, industrial specialization accounts for eight percent (aprox.) and senior/junior specialization split the rest.

Our model requires that PO possess three characteristics in order to benefit society: easier social learning, small size and exclusivity. In the discussion that follows, the lower costs of social learning of the PO of the EA become evident; therefore, we do not expand on it. Regarding small size, the very definition in the EA specifies a “small minority". Exclusivity requires more care to be mapped to the EA. Industrial specialization and senior/junior specialization comply with the exclusivity criteria. In the former, exclusivity is predicated on employment, and in the latter, it is defined by age. To understand exclusivity in craft specialization, we sampled the ethnographies in the EA. Roughly, there are three types of craft specialization. First, the original ethnographies describe organizations that could be described as “proto-guilds" --the most common type of craft specialization in the EA. Similar to medieval guilds, these organizations had experts, sometimes called “masters", and apprentices, which would come together regularly --or seasonally, in the case of fishing at high latitudes-- in order to exchange work for teaching and to learn from each other. Apprentices typically needed to prove their capacity in order to fully access the community of experts, so access was not freely granted. Being a master often carried prestige in the society. Frequently, the right to execute a particular craft/activity was hereditary (e.g., the Chekiang society in China for fishing), generating occupational castes (or a specialized clan). This

[25] In addition to providing a proxy of POs, the EA allows us to align the empirical test with our evolutionary argument. We posit that the first POs evolved, in some ancestral time, because of their ability to foster the accumulation of culture. This argument about “origins" requires evidence coming from pre-modern societies.

hereditary element compounded exclusivity because, even in this case, skill was also a frequent pre-requisite to enter the "proto-guild". The second type of "specialized minority" were the small and scattered workshops, where one or more skilful specialists, with the help of a handful of workers, would serve the needs of a portion of the society, typically the local town or region (for example, metalworking in the Rifian culture in Northern Africa). The third type is "attached specialists" where skilled crafstmen were appointed and funded by the rulers of the society (e.g., metalworking in the Inca Empire). Either by the selection of workers or their funding, the second and third types also seem to ensure exclusivity. All considered, even though there is heterogeneity in the "craft specialization" of the EA, the basic idea exclusivity in these organizations seems to hold ground.[26]

This characterization of pre-modern craft specialization in the EA --and our use of it as a proxy for PO-- is consistent with the broader archaeological literature. For example, in its review of the evidence Sterelny (2012) indicates that "craft expertise --the kind of skill sets that forager lives depend on-- is fine-tuned at a generation and reliably transmitted across generations by this mode of organized human learning environments" (p. 35; emphasis added). When discussing the prominent example in Stout (2002), an ethnographic study of stone adze making in Irian Jaya, Sterelny indicates: "The social and informational organization of adze making is strikingly akin to a medieval guild. The apprenticeship system is quite formal. There is a master adze maker who has at least formal authority over the distribution of raw material to adze makers. Apprentices have to be accepted by a recognized master, and while apprentices are typically close relatives of their master, that is not sufficient. [....] The parallels with the formal, institutionalized system of apprentice guilds could hardly be clearer." (p. 40-41). In metalworking, this guild like structure is prevalent in the literature (Rowlands, 1971).

There is a final correspondence between the POs of our model and the measure we use. In section 2.3.2 we indicate that our model can accommodate multiple POs which, in the presence of

[26] However, in roughly a third or a quarter of the cases we sampled, the EA doesn't provide indication about the presence of a (exclusive) group of specialists; instead, it might indicate, for example, that in each clan there was an specialist in a particular craft without specifying whether these specialists would come together as a group (or which were the rules of entry). This is referred to in the literature as "home production" (Costin, 2001). We believe that this might introduce a downward bias in our empirical analysis.

multiple technologies, would specialize in a particular technology. This is reflected our measure: the EA identifies minorities that are specialized in a particular technology.

2.4.2. Baseline model

We computed two variables: the percentage of activities that are present in the society ("% presence") and the percentage of those activities that are executed within a PO ("% within PO"). In the dataset, there is missing information on the activities due, for example, to the fact that the ethnography did not study productive activity. Only 263 societies had complete information on the eleven activities. The variable "% presence" is computed as the division of the count of activities that were present in the society over the count of activities for which we had available information. The variable "% within PO" is computed as the division of the count of activities "largely performed by a small minority who possess specialized skills" over the count of activities that were present in the society. The relationship between "% presence" and "% within PO" is positive, with a correlation coefficient of 0.4 (see figure 4A). The variable "% presence" captures, partially, the cultural complexity of a society. A society with more activities has accumulated more culture over time.

To test the impact of the presence of activities and PO on the fitness of the individuals in society i, we use the following econometric model[27]:

$$Population_i = b1 + b2 \times \%Presence_i + b3 \times \%Presence_i \times \%withinPO_i + Controls_i + Error_i$$

Population as a dependent variable captures the standard notion of fitness as reproductive success. It also captures the fact that in pre-modern Malthusian economies, progress translates into increases in population and not into per-capita wealth (Spolaore and Wacziarg, 2013; De La Croix et al., 2017).

[27] In this model, we do not include an individual term for "% within PO" because this variable is nested within the presence of technologies (e.g., when "%presence" is zero, then "%withinPO" is zero as well). If this individual term were included, then it would mean that even if "%presence" is zero, "%withinPO" can have an impact on population, and this would be contradictory. As a result, the marginal effect of "%withinPO" is scaled by the variable "%presence". Even if the variable "%presence" is different from zero, an individual term for "% within PO" would still be very difficult to interpret, and it would be capturing an effect that is distinct to the one in our theory, which requires the scaling. We executed an estimation adding the individual term (available upon request) and we could verify that the majority of the impact of "% within PO" is exerted via the interaction term.

We use "size of local communities" which is a categorical variable with 8 categories: 1 is "less than 50 people" 2 is "from 50 to 99 persons", 3 is "from 100 to 199 persons", 4 is "from 200 to 399 persons", 5 is "from 400 to 1,000 persons", 6 is "more than 1,000 in the absence of indigenous urban aggregations", 7 is "one or more indigenous towns of more than 5,000 inhabitants but none more than 50,000", and 8 is "one or more indigenous towns with more than 50,000 inhabitants". In figure 4B and 4C we plot the "size of local communities" against the variables "% presence" and "% within PO".

As controls, we added geographical variables (absolute latitude, average temperature, distance to coast, slope of terrain), resource endowment variables (amphibian richness, bird richness, mammal richness, vascular plants richness), intensity of agriculture dummies (complete absence, casual, extensive/shifting, semi-intensive, intensive), regions dummies (36 regions in our final sample), type of settlement dummies (e.g., nomadic, semi-nomadic, etc.) and year of the ethnographic record. From the 263 societies with complete information on the technologies, we lost some societies due the missing data on the dependent variable (54 societies) and in some controls (mostly, resource endowment), leading to a final sample of 173 societies. In figure 3, we display these societies on a geographical map.

Figure 3. Societies included in table 1.

The ordered probit estimates are presented in table 1. Assuming that there are no POs, the results of column 2 of table 1 show that moving from 0% to 100% presence of activities does not generate an increase of local population. This is depicted in the blue line with circle markers in figure 4D. The presence of a wider set of activities in most of the adult population does not translate to a larger population. Although this might seem surprising, it is consistent with Rogers' paradox, in the sense that culture does not necessarily leads to increased fitness. However, consistent with the proposition 1 of the model, activities do increase the local population when POs are present in the society. This increase is economically and statistically significant and is depicted in the red line with triangles in the figure 4D.

Table 1. Impact of presence of technologies and PO on the size of local population

	Dependent variable: Size of local population	
	1	2
% presence	1.137 (0.818)	0.016 (0.846)
% presence x % within PO		4.298*** (1.278)
% within PO		
Geographic controls?	Yes	Yes
Resource endowment controls?	Yes	Yes
Year of ethnography?	Yes	Yes
Agriculture intensity dummies?	Yes	Yes
Continent dummies?	Yes	Yes
Type of settlement dummies?	Yes	Yes
Observations	173	173
Pseudo R Square	0.329	0.352
We execute ordered probit regressions. The dependent variable is the size of local population. Robust standard errors are used in all regressions and are reported in parentheses. *** indicates p-value<0.01.		

2.4.3. Endogeneity

As difficult as it is with this type of data, in this section we address concerns about endogeneity in our key variables.

Omitted variables

The first threat to identification of causality is omitted variables. We executed a test that uses selection on observables to assess the extent to which selection on unobservables would need to be in order to overthrow the results (Oster, 2016). In the table 2 we replicate column 2 of table 1, and columns 2, 5, and 8 of table 6 (see section 2.4.5 for a robustness check with different dependent variables). We report the "Oster delta" in the last two rows of the table, assuming a maximum r-square of 1 and 0.95 respectively. Given the inherent measurement error of ethnographic data, the assumption of 0.95 is a good benchmark for the test (and perhaps conservative). This test assumes a linear model, so we estimate columns 1 and 3 using OLS (in section 4.4.1 we use the more appropriate ordered probit estimation for these dependent variables). The results show that the Oster delta is on average 0.97 when an R-square of 0.95 is assumed. This means that selection on unobservables would need to be at least 0.97 times the selection on observables in order to overthrow our results. A delta of 1 is

a good indication against the threat of omitted variables, particularly if a comprehensive set of controls is used (Oster, 2016).

Figure 4. The impact of the presence of activities and POs on the size of local population. The figures use the sample used in table 1. (A / B / C) Scatter plots where the size of the bubbles represents frequency of societies. (D) Here we plot of the second column of results in table 1. We evaluate how much the probability of each one of the eight size categories changes if the presence of activities goes from 0 to 1. We present the average of the marginal effects. To explore how this impact varies with POs we set the variable "% within PO" equal to zero and equal to 0.5.

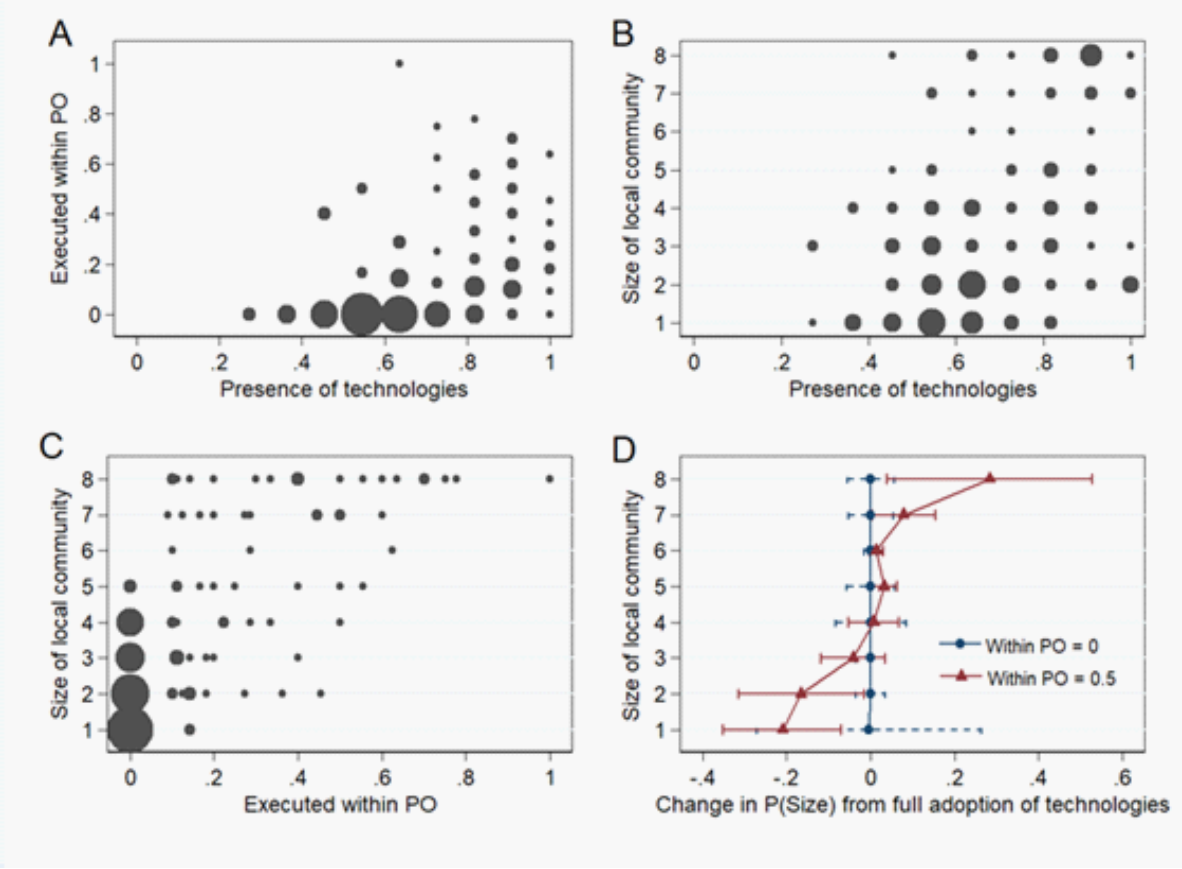


Table 2. Selection on observables versus selection on unobservables

	1	2	3	4
	Local community size	Total population size	Population density	Cultural complexity
% presence	-0.143 (1.166)	-0.564 (1.997)	-0.669 (0.939)	-0.029 (5.155)
% presence x % within PO	5.238*** (1.592)	5.818*** (2.241)	3.159*** (1.202)	14.887*** (5.650)
All controls of table 1?	Yes	Yes	Yes	Yes
Observations	173	153	125	125
R-Square	0.771	0.819	0.409	0.888
Oster delta (R^2 max = 1)	0.41	0.52	0.71	0.93
Oster delta (R^2 max = 0.95)	0.51	0.69	0.98	1.69
Robust standard errors are used in all regressions and are displayed in parentheses. *** indicates p-value<0.01, ** indicates p-value<0.05, * indicates p-value<0.1. The Oster delta is computed for the interaction term "%presence x within PO"				

Reverse causality

The second threat to identification is reverse causality, both in the presence of activities and in the use of PO. We analyse each one in turn.

An important proposition of Cultural Evolution is that a larger and more inter- connected population would generate more cumulative culture (Henrich, 2015). In our case, this would translate into a higher presence of activities in the society which could be channelled through the interaction with POs (assuming, for now, exogeneity in POs). In order to address this issue we instrumented "% of presence" using two variables: "sex differentiation" (which we detail in section 4.3.2) and the index of "kinship tightness" developed by Enke (2017). On the first instrument, there is evidence that the presence of activities coevolved with sex differentiation in pre-modern societies (Haun and Over, 2013). For example, men specialize in large game hunting while women specialize in gathering. Sex differentiation might affect population size if it affects fertility. We con- tend that conditional on the total amount of activity executed by women, and therefore controlling for the time restriction that differentiation might place women fertility, the exclusion restriction should hold.

Kinship systems regulates the pattern of relatedness in society through family structure (e.g., independent nuclear families vs. extended families, post marriage residence in wife or husband's

group vs. independent residence), marriage patterns (e.g., cousin marriage allowed vs. forbidden, polygamy vs. monogamy), and descent (e.g., unilineal vs bilateral descent group, presence of clans sharing a geographical location). Kin- ship tightness is a key variable affecting social organization of a society (Enke, 2017). Tight kinship (i.e., extended families, cousin marriage allowed, clans, polygamy, unilineal descent, etc.) generates high in-group bias, less cooperation with outsiders, strong conformism, and local institutions. The opposite occurs with loose kinship, with the consequence of being much more open to external groups. We argue that Kinship tight- ness affects the presence of activities in society. A tight kinship system should increase the presence of basic and widely known activities through less reliance on sourcing activities from neighbouring societies[28]. A "closed" society does not have an alternative but to provide the basic activities internally. A society with loose kinship, and therefore open, can source part of the basic activities from neighbours. The exclusion restriction for "kinship tightness" is sustained on the documented ancestral origins of kinship systems (Passmore and Jordan, 2017). Kinship systems can be traced back into the societies from which the focal society descent from. Thus, this element of societies can be treated largely as an exogenous variable, particularly when controlling for agriculture intensity and settlement type[29]. Furthermore, there are no a-priori reasons to think that Kinship tightness might generate larger or smaller populations through changes in fertility. For example, the polygamy-fertility literature is not at all conclusive. Accordingly, and consistently with Enke (2017), we do not find a relationship between kinship tightness and population in our data, conditional on covariates.

In the table 3, we present the instrumental variables estimations. In the column 1 we present the first stage. As expected, both sex differentiation and kinship tightness are positively related to the presence of technologies. (Kinship tightness has a p-value of 0.16; if we drop sex differentiation, the p-value increases to 0.04.) By comparing the Cragg-Donald F-test of the first stage (reported

[28] A society with high kinship tightness tends to be more isolated from neighbouring societies. This would increase the need for having all the activities provided inside the society. A society with low tightness would be much more willing to provide some of the activities from abroad. Notice that this would hold for basic activities that have long being invented and diffused. For the case of innovation of newer and more complex activities, the impact of kinship tightness is detrimental to the adoption of innovations from other societies, as shown in Enke (2017).

[29] There are arguments and evidence that indicates that kinship tightness evolved to optimally match the needs of agricultural subsistence, away from nomadism (see Enke, 2017).

in columns 2 and 3) with the values in Stock and Yogo (2002) we can conclude that our instruments are not weak. The Hansen-test indicate that the instruments are indeed exogenous, in line with the theoretical arguments laid out above.

In columns 2 and 3 we present the second stage, with and without the interaction with % within PO respectively. The results do not change from those of the table 1: the presence of technologies increases the local community size, but only when PO are in place. In columns 4 and 5 we use total population of the society as the dependent variable (see the section 4.4 below for the details of this alternative dependent variable). The results show that the presence of technologies has a positive impact independent of the percentage executed within PO. However, consistent with our prediction (and column 2 of table 6), this impact is larger when POs are in place. In column 5, if we assume absence of POs, then the impact of presence of technologies is statistically not different from zero.

Table 3. Instrumenting the presence of technology

Stage:	1st stage	2nd stage		2nd stage	
Dependent variable:	% presence	Local community size		Population size	
	1	2	3	4	5
% presence		4.793 (4.219)	2.958 (3.791)	12.982** (6.492)	6.221 (5.114)
% presence x % within PO			5.475*** (1.217)		10.865*** (2.669)
Kinship Tightness	0.061 (0.044)				
Sex differentiation	0.119** (0.053)				
Same controls as in table 1?	Yes	Yes	Yes	Yes	Yes
Region dummies and resource endowment controls?	No	No	No	No	No
Observations	194	194	194	160	160
Cragg Donald f-test first stage (p-value)		9.55*** (0.008)	9.62** (0.022)	8.884** (0.012)	8.085** (0.044)
Hansen test (p-value)		0.309 (0.578)	0.707 (0.702)	3.438* (0.064)	3.843 (0.146)

Robust standard errors are used in all regressions and are displayed in parentheses. *** indicates p-value<0.01, ** indicates p-value<0.05, * indicates p-value<0.1. In columns 3 and 5, we use "kinship tightness x % within PO" and "sex differentiation x % within PO" as instruments for "% presence x % within PO". We drop the controls of geographical region because the local geographical variation in our instruments is not high. Given that we rely on inheritace from ancestral societies, the societies occupying a particular region tend to share the several cultural traits from their common ancestor. We drop the resource endowment variables in order to avoid data loss and to avoid small sample bias in the IV estimation (results are consistent if we include these controls).

The third and main threat to identification is endogeneity problems with the variable "% within PO". In particular, in case there is a minimum size to POs, large populations might make it easier to have POs. In addition, a common argument is that specialization is favoured by the extent of the market. If any of these cases is correct, we might have a reverse causality problem. We address this issue by instrumenting the presence of POs following the idea of Depetris-Chauvin and Ozak (2017). These authors explore the drivers of the presence of POs in pre-modern societies using the ethnographic atlas. They explore the extent to which diversity in the population of a society -- measured by genetic diversity in the societies of the Atlas-- drives the presence of POs. The theoretical argument is that a genetically diverse population has many different skills in place which would lead to the creation of specialized groups. These authors instrument genetic diversity using the distance of the society from East Africa (specifically, modern day Ethiopia) which is the

origin of the spread of the human species out of Africa (starting approximately 80,000 years ago). As the distance from Africa increases, the diversity within a society goes down because migrant societies are not a random sample of the society of origin[30]. The authors find substantial evidence in favour of their arguments: distance reduces genetic diversity which in turn decreases the presence of POs.

In table 4, we follow these authors and we use "Distance from Africa" as an instrument for "% within PO". We measure the distance from Addis Adaba in east Africa to the focal society; for societies in America, we calculate the distance going through the Bering strait. We do not use the mediating variable of genetic diversity, and thus, we implement a "reduced form" model of Depetris-Chauvin and Ozak (2017)[31]. In the column 1 we present the first stage. Consistent with Depetris-Chauvin and Ozak (2017), the distance from Africa reduces the presence of POs in societies. Although the Cragg-Donald test (reported in columns 2 and 3) indicate that the instrument is relevant, by comparing the values with Stock and Yogo (2002) we cannot rule out weakness in the instruments. In order to address this issue, in the second stage of column 2 and 3 we use the limited information maximum likelihood (LIML) technique, which partially mitigates the problem of weak instruments. The results we obtain with both dependent variables are consistent and supportive of our predictions. By comparing the values of the coefficient with those of table 1 and table 6, we also find that the coefficients display an increase in their size.[32]

[30] This also implies that cultural heterogeneity across societies would increase with the migratory distance to a common ancestor. Becker et al (2018) corroborate this prediction.

[31] This reduced form allows other mechanism to impact the presence of POs. For example, given the non-randomness of the migratory sampling process it could also be the case that traits are lost. As migrant groups are typically small, the likelihood of loss increases due to drift.

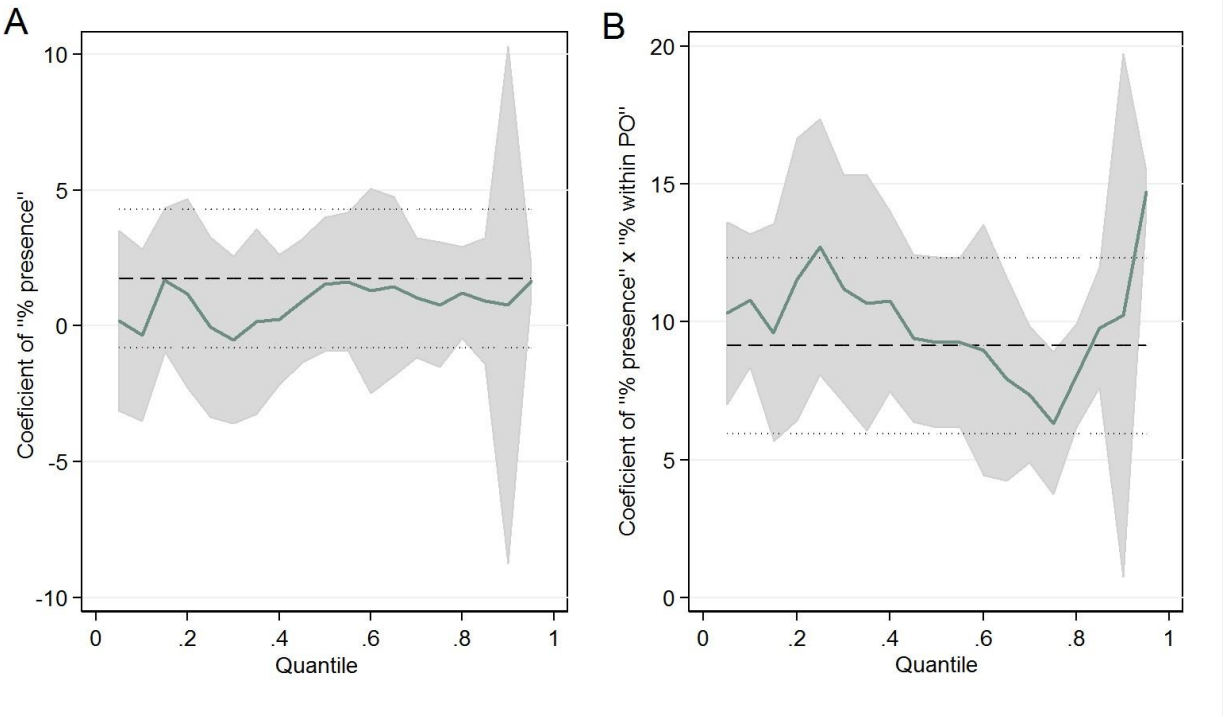
[32] We also instrumented "% presence" and "% within PO" at the same time. We used the three instruments simultaneously following Wooldridge (2010, chapter 8). The results are consistent with table 3 and table 4. The coefficient for "% presence" is 1.32 and for "% within PO" is 3.83. However, statistical significance is lost. The instruments retain their properties: for strength, they surpass Stock and Yogo thresholds on strength; for exclusion, the Hansen test indicates that the instruments are valid.

Table 4. Instrumenting the percentage within PO

Stage:	1st stage	2nd stage	2nd stage
Dependent variable:	% presence x % within PO	Local community size	Total population
	1	2	3
% presence	0.278*** (0.082)	-0.408 (1.223)	-1.316
% presence x % within PO		6.606* (3.751)	17.70*** (6.071)
% presence x Distance from Africa	-5.37e-06** (2.73e-06)		
Same controls as in table 1?	Yes	Yes	Yes
Region dummies?	No	No	No
Observations	173	173	153
Cragg Donald f-test first stage (p- value)		5.228** (0.022)	5.790** (0.016)
Hansen test (p-value)		n/a	n/a
Robust standard errors are used in all regressions and are displayed in parentheses. *** indicates p-value<0.01, ** indicates p-value<0.05, * indicates p-value<0.1. We use LIML in the estimations. We exclude region dummies because the variation of our instruments within regions is low.			

An additional way to address the problem of reverse causality is that both theoretical arguments (i.e., minimum size of PO and extent of the market) would predict that the positive relationship between population and the presence of PO would be much stronger at higher levels of population. However, in figure 4D we show that this is not the case. The impact of PO on the increase in population size is exerted throughout the different size categories of the variable "size of local community". For the case of the dependent variable of total population, we replicated column 2 of table 6 using a quantile regression estimation (and dropping the region controls, as they limit the estimation). In figure 5A we display the value of the coefficient related to "% presence" as it varies across the dependent variable. In figure 5B we do the same for the interaction term "% presence x % within PO". In both cases, the graphs show that the positive impact of POs on population is exerted evenly across different population sizes, reducing the concern for reverse causality.

Figure 5. Coefficients in quantile regression.



2.4.4. Comparative statics

In this section, we test the comparative statics derived from our model and summarized in figure 2. The econometric model we use is the following[33]:

$$Population_i = b1 + b2 \times \%Presence_i + b3 \times \%Presence_i \times \%withinPO_i + b4 \times \%Presence_i \times Z + b5 \times \%Presence_i \times \%withinPO_i \times Z + Controls_i + Error_i$$

In this model we generate a triple interaction to explore whether the impact of POs is affected by the variable Z. We use different variables as Z in order to capture the different parameters p, c and c, as well as the prevalence of secrecy. If the coefficient b5 is positive (negative), then the impact of PO is enhanced (diminished) by the variable Z.

[33] For the same reasons explained in the footnote 8, the interaction term between "%withinPO" and Z is not included.

Uncertainty

We first test the impact of environmental uncertainty, the parameter p of the model, by using climate unpredictability as the proxy. Climate data has already been successfully used to empirically test the parameter p in cultural evolution models (Giuliano and Nunn, 2017). The D-PLACE dataset reports "temperature unpredictability" and "precipitation unpredictability" which are measured using yearly data between 1901 and 1950, the period that has the largest proportion of ethnographies in the ethnographic atlas. The measure of unpredictability captures both the extent to which temperature or precipitation patterns are predictable because these conditions are constant or whether they oscillate in a predictable manner (Colwell, 1974). We multiply these two measures to obtain our measure of climate unpredictability (if used individually results don't change). Consistent with the comparative statics of the model, the results from the column 1 of table 5 show that the impact of POs on population decreases when climate unpredictability is high. A joint t-test shows that the impact of PO is again highly significant and importantly moderated by climate unpredictability. This result is graphed in the figure 6A.

Social learning costs

We studied three variables that decrease the costs of social learning. First, the SCCS provides information on how rooted apprenticeship and teaching are in the society. "Apprenticeship" is a dummy variable we computed from the variables v427 and v428 of the SCCS that measure the extent of guidance and/or formal schooling in late boys and girls respectively. The dummy takes the value of 1 when either variables indicate that the society displays "predominant apprenticeship" or when "formal schooling is frequent and typical", and zero otherwise. Clearly, if schooling and apprenticeship is predominant in society, this will decrease both c and c. Given that a lower c has an ambiguous impact on the fitness of POs (see figure 2A), but a lower c has a unequivocal increase in the fitness of POs (figure 2B), we predict that "apprenticeship" should boost the impact of POs on population size. (It could also be argued that this dummy would be more tightly connected to a decrease in c, than in c, because teaching and apprenticeship probably coevolved with POs.) Many hunter-gatherers societies --which lack POs-- possess less teaching (relative to more advanced sedentary societies) and it is restricted mostly to kin (Hewlett and Roulette, 2016). This possibility would reinforce our prediction. The results are displayed in column 2 of table 5 and are consistent with our prediction: the positive impact of POs on population are higher if apprenticeship is predominant. We graph this result in figure 6B.

The second variable that reduces cost of social learning is "sex differentiation". The variables v44 to v54 of the EA provide information on the extent to which the eleven activities are executed by women and/or men. For each activity we coded a dummy that took the value of 1 in case the activity was executed by "males only or almost alone" or by "female only or almost alone". Then we added these dummies and divided the result by the total number of activities that have available information. We label this variable "sex differentiation" and it captures the percentage of activities that are done by either sex exclusively. There is plenty of evidence that social learning is facilitated by similarity, in which sex plays an important part (Haun and Over, 2013; Henrich, 2015; Rendell et al, 2011). Similar to "apprenticeship" this variable reduces both c and c, and therefore our model predicts more sex differentiation would lead to an increase in the impact of POs. This is what we find in our estimations. In the third column of table 5 we obtain a positive and significant coefficient in the interaction term.

The third variable that reduces the cost of social learning is "Trust". We use the variable v335 of the SCCS which measure the degree to which trust in inculcated in childhood in the society. This variable is ordinal, with 0 meaning "no inculcation or opposite trait" to 9 meaning "extremely strong inculcation". As with "apprenticeship" and "sex differentiation", high "trust" decreases both in c, than in c leading to the prediction of a higher impact of POs. The result is displayed in column 4 of table 5 and is consistent with the prediction from the model: POs have a larger positive impact on population when trust is high. This result is graphed in figure 6C.

Secrecy

Finally, we analyse the impact of the variables "Honesty" and "Generosity". These are the variables v336 and v334 of the SCCS and are analogous to v335, namely, a categorical variable identifying inculcation of honesty and generosity in childhood. It is possible to identify these variable with a decrease in the degree of secrecy in the POs and therefore a boost in their fitness. Oftentimes, secrecy is related to selfish behaviour, a desire to keep useful knowledge proprietary. The zeal to maintain secrecy could also benefit from dishonest behaviour, deflecting requests to share knowledge with negation of its possession. It would also be possible to relate "honesty" and "generosity" to a decrease in the costs of social learning. When agents are generous and honest it is very likely that communication and learning would improve. In any of both cases, the prediction from our model is clear, these variables should increase the fitness benefits of POs. The results are presented in the

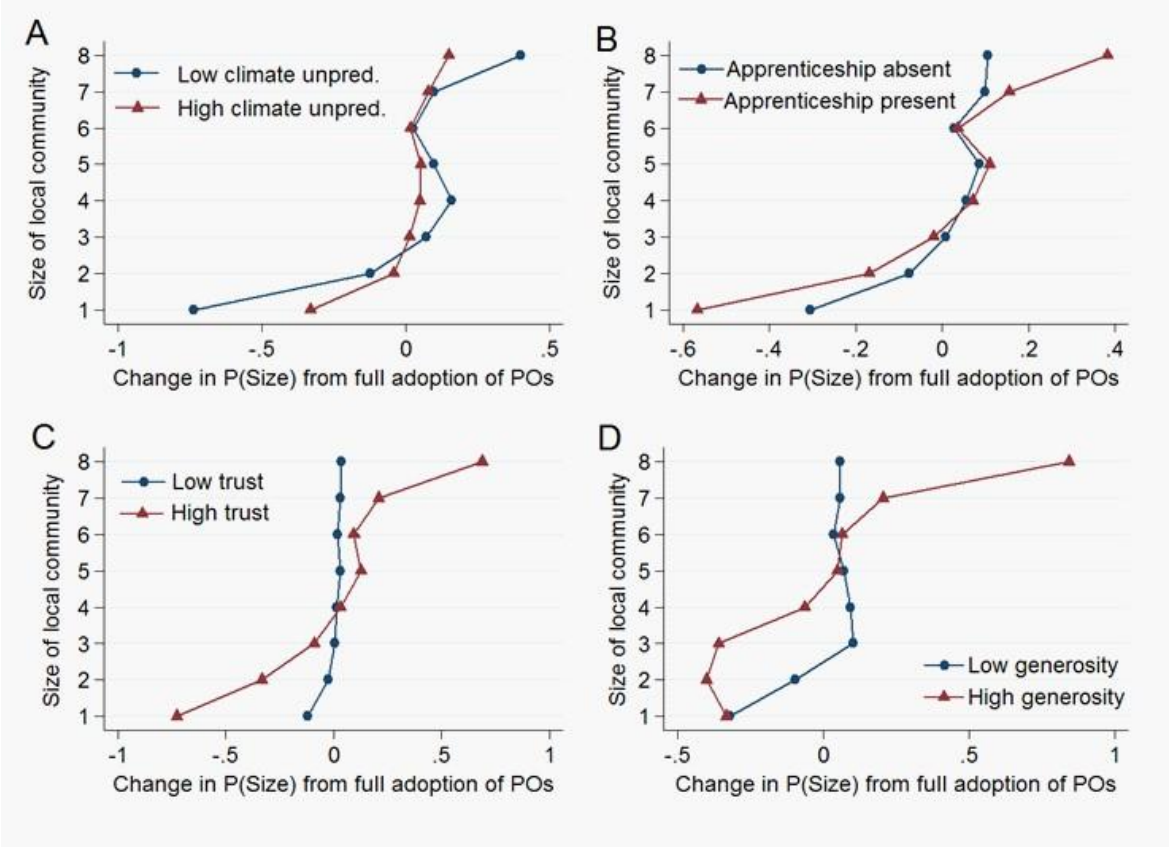
columns 5 and 6 of table 5 are consistent with the prediction of our model. In the figure 6D we graph the result for generosity.

Table 5. Heterogeneity in the impact of POs

	Dependent variable: Size of local population					
	1	2	3	4	5	6
% presence	-3.24 (3.62)	-0.536 (1.590)	0.125 (1.241)	1.112 (1.694)	1.364 (1.701)	0.973 (2.679)
% presence x % within PO	12.28# (8.56)	2.629*# (1.547)	-0.811# (4.684)	0.066 # (2.920)	0.690 # (2.999)	1.898 # (3.231)
Climate unpredictability	-3.25 (4.06)					
Clim. unp. x % presence	4.75 (5.03)					
Clim. unp. x % pres. x % within PO	-10.66# (11.38)					
Apprenticeship		-0.553 (1.129)				
Appren. x % presence		0.237 (1.556)				
Appren. x % pres. x % within PO		3.336*# (1.860)				
Sex differentiation			not included			
Sex differ. x % presence			-0.078 (1.216)			
Sex differ. x % pres. x % within PO			6.955# (6.468)			
Trust				0.250 (0.179)		
Trust x % presence				-0.275 (0.314)		
Trust x % pres. x % within PO				0.768 # (0.644)		
Honesty					0.313 (0.271)	
Honesty x % presence					-0.622 (0.392)	
Hon. x % pres. x % within PO					1.148*# (0.652)	
Generosity						0.574** (0.289)
Gen. x % presence						-0.517 (0.426)
Gen. x % pres. x % within PO						0.598 # (0.615)
Resource endowment control?	Yes	No	Yes	No	No	No
All other controls of table 1?	Yes	Yes	Yes	Yes	Yes	Yes
Observations	173	136	173	101	84	79
Pseudo R Square	0.354	0.335	0.353	0.239	0.281	0.299

Notes: Robust standard errors are used in all regressions and are displayed in parentheses. *** indicates p-value<0.01, ** indicates p-value<0.05, and *** indicates p-value<0.1 . # indicates p-value<0.01 for joint two tailed F-test of dsize/d%withinPO. To avoid losing excessive observations by using the SCCS variables, the sample of columns 2, 4, 5, and 6 allow at least 6 technologies with available information.

Figure 6. Heterogeneity of impact. In this figure, we analyse the predictions from the comparative statics of our model. We use the estimation of table 5. (A) We plot the average of the marginal effects of "% within PO", that is, the change the probability of each one of the eight size categories is this variable goes from 0 to 1. To show how this impact varies with climate unpredictability we set this variable to the minimum and maximum values in the sample used in the estimation. (B / C / D) Analogous to graph A. Confidence intervals are not displayed due to multicollinearity in the estimates (see legend of table 5).



Comparison with Transaction Cost Economics

The results we obtain for the comparative statics --except for apprenticeship-- are opposite to the ones that TCE would predict (Williamson, 1991; Tadelis and Williamson, 2013). TCE predicts that uncertainty would increase the benefits and the frequency of POs as opposed to market exchange; trust and reputation would move transactions away from hierarchical POs and towards market or hybrid governance; and the risk of leakage (and thus, secretive behaviour) would lead to heavier

reliance on POs[34]. As discussed in sections 2.2 and 3.1 above, our theory provides opposing predictions which, at least in our data, are supported.

Given the empirical support for the mechanisms proposed by TCE (Lafontaine and Slade, 2007), how can we make sense of this divergence? We speculate that this might indicate that when studying the evolution POs, the explanatory logic changes. In our theory, we do not consider incentives and governance issues in order to focus on the impact that POs have on cumulative culture, that is, on the increase in the pool of useful knowledge and technologies. In contrast, TCE assumes a knowledge pool in order to focus on the governance of the transactions. Consistently, the unit of analysis in cultural evolution is the population, while in TCE it is the PO and its transactions. Thus, these theories need not be contradictory, instead they can complement each other by operating at different levels and time scales. For example, the advantage in social learning costs of POs might have its roots not only in identity and self-enforced cooperation, but also in hierarchical governance devices that minimize the hazards of knowledge transmission. The exploration of these different explanatory logics, and how they interact, is a topic for further research[35].

2.4.5. Robustness checks

Robustness to alternative dependent variable

The results of table 1 are robust if the following three alternative dependent variables are used: "total population", "population density", and "cultural complexity". We address each one in turn. The variable "total population" is obtained from the Ethnographic Atlas and is a continuous variable that indicates the total population of the society. We use natural logarithms to normalize

[34] Higher generosity and honesty could also be related to lower opportunism. Transaction costs economics would again predict that using POs would provide less benefits, the opposite of what we find.

[35] One possible connection that can be done is in the type of "comparative assessment" that TCE and our theory executes. TCE evaluates under which circumstances a particular organizational arrangement –for example, in-house versus outsourcing– is performing better. Therefore, it is a contingent comparative assessment: depending on the circumstances, one or the other is better. In our theory, we aim at establishing an absolute advantage of PO over its alternative (i.e., the execution of the activities/technologies by independent individuals). If an absolute advantage can be established (which in our case happens if $p>0$, so almost always), then we can argue that the evolutionary processes of cultural selection will slowly select and diffuse the PO, which will ultimately invade the landscape. In a way, we are explaining the object that TCE plays with: executing the transaction 'x' inside the firm versus supply it from another firm; in both cases it is a firm.

its distribution. In the columns 1, 2 and 3 of table 6 we display the results. In order to assess robustness of the comparative statics, in column 3 we include the interactions with climate unpredictability (other interactions were also robust to change in the dependent variable across table 6; these estimations are available upon request). The results show that total population increases with the presence of technologies, but only when these are executed by POs. The positive impact of POs on total population increases when climate unpredictability is low.

"Population density" is the variable v1130 of the SCCS and is a categorical variable with 1 equal to "less than 1 person per square mile", 2 equal to "1 - 4.9 persons per square mile", 3 equal to "5 - 24.9 persons per square mile", 4 equal to "25 - 99.9 persons per square mile", 5 equal to "100 - 499.9 persons per square mile", and 6 equal to "500 or more persons per square mile". The results of column 4, 5 and 6 of table 6 show that our findings are also robust to the use of population density as dependent variable.

Finally, "cultural complexity" is the variable v158.1 of the SCCS, where they sum the scores of 10 variables that capture the degree of cultural sophistication: writing and records, fixity of residence, agriculture, urbanization, land transport, money, density of population, political integration, social stratification, and specialization in metal working, weaving and pottery. In our case, we subtracted from the variable the last component of specialization. Overall, the results of columns 7, 8 and 9 of table 6 show that the results presented are robust to this alternative dependent variable.

Table 6. Robustness to other dependent variables

Dependent variable:	Total population			Population density			Cultural complexity		
	1	2	3	4	5	6	7	8	9
% presence	0.773 (2.253)	-0.564 (1.997)	-21.167*** (5.947)	0.789 (0.859)	-0.116 (0.877)	-6.213 (3.694)	4.660 (4.794)	0.313 (4.683)	-8.208 (13.050)
% pres. x % within PO		5.818*** (2.241)	24.996# (14.409)		3.584*** (1.101)	8.537# (5.423)			64.518*** (18.426)
Climate unpredictability			-12.314** (5.659)			-1.256 (4.509)			-2.943 (14.188)
Climate unp. x % presence			29.596*** (8.307)			8.808 (5.519)			11.606 (20.434)
Climate unp. x % presence x % within PO			-27.183# (19.840)			-7.028# (6.976)			-61.207*** (23.719)
All controls of table 1?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	153	153	153	128	128	127	128	128	127
Pseudo R-Square (r-square)	0.801	0.819	0.835	(0.341)	(0.373)	(0.388)	0.807	0.828	0.835

Robust standard errors are used in all regressions and displayed in parentheses. *** indicates p-value<0.01 and # indicates p-value<0.01 for joint two tailed F-test of 8size/8%withinPO in columns 3, 6 and 9. High multicollinearity-frequent in interaction models- requires a joint test. Columns 4, 5 and 6 use Ordered Probit, the rest OLS. For the columns 4 to 9 we use societies with information in at least 8 technologies in order to accommodate for the smaller sample size in these dependent variables (in the table 6 below we show that the results are robust to data stringency).

Alternative explanations

There are three main alternative explanations for our empirical results. We address each one in turn. First, it could be argued that the positive impact of POs is due to POs generating improvements on the costs of individual learning, rather than on the costs of social learning. To assess this possibility we studied a model where a PO decreases C instead of c. A model with this characteristic generates POs populated entirely by individual learners and, importantly, their benefits are increasing in the uncertainty parameter p. The latter implication is directly contradicted by our interaction with uncertainty: we find that uncertainty reduces the impact of POs on population (column 1 of table 5 and figure 5A). Of course, this mechanism could also be present, but the results show that the opposing mechanism, that is, POs decrease social learning costs, is stronger. The former implication --POs are populated by individual learners-- is rebutted by simple perusal of organizational reality: in general, social learners dominate individual learners inside POs (and the opposite occurs in the market). Guilds are a good example.

A second alternative explanation to our empirical findings is related to trade. Specialized POs might have a positive impact on population because they are a marker for the presence of trade in the society. And trade could be the fundamental driver of larger populations and the key force behind the evolution of specialized POs. To test this alternative explanation we use as controls several variables from the SCCS that proxy for the presence of trade in the society. Specifically we use four variables: i) "inter-community trade" (v1 of the SCCS) is a categorical variable that measures the extent to which inter-community trade is a source of food (from "no trade" to "food imports present and contribute more than 50%"), ii) "presence of money" (v17) is a categorical variable with five categories (from "no media of exchange or money" to "indigenous coinage or paper currency"), iii) , "presence of credit" (v18) is a categorical variable with four categories (from "Personal loans between friends or relatives" to "banks or comparable institutions"), iv) and "importance of trade" (v819b) is a continuous variable measures the percentage that trade contributes to subsistence in the society (i.e., food provision)[36].

[36] This variable is computed by the SCCS from using v1 and other five variables that provide categorical information on the extent that agriculture, fishing, gathering, animal husbandry and hunting contribute for subsistence (the mean is 8% with a maximum of 65%, a median of 5% and a 90th percentile equal to 25%).

In column 2 of table 7 we add "Inter-community trade" as a control; in column 1 we use the same societies used in column 2. This allows to cleanly assessing the impact of the control variable on the impact of PO. The same is done in columns 4 and 3 for "money" and "credit" and in columns 6 and 5 for "importance of trade". Across these three comparisons, the coefficients decrease in size by an average of 16% but remain statistically significant. The largest decrease is experienced in the model that includes "money" and "credit". These reductions in the size of the coefficients indicate that some of the impact of POs is indeed generated through trade benefits, but that it is not the main mechanism. Instead, this result is consistent with our proposition 3 that states the origin of specialized POs is driven by the need to make social learning useful in society, without requiring trade as a force for its evolution. Thus, even in societies without trade specialization within POs would be beneficial A secondary way to assess the alternative explanation of trade is by exploring the interaction with uncertainty. The literature on trade has proposed and documented that trade (and thus the specialization it drives) is particularly useful to mitigate the effects of shocks to local productivity, such as weather changes (Burgess and Donaldson, 2010). Therefore, if the benefits of trade is the key driver of the impact of POs we should find a positive interaction of POs with uncertainty; however, we find the opposite in our results (column 1 of table 5 and figure 6A).

Table 7. Robustness to trade.

	Dependent variable: Local community size					
	1	2	3	4	5	6
% presence	1.079 (1.235)	0.762 (1.267)	0.751 (1.243)	1.892 (1.421)	0.980 (1.227)	0.925 (1.259)
% presence x % within PO	4.237*** (1.530)	3.897*** (1.524)	3.978** (1.551)	2.630* (1.559)	4.261*** (1.532)	4.072*** (1.503)
"Intercommunity trade" dummies?	No	Yes	No	No	No	No
"Money" and "Credit" dummies?	No	No	No	Yes	No	No
"Importance of trade" control?	No	No	No	No	No	Yes
Resource endowment controls?	No	No	No	No	No	No
All other controls?	Yes	Yes	Yes	Yes	Yes	Yes
Observations	130	130	124	124	131	131
Pseudo R-Square	0.367	0.379	0.368	0.422	0.367	0.375

Notes: We execute ordered probit regressions. Robust standard errors are used in all regressions and are displayed in parentheses. *** indicates p-value<0.01, ** indicates p-value<0.05, * indicates p-value<0.1. In order to avoid data loss, in all regressions we use societies with information in at least 8 technologies and drop resource endowments controls.

The third alternative explanation for the origin of POs is that they emerge as a result of political complexity. The idea is that having a complex political organization in the society, allows to better define, monitor and enforce POs. Thus, it might be that political complexity is really driving both the presence of POs and a larger population. In table 8 we show that our results are robust to adding "political hierarchy" as a control. This variable is a categorical variable from the Ethnographic Atlas that indicates whether the society has political authority, and if it does, the reach of this authority (local chiefdoms, large chiefdoms, small states, large states). In these models we replicate column 2 of table 1 and columns 2, 5 and 8 of table 6 but with the addition of the control of political hierarchy. Comparing the estimated coefficients with those of table 1 and table 6, the results shows that the coefficients are reduced by 23% on average (across dependent variables). This results indicates that this alternative explanation carries some weight, but not enough to overthrow our results. Of course, it could also be argued that political complexity is driven by POs in the first place. If that is the case, including this control would be biasing downward the true impact of POs.

Table 8. Robustness to political hierarchy

	1 Local community size	2 Population size	3 Population density	4 Cultural complexity
% presence	-0.190 (0.861)	-0.253 (2.053)	-0.669 (0.939)	-0.020 (3.350)
% presence x % within PO	3.790*** (1.428)	4.851** (2.490)	3.159*** (1.202)	8.964** (4.360)
Political hierarchy dummies?	Yes	Yes	Yes	Yes
Region dummies?	Yes	Yes	No	No
All other controls from table 1?	Yes	Yes	Yes	Yes
Observations	171	151	125	125
R-Square (Pseudo R-Square)	(0.358)	0.830	(0.409)	0.888

For "Population density" and "Cultural complexity" we don't use region dummies and we use societies with information in at least 8 technologies in order to accommodate for the smaller sample size in these dependent variables. Robust standard errors are used in all regressions and displayed in parentheses. *** indicates p-value<0.01. Columns 1 and 3 use OLS, columns 2 and 4 use Ordered Probit.

Data stringency

The results are also robust to being less restrictive on the information available on the activities (table 9). In many societies, there is information only for portion of the activities. In columns 1 to 4 we change the minimum number of activities have available information in the society and the results do not change. In addition, in columns 5 and 6 we restrict the sample to regions that have at least 2 and 3 societies in them, leading to a loss of 7 and 19 societies respectively. The results are robust to changing both of these information criteria.

Table 9. Robustness to available information on activities and regions

	Dependent variable: Size of local population (Ordered probit)					
	Number of activities				Societies per region	
	1	2	3	4	5	6
sample:	at least 7	at least 8	at least 9	at least 10	at least 2	at least 3
% presence	0.998** (0.472)	0.956* (0.525)	0.912 (0.590)	-0.047 (0.647)	0.013 (0.828)	-0.246 (0.832)
% presence x % within PO	3.230*** (0.715)	3.346*** (0.759)	3.201*** (0.837)	3.412*** (0.978)	4.185*** (1.243)	3.513*** (1.261)
All controls of ta-	Yes	Yes	Yes	Yes	Yes	Yes
ble 1 included?						
Observations	430	393	330	269	166	154
Pseudo R-Square	0.287	0.286	0.295	0.306	0.342	0.3
We execute ordered probit regressions. Robust standard errors are used in all regressions and are displayed in parentheses. *** indicates p-value<0.01, ** indicates p-value<0.05, * indicates p-value<0.1.						

2.4.6. Testing the impact on the contemporary importance of tradition

The final empirical test we execute is drawn from Giuliano and Nunn (2017). These authors test an important implication of the baseline model: the use of social learning decreases with environmental uncertainty (see footnote 7). When the environment changes frequently, people increase the use of individual learning and decrease their reliance on inherited tradition (i.e., social learning). To test this idea, Giuliano and Nunn generate a measure of the environmental instability that the ancestors of a country's population were subject to. First, they create a mapping that breaks down a country's population according to their ancestry in different societies present in the ethnographic atlas. Second, they use the intergenerational temperature variability that was in place between 500 and 1900 in the regions of the country's ancestors to generate a measure of climate

instability at the country level. Third, they analyse the impact of this measure on several measures that capture the use of tradition. Using several empirical approaches across countries, individuals, and descendants of immigrants, they find extensive support for an increase in the reliance on tradition if the ancestors lived in a stable environment.

We use the same idea but instead of relying on weather variability, we rely on the extent of use of PO in the pre-modern societies of the Ethnographic Atlas. Our model predicts that the introduction of POs leads to higher use of social learning in the population, particularly when their advantage in social learning is high[37]. Thus, countries with ancestors that used POs intensively would display a higher use of tradition today. In table 10, we replicate the table 1 of Giuliano and Nunn (2017). The dependent variable is the country level average of likert scale from the World Values Survey (WVS) that assess the agreement with the sentence "Tradition is important to this person; to follow the family customs handed down by one's religion or family" (we use last two waves of the WVS). Using the mapping of Giuliano and Nunn (2017), the independent variables "% presence" and "% within PO" measure the presence of activities and POs for a country's ancestors (which typically comprise more than one society from the Atlas). The column 1 of table 10 shows that POs are associated with less tradition, contrary to the model's prediction. However, the presence of POs in ancestors is strongly associated with modern economic success (Depetris-Chauvin and Ozak, 2017), which in turn is strongly associated with less tradition. In column 2, we introduce the natural logarithm of current GDP per-capita as a control and, as expected, the coefficient related to POs reduces its size and loses its significance. In column 3 we expand the controls using controls built from the EA that measure progress in pre-modern times and that are also correlated with modern success (Depetris-Chauvin and Ozak, 2017). The impact of POs on tradition becomes positive, but not statistically significant.

The prediction of our model suggest that lower social learning costs within ancestral POs would further the reliance on tradition. In column 4 we test this idea by adding the interaction with "sex differentiation". Social learning improves dramatically when learning from the same sex (Haun

[37] The exception to this prediction is in the case of no-secrecy when the size of the PO is below it point of maximum fitness (see figure 1A). In this case, an expansion of the PO reduces the share of social learners outside, leaving the total share of social learners unchanged at the population level. Thus, our prediction is true on average.

and Over, 2013; Rendell et al, 2011) and increases the impact of POs (column 3 of table 5). This would expand the use of social learning in ancestors and translate into increased more reliance of tradition. The result of column 5 is consistent with this logic: when ancestors had POs that were differentiated by sex (and thus boost social learning), countries rely more on tradition today. This correlation is not small: using the mean of presence of technology, and setting sex differentiation in its 90th percentile, the impact of one standard deviation increase in the use of POs increases the importance of tradition by a bit more than half standard deviation.

Table 10. Impact of POs on the contemporary importance of tradition

	Dependent variable: Importance of tradition			
	1	2	3	4
% presence	0.824 (0.687)	0.449 (0.668)	0.283 (0.653)	0.484 (0.604)
% presence x % within PO	-0.840** (0.345)	-0.122 (0.453)	0.256 (0.445)	-2.926** (1.364)
% presence x Sex differentiation			-0.811* (0.434)	-1.557*** (0.451)
% pres. x % within PO x Sex diff.				5.646** (2.406)
Absolute latitude		-0.001 (0.004)	-0.006 (0.004)	-0.007 (0.004)
Complexity of settlements		-0.041 (0.041)	-0.019 (0.047)	-0.005 (0.044)
Political hierarchies		0.028 (0.099)	0.003 (0.104)	0.056 (0.102)
Ln (GDP per capita)		-0.134** (0.060)	-0.134** (0.057)	-0.139** (0.053)
Observations	72	72	72	72
R-squared	0.124	0.248	0.280	0.355

We use OLS estimation. The dependent variable is an ordinal variable (1-6) taken from the WVS. The control GDP per capita is contemporaneous. The rest of variables are constructed from the EA. Robust standard errors are used in all regressions and are displayed in parentheses. *** indicates p-value<0.01, ** indicates p-value<0.05, * indicates p-value<0.1.

In table 11, we also replicate the tables 3, 4 and 5 of Giuliano and Nunn's paper. In these tables, they analyse the impact of temperature variability on the persistence of cultural traits. Again, instead of using temperature variability, we use the presence of POs. In the columns 1, 2 and 3 we study how the presence of POs shifts the degree of persistence in female labour participation (FLP). The data on FLP is drawn from the World Bank Development Indicators. In column 1 we show that

the persistence is 0.32 between 1970 and 2012 (1 being the maximum and 0 the minimum). In column 2 we show that this persistence varies systematically with the presence of POs. Consistent with table 10, we find that POs generate a significant increase in the persistence of female labour participation (positive coefficient on the term "female labour participation in 1970 x % presence x % within PO"). The size of this correlation is not small: setting "% presence" at its mean, a one standard deviation increase in "% within PO" increases the coefficient of persistence by 0.17 (e.g., from 0.3 to 0.47), which equals half of the baseline coefficient in column 1 (we analyse the expression [*d*FLP12 / *d*FLP70 *d*%withinPO] x St. Dev %withinPO). In column 3, we control for all the interactions between FLP in 1970 and the other variables in the model. Although this addresses the impact that ancestral POs have on tradition persistence through their impact on development, it also generates high multi-collinearity, which weakens the statistical significance in this model.[38]

In columns 4, 5, and 6 we repeat the analysis but now we study the persistence of FLP in ancestors on FLP in 2012. Following Giuliano and Nunn, we use the v54 of the EA that measures the presence of females in pre-industrial agriculture. We normalize this variable to make it between 0 and 1. Column 4 shows that the persistence is 0.164 and statistically significant. Compared to column 1, a lower persistence is expected due to the longer time span. The results displayed in column 5 and column 6 indicate that the presence of POs has a positive impact on the persistence of FLP from ancestry to modernity. However, we only obtain statistical significance on the column 6. This could be due to fact that we explore the persistence from ancestry to modernity and most of the interactions added to column 6 control for the impact of other EA variables on persistence. The effect size is large: setting "% presence" at its mean, a one standard deviation increase in "% within PO" increases the coefficient of persistence by 0.19 (e.g., from 0.1 to 0.29), which is slightly larger that the baseline coefficient in column 4. Significance is only at 10% because multicollinearity in high in this model.

[38] In the models of table 11 we don't explore the heterogeneity of impact with respect to "sex differentiation". Sex differentiation in productive activities can impact female labour participation (or polygamy) through many different ways other than the channel we care about, social learning within POs. Thus, estimations would not be reliable. That said, we do find, particularly for Polygamy, that of Sex differentiation boost the positive impact of POs on the reliance on tradition.

In columns 7, 8 and 9 we analyse the persistence of polygamy in ancestors on polygamy in 2009. Polygamy in 2009 is a dummy variable drawn from the OECD Gender, Institutions and Development Database and we follow Giuliano and Nunn for it operationalization: it takes the value of 1 if having more than one spouse is accepted or legal. For polygamy in ancestors, we follow Giuliano and Nunn and use the variable v9 of the EA. We build a dummy that takes the value of 1 if there is presence of polygamy in the ancestors and 0 if the society is monogamous. In column 1 we find a statistically significant persistence coefficient of 0.33. In columns 8 and 9 we find a positive impact of the presence of POs on the persistence of polygamy. However, due to high multicollinearity significance is not present but the effect size is large: using column 9 and setting "% presence" at its mean, a one standard deviation increase in "% within PO" increases the coefficient of persistence by 0.14 (e.g., from 0.2 to 0.34), which is roughly half of the baseline coefficient displayed in column 7.

Although these tests are subject to confounds, the correlations we document are consistent with the prediction of our model. Two features of this exercise provide additional confidence to our account of POs. First, the dependent variable are not drawn from the EA making this an "out-of-sample" exercise. Second, we test a different prediction of our model which is not about fitness but about changes in the share of social learning (and therefore about the reliance on inherited culture).

Table 11. Impact of POs on the persistence of cultural traits

Dependent variable:	Female labour participation in 2012			Female labour participation in 2012			Poligamy in 2009		
Persistence variable "P":	Female labour participation in 1970			Female labour participation in ancestors			Polygamy in ancestors		
	1	2	3	4	5	6	7	8	9
P	0.324***	-1.338*	-1.673	0.164**	0.158	0.062	0.337**	0.497	2.815
	(0.123)	(0.708)	(1.062)	(0.068)	(0.302)	(3.921)	(0.146)	(0.899)	(4.537)
P x % presence		1.573**	1.624*		0.024	-0.290		-0.333	-0.434
		(0.756)	(0.895)		(0.369)	(0.467)		(1.024)	(0.969)
P x % pres. x % within PO		0.956**	0.662		0.174	1.032*		0.626	0.752
		(0.435)	(0.683)		(0.396)	(0.579)		(0.985)	(1.168)
% presence, % within PO and their required interactions?	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Controls of table 10?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
% presence x Sex differentiation?	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Interactions between P and controls of table 10 and "% presence x Sex diff."?	No	No	Yes	No	No	Yes	No	No	Yes
Region dummies?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year of ethnography?	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Observations	74	74	74	149	149	149	96	96	96
R-Square	0.545	0.695	0.749	0.451	0.506	0.543	0.634	0.675	0.722

Notes: Robust standard errors are used in all regressions and displayed in parentheses. *** indicates p-value<0.01, ** indicates p-value<0.05, * indicates p-value<0.1. In models 2, 3, 5, 6, 7, and 8 there is high collinearity (this is frequent in models with interaction terms). Following Nunn and Giuliano (2017), in models 1 to 6 we added the square of ln(GDP per capita) and in models 4 to 9 we added the year of the ethnography and we excluded ethnographies that occurred before 1800 (many ethnographies are dated BC -these are drawn from historical studies). For FLP in 1970 we used the average of a five-year window around the year 1970 (the data is sparser in 1970 than in 2012).

2.5. Conclusion

In this article, we have developed a theory that explains the evolution of Productive Organizations (POs). We used a cultural evolution model to show that improvements in social learning within POs can favour the hard-to-propel process of cumulative culture. Under this account, POs evolved because they facilitate the transmission of knowledge between individuals, particularly if the PO specializes in a specific activity or technology. If access to POs is restricted, as is typical, then this knowledge transmission advantage leads to higher fitness of societies and therefore, to the selection and invasion of POs. We provide evidence from a sample of pre-modern societies that is largely consistent with the predictions of our model. The theory applies straightforwardly to pre-modern POs such as guilds and long-standing POs such as partnerships; as a descendant of these older POs, our theory also informs the origin of modern firms.

Our findings provide several main contributions. First, we are the first to show that social learning can be beneficial to society even if it does not generate a positive externality on individual learning (cf., Boyd and Richerson, 1995). Second, we provide an explanation for the origins of POs based on social learning, knowledge transmission and cultural accumulation. Mainstream theories of POs focus on governance and incentive, assuming a pre-determined pool of knowledge and culture. Knowledge issues have been addressed but mostly in relation to knowledge integration and problem solving (Garicano, 2000; Grant, 1996). Third, we provide a theory for the origins of specialization within POs that does not rely on trade and comparative advantage as the driving force; specialization evolves because it favours the social learning benefits of POs. Fourth, as our comparative statics and empirical results attest, our findings on the role of uncertainty, trust, and secrecy run counter the conventional wisdom. This suggest that an evolutionary lens changes the predictions that one would derive when the problem of knowledge and culture accumulation is not considered.

As with any trait that has been selected in a population, a full explanation of the nature of POs requires adding an evolutionary perspective to the mix. We need to consider not only the mechanisms that explain the inner workings and immediate benefits of POs, such as governance and protection from hazards, but also the evolutionary motive as to why they might have increase in frequency in the first place. As a first step into this direction, we hope that this paper stimulates further research on this important evolutionary foundation.

We can point to several limitations in our paper, all suitable for future research efforts. First, although we test our model on pre-modern societies using a good proxy for POs, it would be interesting to test the model using data on guilds or early partnerships. There are interesting new datasets that could be

used for this purpose (e.g., Comino et al, 2017). Second, the link of our theory with modern firms is derivative, mainly as descendant of early POs. Knowledge and technology in our model is transmitted across individuals, that is, we deal with accumulation of individual level traits. However modern firms combine specialized knowledge to generate complex technologies that are beyond the capacity of any single individual to produce or imitate. It would be very interesting to study how our model can be extended to study the evolutionary origin of modern firms. Third, the behaviour in our model is simplified to copying by social learners and "radical innovation" by individual learners. The model can be enhanced by introducing incremental innovation: agents could improve the technology while the state of nature remains unchanged. We suspect that the introduction of this element will further expand the beneficial impact that POs bring to society.

2.6. References

Alesina, A., Giuliano, P., 2015. "Culture and Institutions" Journal of Economic Literature 53 (4): 898-944.

Aldrich, H. 1999. Organizations Evolving. Sage Publications, London.

Allaire, Y., & Firsirotu, M. E. (1984). Theories of organizational culture. Organization studies, 5(3), 193-226.

Anderson, R., 1971 "Voluntary Associations in History". American Anthropologist, 73: 209-222

Apel, J. (2008). Knowledge, know-how and raw material-the production of Late Neolithic flint daggers in Scandinavia. Journal of archaeological Method and Theory, 15(1), 91-111

Argote, L., Miron-Spektor, E. 2011 "Organizational Learning: From Experience to Knowledge." Organization Science 22 (5): 1123-1137

Atalay, E., Hortacsu, A., Syverson, C. 2014. "Vertical Integration and Input Flows". The American Economic Review 104(4): 1120-1148.

Becker, A., Enke, B., Falk, A. 2018 "Ancient Origins of Global Variation in Social Preferences" NBER Working Paper No. 24291.

Benkler, Y., 2006. The wealth of networks: How social production transforms markets and freedom. Yale University Press.

Bisin, A., Verdier, T., 2011. "The Economics of Cultural Transmission and Socialization" In Handbook of social economics (Vol. 1, pp. 339-416). North-Holland.

Blau, P., Scott, R. 1963. *Formal Organizations: A Comparative Approach*. Routledge and Kegan Paul, London.

Boyd, R., Richerson, P.J. 1985. Culture and the Evolutionary Process. Chicago: University of Chicago Press.

Boyd, R., Richerson, P. J. 1995. "Why Does Culture Increase Human Adaptability?" Ethology and Sociobiology 16(2), 125-143.

Boyd, R., Richerson, P.J. 2005. The Origin and Evolution of Cultures. Oxford: Oxford University Press.

Boyd, R., Richerson, P. J., Henrich, J. 2013. "The Cultural Evolution of Technology: Facts and Theories." In Cultural Evolution: Society, Technology, Language, and Religion, ed. Peter J. Richerson and Morten H. Christiansen.

Bowles, S., 2009. "Did Warfare Among Ancestral Hunter-Gatherers Affect the Evolution of Human Social Behaviors?" Science 324 (5932): 1293-1298.

Burgess, R., Donaldson, D. 2010. "Can Openness Mitigate the Effects of Weather Shocks? Evidence from India's Famine Era 449-53." American Economic Review 100 (2): 449-453

Chandler, A. D. 1990. Scale and Scope: The Dynamics of Industrial Competition. Cambridge, MA, Harvard Business School.

Coase, R. 1937 "The Nature of the Firm." Economica 4 (16): 386-405.

Colwell, R. 1974. "Predictability, Constancy, and Contingency of Periodic Phenomena" Ecology 55 (5): 1148-1153.

Comino, S., Galasso, A., Graziano, C. 2017. "The Diffusion of New Institutions: Evidence from Renaissance Venice's Patent System" NBER Working Paper No. 24118

Costin, C. L. 2001. "Craft Production Systems". In *Archaeology at the millennium* (pp. 273-327). Springer, Boston, MA.

Coto-Sarmiento M., Rubio-Campillo, X., Remesal, J. 2018. Identifying social learning between Roman amphorae workshops through morphometric similarity, Journal of Archeological Science, 96: 117-123

Crook, R., Combs, J., Ketchen, D., Aguinis, H. 2013. "Organizing Around Transaction Costs: What Have we Learned and Where do we go From Here?" Academy of Management Perspectives 27 (1): 63-79

Dari-Mattiacci, G., Gelderblom, O., Jonker, J., Perotti, E. 2017. "The Emergence of the Corporate Form" Journal of Law, Economics and Organization 33 (2): 193-236.

De la Croix, D., Doepke, M., Mokyr., J. 2017. "Clans, Guilds, and Markets: Apprentice- ship Institutions and Growth in the Pre-Industrial Economy." The Quarterly Journal of Economics Forthcoming.

Dean, L.G., Kendal, R.L., Schapiro, S.J., Thierry, B., Laland, K. N. 2012. "Identification of the Social and Cognitive Processes Underlying Human Cumulative Culture." Science 335 (6072): 1114-1118.

Demsetz, H. 1988 "The Theory of the Firm Revisited" Journal of Law, Economics and Organization 4 (1): 141-162.

Depetris-Chauvin, E., Ozak, O, 2017. "The Origins and Long-Run Consequences of the Division of Labor" Working paper.

Enke, B. 2007. "Kinship Systems, Cooperation and the Evolution of Culture." Working paper.

Epstein, S. R. 1998. “Craft Guilds, Apprenticeship, and Technological Change in Preindustrial Europe.” Journal of Economic History 58(3): 684–713.

Epstein, S. R. 2008. “Craft Guilds in the Premodern Economy: A Discussion.” Economic History Review 61(1): 155–74.

Epstein, S. R., and Maarten Prak. 2008. “Introduction: Guilds, Innovation and the European Economy, 1400–1800.” In Guilds, Innovation and the European Economy, 1400–1800, edited by Stephan R. Epstein and Maarten Prak, 1–24. London: Routledge.

Espinosa, M. 2017. "Sourcing of Expertise and the Boundaries of the Firm: The Case of Lobbyists" Working paper, LSE.

Fogarty, L., Strimling, P., Laland, K. N. 2011. "The Evolution of Teaching." Evolution 65 (10): 2760-2770.

Forbes, S., Lederman, M. 2009. "Adaptation and Vertical Integration in the Airline Industry" American Economic Review 99 (5): 1831-1849.

Garicano, L. 2000 "Hierarchies and the Organization of Knowledge in Production" Journal of Political Economy 108 (5): 874-904.

Geyskens, I., Steenkamp, J., Kumar, N., 2006. "Make, Buy, or Ally: A Transaction Cost Theory Meta-Analysis". The Academy of Management Journal 49 (3): 519-543.

Gibbons, R., Henderson, R. 2012. "Relational Contracts and Organizational Capabilities." Organization Science 23 (5): 1350-1364.

Giorgi, S., Lockwood, C., & Glynn, M. A. (2015). The many faces of culture: Making sense of 30 years of research on culture in organization studies. *The academy of management annals*, 9(1), 1-54

Giuliano, P. 2016 "Review of Cultural Evolution: Society, Technology, Language, and Religion edited by Peter J. Richerson and Morten H. Christiansen" Journal of Economic Literature 54 (2): 522-533.

Giuliano P, Nunn N. 2017 "Understanding Cultural Persistence and Change." Working Paper

Grant, R. 1996. "Toward a Knowledge-Based Theory of the Firm." Strategic Management Journal, 17 (S2): 109-122.

Greif, Avner, Paul Milgrom, and Barry Weingast. 1994. "Coordination, Commitment, and Enforcement: The Case of the Merchant Guild." Journal of Political Economy 102(4): 912–50

Guinnane, T. W. , Harris, R., Lamoreaux, N., Rosenthal, J. 2007 "Putting the Corporation in its Place" Enterprise and Society 8: 687-729.

Hannan, M.T., Freeman, J. 1977. "The Population Ecology of Organizations." American Journal of Sociology 82(5): 929-964.

Hansmann, H., Kraakman, R., Squire, R. 2006. "Law and the Rise of the Firm." Harvard Law Review, 119: 1335-1403.

Hart, O., Moore, J. 1990. "Property Rights and the Nature of the Firm." Journal of Political Economy 98 (6): 1119-1158.

Haun, D., Over, H. 2013 "Like Me" In Cultural Evolution: Society, Technology, Language, and Religion, ed. Peter J. Richerson and Morten H. Christiansen.

Henrich, J. 2015. The Secret of Our Success: How Culture is Driving Human Evolution, Domesticating our Species, and Making us Smarter. Princeton University Press.

Hewlett, B., Roulette, C. 2016. "Teaching in Hunter-Gatherer Infancy". Royal Society open science 3 (1): 150403

Holmstrom, B., Milgrom, P. 1994. "The Firm as an Incentive System." The American Economic Review 84 (4): 972-991.

Josefy, M., Kuban, S., Ireland, R. D., Hitt, M. A. 2015. "All Things Great and Small: Organizational Size, Boundaries of the Firm, and a Changing Environment." Academy of Management Annals 9 (1): 715-802.

Kirby, K., Gray, R., Greenhill, S., Jordan, F., Gomes-Ng, S., Bibiko, H., Blasi, D. et al. 2016. "D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity." PloS one 11(7): e0158391.

Klepper, S. 1996. "Entry, Exit, Growth, and Innovation over the Product Life Cycle." The American Economic Review, 86(3): 562-583

Kogut, B., Zander, U. 1992. "Knowledge of the Firm, Combinative Capabilities and the Replication of Technology." Organization Science 3 (3): 383-397

Lafontaine, F., Slade, M. 2007. "Vertical Integration and Firm Boundaries: The Evidence". Journal of Economic Literature 45 (3): 629-685.

Laland, K. 2017. Darwin's Unfinished Symphony: How Culture Made the Human Mind. Princeton University Press

Lamoreaux, N.R., 1998. "Partnerships, Corporations, and the Theory of the Firm" The American Economic Review 88 (2): 66-71.

Levin, J and Tadelis,S.,2005 "Profit sharing and the role of professional Partnerships" The Quarterly Journal of Economics, Volume 120, Issue 1: 131-171,

Levinthal, D. 1997. "Adaptation on Rugged Landscapes." Management Science 43 (7): 934-950.

Levitt, S.D., List, J.A.,Syverson, C., 2013. "Toward an Understanding of Learning by Doing: Evidence from an Automobile Assembly Plant" Journal of Political Economy, 121 (4): 643-681.

Lewis, H. M., Laland, K. N. 2012. "Transmission Fidelity is the Key to the Build-up of Cumulative Culture." Philosophical Transactions of the Royal Society of London B: Biological Sciences 367 (1599): 2171-2180.

Lowie, R. 1948. *Social Organization*. New York Rinehart and Co.

Meek, V. L. (1988). Organizational culture: Origins and weaknesses. Organization studies, 9(4), 453-473.

Mokyr, J., 2016. A Culture of Growth: the Origins of the Modern Economy. Princeton University Press.

Morrison AD, Wilhelm WJ. "The Demise of Investment Banking Partnerships: Theory and Evidence." The Journal of Finance. 63 (1):311-50.

Murmann, J., Aldrich, H., Levinthal, D., Winter, D. 2003 "Evolutionary Thought in Management and Organization Theory at the Beginning of the New Millenium" Journal of Management Inquiry 12 (1): 22-40.

Nelson, R., Winter, S. 1982. An Evolutionary Theory of Economic Change. Belknap Press of Harvard University Press, Cambridge, Massachusetts

Nickerson, J., Zenger, T. 2004. "A Knowledge-Based Theory of the Firm - The Problem Solving Perspective." Organization Science 15 (6): 617-632.

North, D.C., 1991. "Institutions" Journal of Economic Perspectives 5 (1): 97-112.

Nowak, M. A. 2006. "Five Rules for the Evolution of Cooperation." Science 314 (5805): 1560-1563.

Nunn, N. 2009. "The Importance of History for Economic Development" Annual Review of Economics 1: 65-92.

Nunn, N., 2012. "Culture and the Historical Process" Economic History of Developing Regions, 27 (1): 108-126.

Ogilvie, S. 2014. "The Economics of Guilds." The Journal of Economic Perspectives, 28 (4): 169-192.

Ogilvie, S. 2019 The European Guilds: An Economic Analysis, The Princeton Economic History of the Western World - Princeton University Press

O'Reilly, C. A., & Chatman, J. A. (1996). Culture as social control: Corporations, cults, and commitment

Oster, E., 2016. "Unobservable Selection and Coefficient Stability: Theory and Evidence." Journal of Business €3 Economic Statistics DOI: 10.1080/07350015.2016.1227711

Passmore, S, Jordan, F. 2017. "Shared Ancestry Shows Kinship is Conserved." Pre- sented at the 1st Cultural Evolution Society Conference.

Padgett, J. F., McLean, P. D. 2006. "Organizational Invention and Elite Transformation: The Birth of Partnership Systems in Renaissance Florence." American journal of Sociology 111(5): 1463-1568

Poblete, J,. 2015 "Profit sharing and Market Structure" International Journal of Industrial Organization,39: 10-18

Rand, D. G., Nowak, M. A. 2013. "Human Cooperation." Trends in Cognitive Sciences 17 (8): 413-425

Rendell, L., Fogarty, L., Hoppitt, W., Morgan, T., Webster, M., Laland, K. 2011 "Cognitive Culture: Theoretical and Empirical Insights Into Social Learning Strategies." Trends in Cognitive Sciences 15(2): 68-76.

Rogers, A. R. 1988. "Does Biology Constrain Culture?" American Anthropologist 90 (4): 819-831.

Rowlands, M. J. 1971. "The Archaeological Interpretation of Prehistoric Metalworking" *World Archaeology* 3 (2): 210-224.

Schein, E. 2010. *Organizational culture and leadership*. John Wiley & Sons; 4th edition

Scott, R. 2003. *Organizations: Rational, Natural and Open Systems*. Fifth Edition, Prentice Hall.

Spolaore, E., Wacziarg, R. 2013. "How Deep are the Roots of Economic Development?" Journal of Economic Literature, 51 (2): 325-69.

Spulber, D., 2009 "The Theory of the Firm" Cambridge University Press.

Sterelny, K. 2012. *The Evolved Apprentice: How Evolution Made Humans Unique*. MIT Press.

Stock, J., Yogo, M. 2002. "Testing for Weak Instruments in Linear IV Regression" NBER Working Paper No. 284

Stout, D. 2002 "Skill and Cognition in Stone Tool Production" *Current Anthropology*,

43 (5): 693-720.

Tabellini, G., 2008 "Institutions and Culture" Journal of the European Economic Association 6 (2-3): 255 - 294.

Tadelis, S., Williamson, O. 2013. "Transaction cost economics" Chapter 4 of the Handbook of Organizational Economics (eds. Robert Gibbons and John Roberts).

Turchin, P. 2016. Ultra-Society: How 10,000 Years of War Made Humans the Greatest Cooperators on Earth. Beresta Books, LLC.

Weber, K., & Dacin, M. T. (2011). The cultural construction of organizational life: Introduction to the special issue. Organization Science, 22(2), 287-298

Williamson, O. E. 1991. "Comparative Economic Organization: The Analysis of Discrete Structural Alternatives." Administrative Science Quarterly 36 (2): 269-296.

Winter, S., Szulanski, G. 2001. "Replication as Strategy." Organization Science 12 (6): 730-743.

Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data. MIT press.

2.7. Appendix

Proof that there exists a unique ESS (r_I, r_s) and in that equilibrium $f_s = 1-C$. *The proof has two parts, we first show that there exists a unique Nash Equilibrium,*

moreover in this equilibrium $f_s = 1 - C$. *In the second part we show this equilibrium constitutes an ESS. There can't exist other ESS because ESS is a refinement of the Nash Equilibrium.*

Lemma 1: *In any Nash equilibrium both strategies are played.*

Proof. Suppose not and remember that $f_I = 1 - C$. If $r_I = 0$ then $f_S = -c < f_I = 1 - C$.So r_I must be positive in any equilibrium. If $r_I = 1$ then $f_S = (1 - c)(1 - p)$. In order for f_S to be larger than $f_I = (1 - C)$, it is needed that $(C - c) > p - pc$. Since this condition will always be met according to assumption 1, then r_I can never be one because deviating to r_S would be profitable. ■

Lemma 2: *There is a Unique Nash Equilibrium* (r_I^*, r_S^*).

Proof. By lemma 1, in equilibrium we must have $f_I = f_S$ but f_I is decreasing in r_I and f_S is increasing in r_I therefore there exists a unique r_I compatible with $f_I = f_S$ and therefore the equilibrium is unique. ■

Lemma 3: *The strategies played in the Nash Equilibrium* (r_I^*, r_S^*) *are ESS.*

Proof. Suppose a population of size $\delta < 1$ with a share of individual learners $r_I \neq r_I^*$ invades. 1) If r_I is larger than r_I^* then the resulting share of individual learners would be $\widetilde{r}_I = \frac{1}{1+\delta}(r_I^* + \delta r_I) > r_I^*$. This implies that in the new equilibrium the fitness of social learners is $\widetilde{f}_s > 1 - C$, because the fitness of social learners is increasing in the share of individual learners. The average fitness of the invading population would be $1 - C + (1 - r_I)(f_S - (1 - C))$ while the average fitness of the existing population would be $1 - C + (1 - r_I^*)(f_S - (1 - C))$, strictly larger because $r_I > r_I^*$. 2) If r_I^* is larger than r_I then the resulting share of individual learners would be $\widetilde{r}_I = \frac{1}{1+\delta}(r_I^* + \delta r_I) < r_I^*$. This implies that in the new equilibrium the fitness of social learners is $\widetilde{f}_s < 1 - C$, because the fitness of social learners is increasing in the share of individual learners. The average fitness of the invading population would be $1 - C - (1 - r_I)((1 - C) - \widetilde{f}_s)$ while the average fitness of the existing population would be $1 - C - (1 - r_I^*)((1 - C) - \widetilde{f})$, strictly lower because $r_I < r_I^*$. ■

Because every set of ESS are Nash, we know there exists a unique set of ESS for this game. QED

Proof of Proposition 1: T*he structure of the proof is identical to the previous one, including a PO of size* λ, *for a sufficiently small* λ.

Lemma 4 *In any Nash Equilibrium both strategies are played outside the PO.*

Proof. Suppose not and remember that $f_I = 1 - C$. If $r_I = 0$ then $f_S = \lambda\widetilde{q} - c < f_I = 1 - C$ if λ is sufficiently small. So r_I must be positive in any equilibrium. If $r_I = 1$ then $f_S = \lambda(\widetilde{q} - c) + (1 - \lambda)(1 - c)(1 - p)$. In order for f_S to be larger than $f_I = (1 - C)$ for every $\lambda > 0$, it is needed that $(C - c) > p - pc$. Since this condition will always be met

according to assumption 1, then r_I can never be one because deviating to r_S would be profitable. ■

Lemma 2: *There is a Unique Nash Equilibrium* (r_I^*, r_S^*) outise de *PO*.

Proof. By lemma 1, in equilibrium we must have $f_I = f_S$ but f_I is decreasing in r_I and f_S is increasing in r_I therefore there exists a unique r_I compatible with $f_I = f_S$ and therefore the equilibrium is unique. ■

Lemma 3: Inside the PO all agents play "social learning".

Proof. The fitness of individual learning is $\widetilde{f_I} = 1 - C$, the fitness of social learning inside the *PO* is given by $\widetilde{f_s} = f_s + \lambda(c - \widetilde{c}) = f_I + \lambda(c - \widetilde{c})$. (see equation 2). Because $\widetilde{f_s} > \widetilde{f_I}$ only social learning is played inside the firm. ■

Lemma 4: *The strategies played in the Nash Equilibrium* $(r_I^*, r_S^*, \widetilde{r}_S = 1)$ *are ESS.*

Proof. Suppose a population of size $\delta < 1$ invades. 1) If the invasion is outside the PO's to proof is analogous to lemma 3 of the previous proof and therefore omitted.If the invasion is inside the PO and $r_I \neq 0$, observe that average fitness of invading population is $r_I(1 - C) + (1 - r_I)f_S < f_S$, and thus it has lower fitness than population inside the PO. ■

Because every set of ESS are Nash, we know there exists a unique set of ESS for this game. QED

Proof of Proposition 2 *To prove this result, consider two societies with a productive organization of size* λ, *and assume society a has secrecy and society b does not. Inside the firm (both with and without secrecy), all agents are social learners and therefore in either case their fitness is given by*

$$\tilde{f}_s^i = [1 - \lambda][(1 - p)q^i - c] + \lambda[(1 - p)\tilde{q}^i - \tilde{c}]. \tag{7}$$

The recursive equations that determine the stock of knowledge inside the firm at the steady state is the same regardless wether there is secrecy or not and is determined by

$$\tilde{q}^i(t) = [\lambda\tilde{q}^i(t - 1) + (1 - \lambda)q^i(t - 1)](1 - p). \tag{8}$$

From equations (7) and (8) it is clear that fitness in the firm is strictly increasing on the stock of knowledge outside the firm. Therefore to show that secrecy is detrimental to the firm's fitness we only need to show that secrecy is detrimental to the stock of knowledge oustide the firm (ie that $q^b > q^a$*).*

To show this, is useful to define the function $\tilde{q}^i(q^i)$ *, which is the stock on knowledge inside the firm , a function increasing in the stock of knowledge outside the where* $\tilde{q}^i(q^i) < q^i$.

We know that

$$f_S^a = (1 - p)q^a - c = 1 - C$$

$$f_S^b = (1-\lambda)(1-p)q^b + \lambda(1-p)\tilde{q}^b(q^b) - c = 1 - C$$

To the contrary, suppose that $q(a) \geq q(b)$*; then*

$$1 - C = (1-p)q^a - c > (1-\lambda)(1-p)q^b + \lambda(1-p)\tilde{q}^b(q^b) - c = 1 - C$$

a contradiction. Therefore it must be that $q(b) > q(a)$.QED

Proof of Proposition 3. *Observe from equation (6) that the fitnesss of a social learner that specializes in technology* j*, can be expressed as.*

$$\widetilde{f_S^j} = f_S^j + (c - \widetilde{c}) \cdot \frac{\lambda \widetilde{x}^j}{\lambda \widetilde{x}^j + (1-\lambda)x^j}. \tag{9}$$

Remember that in equilibrium f_S^j *and* x^j *must be constant across* J *outside the productive organization. It is straightforward to check that there are two Nash equilibrium inside the PO. 1) Either* $\widetilde{x}^j$ *is the same for every* j *in* J*, in which case* $\widetilde{f_S^j}$ *is constant in* j *or 2) There is full specialization and* $\widetilde{x}^j = 1$ *for* $j = j^*$ *and* 0 *otherwise.*

Observe now that the first equilibrium is not ESS because if a small population of size δ *that specializes in one activity* $\widetilde{j}$ *invades the PO, then the fitness of the invading population becomes*

$$\widetilde{\widetilde{f_S^j}} = f_S^j + (c - \widetilde{c}) \cdot \frac{\lambda \widetilde{x}^j + d}{\lambda \widetilde{x}^j + (1-\lambda)x^j + d}$$

While average fitness of the average population is

$$E_{j \in J}(\widetilde{f_S^j}) = f_S^j + (c - \widetilde{c}) \cdot \left[\frac{J-1}{J}\right]\left[\frac{\lambda \widetilde{x}^j}{\lambda \widetilde{x}^j + (1-\lambda)x^j + d}\right] + \left[\frac{1}{J}\right]\widetilde{\widetilde{f_S^j}} < \widetilde{f_S^j}.$$

To see that a specialized Nash is ESS, notice that in a specialized equilibrium , fitness inside the PO is given by

$$\widetilde{f_S^{j^*}} = f_S^{j^*} + (c - \widetilde{c}) \cdot \frac{\lambda}{\lambda + (1-\lambda)x^j},$$

And observe that an invasion of size δ *by any other technology will obtain a fitness of*

$$\widetilde{\widetilde{f_S^j}} = f_S^j + (c - \widetilde{c}) \cdot \frac{\delta}{\lambda + (1-\lambda)x^j + \delta},$$

Which converges to $f_S^j < \widetilde{f_S^{j^*}}$ *as* δ *approces 0, and therefore the invading population obtains a lower fitness level.* QED.

3. The Evolution and Impact of Cooperation in Large Groups: Evidence from Administrative Data and a Field Experiment

Abstract

We study the adoption of cooperation in large groups, defined as voluntarily incurring a cost in order to benefit other group members or the group at large. Cooperation is vulnerable to group size, as the benefits of cooperation are diluted over more members while the cost to the cooperator stays the same; that is, cooperation is a social dilemma. A large body of theoretical research in evolutionary biology, anthropology and economics has proposed mechanisms that favour the evolution of cooperation, however, empirical evidence comes mostly from lab experiments with scant evidence from the field. In addition, these mechanisms haven't informed much collaboration research within management. In this paper, we study a workplace safety methodology in which an initial small group of 10 workers is trained first to observe and counsel co-workers on safe behaviour, and then to expand the group by enrolling and training new workers to become observers within the implemented site (e.g., a plant). The methodology leverages cooperation: training and counselling is costly to observers while the benefits of improved safety flows mostly to the observed workers. First, we use archival data from a representative sample of 88 implementations to show that the methodology is effective -- it reduces accidents and improves safety culture, which in turn improves other aspects of culture such as team spirit and the workers' relationship with the organization; however, this positive impact decreases as the number of observers expands beyond 25 about observers -- the data empirically confirm the cooperation breakdown in large groups. Specifically, we show that the cooperation decrease comes from a substantial lower and less sustained cooperative effort by the additional enrolled observers as the observer group increases. Second, we conduct a field experiment in four sites where we manipulate the safety methodology with three treatments aimed at mitigating the cooperation breakdown. We show that: i) the effort of the additional observers is restored when the expansion of observers is structured around small groups (1st treatment – "direct reciprocity"), ii) partially lifting the anonymity of the observed workers is detrimental to observers' effort (2nd treatment – "identity"), and iii) public display of cooperative effort of observers does not change effort (3rd treatment – "reputation") --- but interacts with the 'private enforcement' –measured with administrative data– in subtle ways. We find that these treatment effects on cooperation effort indeed matter: they modify the speed of diffusion of cooperation and the incidence of risky behaviour and accidents in the workforce.

Overall, our study provides novel field evidence of the cooperation breakdown when groups grow large, as well as of a group structure design that supports its recovery. The result on anonymity suggests a not-yet-explored angle in cooperation research: if the cooperation benefit entails pointing at "mistaken" behaviour (i.e., behaviour causing safety risks), then transparency might reduce cooperation (and anonymity be preferred), in contrast to standard prescriptions in the cooperation literature (e.g., indirect reciprocity).

KEYWORDS: Cooperation, Field Experiment, Culture, Evolution, Direct Reciprocity, Workplace Safety

3.1. Introduction

Achieving and sustaining cooperation in large groups is a crucial ingredient in the success of divisions, departments and whole companies. Given that organizations consist of complementary assets, processes and tasks (Argyres and Zenger, 2012; Milgrom and Roberts, 1995), the willingness of organizational members to exert consummate effort that benefits the group and co-workers is a basic condition for effective integration of efforts and high performance (Gibbons and Henderson, 2013; Fehr, 2018; Organ et al, 2005; Puranam, 2018). In addition, several valuable intangible assets, such as the firm's reputation and brand, require collective cooperative effort to be built and maintained. Barnard (1938) long argued that a central role of the CEO is to engineer cooperation in the organization. Consistently, research has documented a strong positive association between cooperative behaviour of workers and the performance of their units/firms (Podsakoff et al, 2009); Grennan (2014) is a rare study that establishes a causal link.

High levels of cooperation within organizations is hard to achieve due to several motives. First, cooperation poses a social dilemma: cooperative behaviour benefits the group, but individuals face an incentive to free-ride, that is, to enjoy the benefits of others' cooperation without incurring the costs of cooperation him/herself. Second, consummate cooperative behaviour is relational and discretionary in nature, and therefore it lies largely beyond the direct control of basic managerial levers –such as top down monitoring or formal contracting (Gibbons and Henderson, 2012 and 2013; Organ, Podsakoff and Mackenzie, 2005). In simple, while perfunctory cooperation can be enforced, consummate cooperation is voluntary. Third, the degree of cooperation in groups is a self-enforcing equilibrium, making it stable and hard-to-change. This is the condition that makes cooperation a key driver of persistent-performance-differentials among seemingly-similar-enterprises (Gibbons, 2006)

The importance and difficulty of achieving high cooperation is reflected in senior executives. In a survey of CFOs/CEOs in 1348 large US Firms (Graham et al, 2018), responses indicate that while

cooperation among workers is the main antecedent to an effective culture and that improving their culture would increase their firm's value, only 16% believe their culture is where it should be. Put simply, CEOs seek cooperation in their workforce but struggle with it.

The literature has pointed at some drivers of voluntary, consummate cooperation in large groups[39]: the role of leaders as examples, guides and enforcers (Barnard, 1938; Schein, 2010; Kosfeld and Rustagi, 2015; Hermalin, 2013); promoting the identification of workers with the organization (Akerlof and Kranton, 2005); eliciting joint goal motivation using symbolic management and organizational design (Lindenberg and Foss, 2015); firm-wide financial incentives coupled with small groups (Knez and Simester, 2001); modifying the relational architecture of jobs such as connecting employees with internal and external beneficiaries (Grant, 2007); and governance that focus on the long term, not the short-term shareholder benefits (Grennan, 2014).

In this paper, we use the notion of an "interaction structure" from evolutionary biology/anthropology (Rand and Nowak, 2013) to inform this issue and guide our archival empirical analysis and the design of our experiment. For cooperation to evolve, whether in nature or society, a mechanism is required that allows favouring cooperators over defectors (Nowak, 2006). In the models of this literature, a mechanism is an interaction structure that specifies *who interacts with whom* in a population –these interactions could be random or structured, and on the latter, the structure could be fixed or flexible– and *how* the agents interact in order to receive payoffs what –for example, information availability, degree of repetition (before random reshuffling), order of play, details of payoff functions, enforcement technology. In this paper we focus on a particular condition, the repetition of play. When this is the case, the player in a social dilemma can condition its behaviour on the past behaviour of the other player(s). There are many ways of conditioning –known as "strategies"– with tit-for-tat being an excellent strategy for sparking cooperation in repeated prisoners' dilemma (Axelrod and Hamilton, 1981; Rand and Nowak, 2013). This strategy starts cooperating and then copies the other player's previous move: cooperate if cooperation was experienced, defect is defection was experienced. Other successful strategies (e.g., grim, win-stay-loose-shift, generous tit-for-tat) share this condition of reciprocating the other player's move: cooperate but punishing defection by withdrawing cooperation.

[39] In small teams, decades of research have unveiled important parts of the cooperative chemistry (Mathieu, 2008). Plenty of research has been done in pairwise infinitely repeated prisoners' dilemma (see Dal Bo and Frechette (2018) for a review). In organizational economics, attention has been given to relational contracts between the principal-agent dyad, both within and across the firm (e.g., Gibbons et al, 2002; Halac, 2011; Chassang, 2010). In strategic management, attention has concentrated on trust and cooperation in repeated interfirm relations (e.g., Vanneste et al, 2014; Elfenbein and Zenger, 2013).

Hence, the mechanism associated with repetition of contact is referred to as "direct reciprocity"[40]. However, in repeated public goods games – where interactions occur over a N-player prisoner dilemma – direct reciprocity breaks down very quickly as the group grows (Boyd and Richerson, 1988). In a public goods scenario, punishing non-cooperators by withholding cooperation not only punishes defector(s), it also punishes co-operators; this makes direct reciprocity inefficient.

Repeated public good games are ever-present in organizations; therefore, understanding how direct reciprocity can be rescued is important for cooperation in organizations. First, direct reciprocity has a higher chance to favour the evolution of cooperation in public good games with continuous choice (instead of the binary choice cooperate-defect) (Takezawa and Price, 2010). Second, prior research has shown that if pairwise targeted reciprocity is introduced in between rounds of the public good game – in the form of costly punishment or reward to a defector– (e.g., Fehr and Gachter, 2000; Rand et al, 2009; Gürerk et al, 2006), then reciprocity regains its power to sustain cooperation in large groups. However, in many cases, such as noisy observability of consummate cooperation, judgement about behaviour will be dichotomized and rewards can be costly or non-feasible. In addition, punishment is frequently inefficient, as it leads to lower total payoff, retaliation spirals or antisocial punishment (Rand et al, 2009; Hermann et al, 2008).

We focus on a simpler solution, the role of formal structure of the organization. An important part of the organizational formal structure defines the grouping of organizational members (i.e., defining teams, areas, divisions). We argue that even grouping can generate a boost in the degree of repeated interactions. For example, if a population of 50 workers get arranged into groups of 5 –and interactions are bound to the group– then, even if grouping and interactions are random, repetition of contact would increase: if players are playing a pairwise prisoners' dilemma, repeated interaction increases by a factor of by a factor of 10; If players are engaged in a public goods game, simulation shows that the odds that cooperation evolves increase exponentially from 50 to 5 (Boyd and Richerson, 1988). This intuition is supported by research: in public goods setting, the odds of direct reciprocity to support cooperation

[40] As indicated above, models repeated interactions and relational contracting in organizational economics and strategic management uses dyadic relations with a principal-agent. Given that we seek to understand large group cooperation, we instead draw on models from evolutionary biology/anthropology. The main advantage using these models is that they brings a population point of view with many agents interacting, which is exactly what is required when addressing large scale cooperation (or any other group level behaviour such as culture). This comes at a cost of a simplified view of the capacities of interacting human agents (i.e., no foresight and fixed strategies in evolutionary game theory), plus an oversimplification of the replication dynamics for social traits. However, at least on the first issue, a long research pedigree in organizational research has been founded precisely on taking these assumptions seriously and working out their consequences (Cyert and March, 1963).

are higher if the population is structured (Boyd and Richerson, 1988), particularly if cooperation is continuous (Takezawa and Price, 2010). The same has been found to be true for a population playing pairwise prisoners' dilemma: van Veelen et al (2012) finds that "a strong dose of repetition and a pinch of population structure" (p. 5) is crucial for cooperation to evolve and that these "findings are noteworthy because human interactions are typically repeated and occur in the context of population structure" (p. 5).[41]

The role of formal organization in solving social dilemmas informs a nascent literature that explores how formal organization impacts informal organization within firms (McEvily et al, 2014; Clement and Puranam, 2018). Our results are particularly complementary to Clement and Puranam (2018); while they show that a minimal and random formal organization helps agents in finding valuable interactions by saving coordination costs, we show that it can also favour cooperation among agents. Together, a compelling theory for organizational structure emerges, that is, the facilitation of coordination and cooperation, the two underpinnings of collaboration.

To empirically explore these issues, we collaborated with DEKRA Insight, a global workplace safety firm, by studying their BAPP methodology. This methodology consists of training a group of 10 workers of a site (e.g., a plant, a warehouse, a store) in observing how their colleagues execute their tasks and then giving them safety feedback. The identity of the observed worker remains anonymous. Then, this group of 'observers' expands within the site by enrolling and training other workers to become observers themselves. Thus, the starting group of observers might become a group of several tens, even above a hundred. Becoming an observer and executing observations is a cooperative act: getting trained and observing workers takes time and effort – on average 5% of their time – and the benefits flow mostly to the workers receiving the feedback. BAPP is voluntary, workers are not obliged to contribute or to remain observers, and it is "for the workers by the workers", with minimal intervention of the managers and supervisors of the site. 'Who observes whom' and 'who becomes an observer' is not structured by the methodology; observing and recruiting are executed in a free form by sites with many sites doing it quasi-randomly. BAPP only specifies the goal of ideally, over time, having all the workers of the site being observed once a month. The absence of an "interaction

[41] Population structure is another independent mechanism driving the evolution of cooperation (Nowak, 2006). This mechanism doesn't require strategic behavior (i.e., players have fixed strategies), and include models of "spatial/network" selection (Nowak and Sigmund, 1992; Ohtsuki et al, 2008) and "group selection" (Traulsen and Nowak, 2006).

structure" in observing and enrolling workers suggests that the expansion of cooperators (observers) and the amount of the effort they exert might be systematically limited.

To study BAPP, we had access to a dataset with a representative sample of 88 implementations. We first document that BAPP indeed promotes cooperation: observers expand within the site reaching 20% of workers in the third year, cooperative effort is exerted, and workers end up being observed once a month. We find that BAPP is associated with a substantial decrease in accidents. However, consistent with the absence of an "interaction structure" we document that the impact of BAPP suffers as the number of observers expands, especially after 20 workers. The additional observers that are enrolled in a site conduct substantially lower numbers of observations and display higher rotation. Exploiting the natural variance in implementations, we also find that BAPP's impact is higher when observers focus on particular areas of the site and therefore interact with fewer workers repeatedly (facilitating direct reciprocity). Overall, cooperation through BAPP is powerful and collectively beneficial, but without an interaction structure in place, its spread and impact are limited by free-riding.

The second part of this study consists of a field experiment. With the support of DEKRA, we collaborated with ACHS –one of the three private non-profit organizations in Chile that provide workplace services to Chilean companies (prevention, medical services, compensations and pensions)– and with one of ACHS's clients, SODIMAC –a Chilean multinational that operates a home improvements stores. We introduced experimental variations into the BAPP implementations taking place in SODIMAC. We conducted three treatments. In the first treatment we randomly assigned half of the workers in the site, typically 250, to 5 groups of 25 workers, which then were assigned to randomly selected 5 observers (out of 10 observers that initiate BAPP in the site). These five observers were restricted to execute observations within their respective groups. The remaining workers and observers constituted the control group, where the BAPP was executed as usual, without structure. New observers that were enrolled from the workforce were also bound to this structure (if a worker of group "a" became observer, he was bound to observe within that group). This treatment generated approximately a fivefold increase in the amount of repeated interactions between a specific observer and a specific worker (as compared to control). In addition, the treatment also led to a dramatic reduction of the group size over which observers play a public goods game of consummate cooperative effort. With the group structure, the public good game is played among few observers, only those within a group (about 3 in our experiment, as compared to approximately 30 in control). This dramatically improves the chances of direct reciprocity in the public good game (Boyd and Richerson, 1988).

We find that this treatment was highly effective: it increased in the number of observations and the received coaching, especially for new observers. Outcomes were also affected by this treatment: the workers within the groups displayed less risky behaviour and a lower likelihood of having an accident. Finally, the workers in the treatment increased the likelihood of becoming observers themselves. The magnitudes of all these estimates are economically meaningful. As a whole, these results suggest that putting in place an "interaction structure" that facilitates direct reciprocity, improves the spread of cooperation, the effort exerted by cooperators and the ultimate impact on outcomes. End-of-experiment interviews executed in the field provided confirming evidence for the mechanisms and our interpretations of the findings, especially for the role of increased reciprocity between observer and worker.

The second and third treatments are 'interaction effects', designed to boost the impact of treatment 1. In the second treatment, we named the groups of treatment 1 and revealed the names of group members within the group. The idea was to promote group identity by creating "minimal groups" that have common knowledge of group membership (Tajfel, 1970 and 1982; Guala et al, 2013; Goette et al, 2006). This treatment also serves to probe the main alternative explanation to the findings in treatment 1: it could be that psychological predisposition to small teams and identification processes, and not direct reciprocity, are driving the results. In the third treatment, the number of observations executed by observers was made public in two stores in the form of posted lists, in which observers were ranked in decreasing order. This treatment tapped into the reputation concerns of observers, possibly triggering the mechanism of "indirect reciprocity" (Nowak and Sigmund, 2006): I cooperate (defect) with you, if I observe that you have cooperated (defected) in the past. This strategy is useful when interactions are not repeated, but reputation (i.e., previous actions) is observable. Roberts (2008) studies the interaction between of indirect reciprocity and direct reciprocity and shows that under high repetition of contact, direct reciprocity will dominate.

Treatment 2 reverted the benefits of treatment 1. This treatment reduced observations and the likelihood of becoming an observer, and it increased risky behaviour and accidents by workers. Exit interviews strongly suggested that this treatment lifted the anonymity condition of observations, generating additional costs for workers in terms of suspicion and distaste for surveillance and blame. The treatment clashed with the motto of BAPP ("no spying, no naming, no blaming") and its voluntary character. We executed several tests that confirmed this interpretation and ruled out alternative explanations. This result raises an interesting new angle for the cooperation literature: when the benefit that is provided entails pointing at and correcting erroneous behaviours, anonymity might be necessary.

Treatment 3 did not affect the impact of treatment 1. This null effect is be consistent with the prediction of Roberts (2008): when repeated interactions are high, using private information (direct reciprocity) dominates using public information (indirect reciprocity). Nonetheless, post-hoc revealed interesting reputational dynamics. Our exit interviews indicated that the observers' effort was displayed in meetings, generating informal peer control; thus, treatment 3 simply made public what was already known privately within the groups. We developed a measure of private reputational enforcement and we found it had a positive impact on cooperative effort. We found that treatment 3 decreased the impact of private enforcement, suggesting that a public reputation mechanism can substitute for private enforcement.

Overall, this study contributes to research on the determinants of successful collaboration in organizations (Puranam, 2018; Gibbons, 2018). We study the anatomy of large scale cooperation breakdown and recovery "in-the-wild", complementing the dominance of lab evidence in the literature (Kraft-Todd et al, 2015; Nowak and Rand, 2013; Dal Bo and Frechette, 2018; Balliet and Van Lange, 2013). Perhaps more importantly, we illustrate how the concept of "interaction structure" from evolutionary biology/anthropology informs the role the formal structure of organizations have.

The rest of the study is organized as follows. Section 2 provides a detailed description of the BAPP method. Section 3 provides evidence using a large sample of previous implementations. Section 4 introduces and analyses our field experiment. Section 5 concludes.

3.2. BAPP Methodology

We collaborated with DEKRA Insight, a company specialized in workplace safety prevention. One of the services that DEKRA provides to its clients is BAPP (Behavioural Accident Prevention Process)[42]. BAPP is a methodology based on co-worker feedback that seeks to improve workplace safety among the employees of a treated site, such as a plant, a store, or a warehouse. Sites are typically large, employing at least 50 employees. This methodology started in the late 1980s and it has been implemented across the world and in many different industries. The BAPP methodology works as follows:

[42] The BAPP methodology was originally developed by BST (Behavioural Science Technologies), the company that was acquired by DEKRA Insight in 2012.

- In the first month, DEKRA consultants meet the site manager and lay out an implementation plan. In the second month, a focus groups and a culture survey are performed in order to assess in detail the status of the site.

- In the third month, a team of 8 to 12 employees (depending on the site's size) is trained by the consultants to become "observers". The selection of employees doesn't follow a pre-defined criteria, other than striving to be representative of the site and focusing on front-line workers (no supervisors or managers are selected). The selection process is steered by the manager of the site and the consultant. Importantly, becoming an "observer" is voluntary.

- A leader of the team is selected (ideally) by consent between the manager of the site, the consultant and the workers of the team. This leader takes on the role of being the "BAPP enabler". Unlike the rest of the team members, the enabler is 100% devoted to the project. The enabler reports directly to the site manager, and he is the direct owner of the implementation of the methodology. The site manager has the role of being the "sponsor" of the BAPP initiative, which includes advocating for and supporting the initiative, providing resources, and participating in the barriers removal committee (see below).

- Over the course of BAPP, the enabler and the team meet regularly, typically once a month, in a "BAPP committee". This committee is led by the enabler, and its purpose is to track progress, assess challenges, define priorities, plan ahead and propose barrier removals.

- In the third month, the consultant and the team develop an inventory of critical behaviours in terms of safety (known as CBI, "critical behavior inventory"). The behaviours are adapted to their site and the inventory typically includes around 25 behaviours (e.g., placing your body in front of the line of fire, not using the safety equipment, clotted workspace).

- On the fourth month, the observers receive training on how to provide constructive feedback to a worker of the site on safe working behaviour. This feedback event is known as an "observation", which is the core unit of BAPP. After training, observers start immediately executing real observations of the site's workforce.

- An observation consists of approaching the worker and, after his consent, observing its behaviour for 10 to 20 minutes. A detailed observation sheet is filled during the observation. This sheet contains at the top general information as date, place of the site, time of the day, and presence of coaching (to be explained below). At the bottom, the list of critical behaviours with a space to indicate whether it was observed as safe or risky. If a risky behaviour is identified, verbal feedback is then provided to the

observed worker. The sheet provides space to provide written details about the behaviour and the interaction with the worker.

- Only front-line workers are subject of observation; this makes BAPP a method that is "by the workers, for the workers". BAPP doesn't establish any criteria regarding which observer observes which workers of the site. This decision, if any, is discussed and worked over time by the enabler and the joint committee.

- DEKRA stresses that a very important feature of BAPP is anonymity, meaning that the identity of the observed worker remains anonymous, it is never recorded by the observer in any shape or form. This is told to workers in advance. Related to this, observers do not "spy", they are open and frank about the act of observing a worker. BAPP has a frequently repeated mantra: "no spying, no name, no blame".

- All the information of the sheets is uploaded to a local data storage system. This feeds the analysis that is performed by the enabler and presented discussed at the committee (e.g., number of observations per observer, number and place of risky behaviours).

- In the fifth month, the consultant monitor the progress of observers as they execute observations in the site. In addition, the workers are trained to become trainers themselves. From the sixth month onwards, the methodology establishes that new workers are enrolled and trained to become observers. This training is executed by the members of the committee.

- The selection of new workers to become observers is similar to the selection of the committee members: it is unstructured and BAPP doesn't impose any criteria, except for the fact that only front-line workers are allowed. The site manager, the enabler and the committee, they are all able to participate; and is voluntary, it cannot be imposed on workers.

- The new observers do not become part of the committee, so they don't attend progress meetings. BAPP doesn't specify a structure of engagement for these new observers. Instead, the enabler (and to some extent the committee) organically defines and executes a way to relate to the new observers.

- In addition to observations, observers also perform coaching. Coaching consist on observing a fellow observer execute an observation and then providing feedback and suggestions for improvement to him. Coaching is provided mostly by committee members with the occasional coaching by non-committee observers. On average, 10% to 20% of observations are executed with a coach.

- Between the 6th and the 12th month, the consultant accompanies the enabler and committee and helps them in: ramping up observations, enrolling new observers, analysing the accumulated sheets data to spot patterns and problems, monitoring observation quality. In the month 12th the consultant performs a sustainability review and report after which the site is left to their own devices. From month 3 to month 12, each observer of the committee receives around 18 days of training by consultants.

- As observations accumulate, systematic barriers to safe behaviour can be identified. A barrier is any impediment to safe behaviour, from lack of safety gear to a managerial practice that is compromising safety. To deal with this, BAPP includes setting up a barrier removal committee which has ad-hoc meetings to decide and take action upon barrier proposals. This committee is constituted by the site's leader, the enabler and (some) area managers of the site.

- The goal of BAPP is to generate good quality observations in the site. A key variable that is tracked is the "contact rate" which is equal to the number of observations in a month divided by the number of workers in the site on that month. The unwritten rule of thumb of BAPP consultants –which is a widely accepted and communicated– is that sites should reach a contact rate of 1 over time (a specific time is not specified as implementations may vary). Given that BAPP is typically applied to sizeable sites, on average 250 workers, getting to a contact rate of 1 requires a fair share of new observers. Without new observers a committee of 10 would need to execute 25 observations per month (plus coaching) which is difficult to accomplish. If 40 new observers were added to the committee, then 5 observations per month would be needed per observer, which is feasible. However, there are many possible other combinations. BAPP methodology doesn't specify nor recommends an execution strategy in terms of number and timing of new observers (nor, as indicated above, specifies how new observer should be incorporated to the process). In practice, there is a lot of variance in the number of new observers and in the observations per observer.

- Over the course of BAPP, being an observer entails: participating in regular meetings (for committee members), training/coaching new enrolled observers (for committee members), executing coaching (mostly to committee members) and executing observations.

- While being part of BAPP, the employees continue to execute their regular work in the site. BAPP is an additional activity that they execute. For observers that are part of the committee, DEKRA estimates that approximately during the first year, 8% of a worker's time is devoted to BAPP. After that, that figure drops to 5%. Non-committee observers spend a bit less, 3% to 5% on average. Although Sites attempt to provide flexibility to workers; however, this is not always achieved, leading to possible role tensions.

3.2.1. BAPP is an ideal setting to study the evolution of cooperation

This setting is well suited to study the cooperation in the field. First, BAPP requires that observers devote time and effort in order to provide a feedback to colleagues. Here the observer bears a cost while the benefits of decreased likelihood of accident flows (mostly) to the worker that received the feedback. This fits into the textbook definition of a cooperative act. (Fellow observers may also receive observations, but as we show below, this doesn't eliminate the cooperative/social dilemma in BAPP.)

Second, the number of observers is sought to grow and expand in the site; in other words, the setting allows to study how cooperative trait increases in frequency in a population. This evolution problem is a central to theoretical and experimental literature on cooperation (Nowak, 2006).

There are three types of social dilemmas occurring in this setting: a public good game between workers in becoming observers; a public good game between observers in contributing to the success of BAPP; and a prisoner's dilemma between an observer and a worker when an observation is conducted. As we explain below in more detail, these games are not independent of each other. Their introduction/distinction serves a practical purpose: it helps delineating and thinking about all the different interactions in BAPP that are subject to social-dilemma tensions, and many results we display in our empirical analysis of sections 3.3 and 3.4 can be better interpreted and understood if we refer to a specific game.

The **first social dilemma** occurs among workers in their decisions of becoming (or not becoming) observers, both when workers are approached to be part of the initial team and also when BAPP grows within the site. This interaction conforms to the structure of a public good game because of the following elements:

Cooperation (exposes the worker to benefit and cost): Becoming an observer and execute observations of other workers.

Defection (exposes the worker to only the benefit): Continue as a worker.

Benefit (flows to everyone): Safety feedback from observations (observers also receive observations).[43] (This matters because BAPP does meaningfully reduce accidents, as we demonstrate below.)

Cost (of cooperating): Effort and time in getting trained and in executing observations, minus the "reputational" benefits to observers for being part of BAPP[44].

The **second social dilemma** occurs among the group of observers that are in place at any point in time during the implementation. This can also be identified as a public good game by the following elements:

Cooperation: Consummate (high) effort in the number and quality of observations (and coaching).

Defection: Perfunctory (low) effort in the number and quality of observations (and coaching).

Benefit: Safety feedback from observations (observers also receive observations), plus the "reputational" benefits to observers for being part of BAPP.

Cost: Additional effort and time in executing consummate effort (more and better observations), minus the marginal "reputational" benefits to observers for implementing BAPP in a consummate way.

In the section 3.3.8 we provide estimations of these costs and benefits and show that these two games are very likely a public goods dilemma. However, we also show that the "reputational" benefits, which are hard to estimate, might generate a relaxed social dilemma such as the snowdrift game. Below we

[43] This benefit is the contact rate (total observations divided by total workers) and the reduction that it generates in on accidents. This number is 0.5 on the 6th month of observations, and approaches 1 by the end of the third year in the average BAPP implementation (see Figure 7 in page 12). In section 3.3.3 we estimate the reduction in accidents. Although that estimation is based on ex-post values, one could expect "a priori" that the expected benefits would be similar to the past experiences of BAPP.

[44] BAPP is very costly to execute as firms pay considerable sums to implement it. Therefore, site managers –who are sponsors of BAPP– are typically invested and hold part of the responsibility of BAPP's success. This means that the implementation of BAPP has non-negligible relevance in the site. This generates "signaling" or "reputational" benefits to be an observer which can flow immediately as status or recognition, or in the future as reciprocity from colleagues or potential advancements in their careers. Importantly, these benefits are more salient or clear to observers that part of the committee than to the additional new observers. For the latter, the benefits also decrease with the number of new observers. For example, career benefits quickly exhaust with more observers: there aren't that many job vacancies within the site to promote all them. This puts a strong limit to the potential career benefits flowing to observers. This means that the social dilemma is more acute for additional/new observers.

detail how this might affect the interpretation of the results of section 3.3 and the predictions/results of the section 3.4.

The **third social dilemma** occurs between the observer and the worker being approached for an observation. This interaction can be identified as a prisoner's dilemma by the following elements:

Cooperation by the observer: Frequent observations of high quality.

Defection by the observer: Few (or no) observations of bad quality.

Cooperation by the worker: Accepting being observed – workers can reject being observed – and if observed, be open, be engaged, and be willing to change his/her behaviour.

Defection by the worker: Dismiss the observation, or be unengaged and unwilling to change his/her behaviour.

Costs and benefits: These are more diffuse for this third dilemma. For an observer, the execution of observations generates a contribution to the team of BAPP observers in achieving their goals, and from it some "reputational" benefits may flow (e.g., status, recognition, career prospects). The cost is the same as before, effort and time of frequent consummate observations. For the worker being observed, the benefit is the safety feedback he receives. The cost is a combination of: i) any annoyance or perceived risk that is generated by being observed while executing his/her work (e.g., they are "spying on my behaviour," workers here (as elsewhere) are always nervous about looking bad or being criticized), ii) devoting time to the feedback, iii) changing his/her behaviour (this can be quite costly if behaviour is engrained and routinized, or imposes adjustment costs on other interacting workers). Also, as part of the costs one could add the "moral debt" that is acquired by being observed. This debt can eventually be collected, for example, when asked to be an observer.

Of course, these games are endogenous. Cooperation in game 2 and 3 is the same decision for the observer. Or consider game 1 and 2: If a worker deciding whether to cooperate in game 1 thinks that the cooperation is low in game 2, he/she will have a lower expected benefit and thus will be less willing to cooperate. This could happen, for example, if the number of observers is already high, and thus cooperation in game 2 is harder. At the same time, an observer deciding whether to cooperate in game 2 will face different incentives to cooperate depending on the equilibrium of the 1st game.

The impact of group size on cooperation. Given that we explore in this study the impact of organization size on cooperation, we now discuss how the number of observers affects these games. These are not predictions, but an exploration of what one might expect. In the empirical section we use the foundation from this discussion to more incisively interpret the results.

The outcome of the game 1 is the number of workers cooperating voluntarily, that is, the number of observers. In order for cooperation to spread beyond the starting group of (typically) 10 observers, a mechanism – an interaction structure – needs to be in place. However, BAPP doesn't include any structural mechanism to generate this expansion of cooperating workers. Thus, in principle, the expansion of observers would need to appeal to the cooperative spirit of workers. This may lead to some penetration of BAPP, but perhaps not enough. The structure of the benefits is such that as the number of observers expands, the additional observer will face lower "reputational benefits". This will cause the likelihood of cooperation to decrease with an increasing number of observers.

In game 2, the impact of size operates through repeated interactions. Given the absence of structural mechanisms in BAPP (which could, for example, a centralized punishment of low effort; see Boyd et al, 2010 or Kosfeld and Rustagi, 2015), a natural mechanisms that is applied in BAPP is peer pressure and reciprocity, particularly within the starting team of observers, the committee that meets regularly. As BAPP is voluntary, "formal" or "direct" pressure are absent (for example, the enabler does not "enforce" effort or "fire" an observer). However, social control can operate in two ways: informal peer to peer communication that puts pressure on observers of low effort, or it can be sustained by the implicit threat by cooperating observers to withdraw their high effort. Both mechanisms would become less effective as the number of observers expands. Informal peer pressure becomes harder to execute, particularly for new observers that aren't part of the monthly committee, and reciprocity in public goods games breaks down fairly quickly as the group grows (Boyd and Richerson, 1988).

In the game 3, as a consequence of game 2 and also as a decrease in the "reputational" benefits when there are many observers, we would expect that a higher number of observers would make observers less prone to cooperate. In addition, in the absence of an interaction structure, having more observers means that a worker is observed by multiple and different observers, hampering the ability of direct reciprocity to operate. This would decrease the workers' incentives to cooperate.

3.3. Evidence from large scale administrative data

In this section we use a representative sample of BAPP projects in order to answer the following questions:

i. How does the number of observers and the contact rate expand over?
ii. What is the impact of BAPP on site accidents and culture?
iii. Does the impact on accidents varies according to the number of observers?
iv. Are there any interaction structures that can be useful to improve the impact of BAPP on accidents?

3.3.1. Data

We worked closely with DEKRA in generating a dataset that would allow us to study the impact of BAPP on accidents. DEKRA had already collected in 2013 an administrative dataset of 1,352 sites with a BAPP implementation. This dataset contains projects executed between 1989 to 2013. Although these projects do not represent the entirety of the projects executed by DEKRA over this period, it covers a substantial percentage of their BAPP activity over the years[45]. We refer to this dataset as the population of BAPP projects. The unit of the dataset is a site-month. For each site and month, we have information on: date, name of site, company of site, industry of company, country of site's location, name of consultant, presence of culture, number of observers, number of observations, number of workers observed (in a minority of cases, an observation is done to two workers at the same time), number of coached observations, method of BAPP implementation, method of training (in a small share of cases, training of new observers is done by DEKRA and not the observers of the committee), number of critical behaviours that are tracked, the number of critical behaviours that were observed, the number of observed critical behaviours that were safely and riskily executed, number of workers on the site, and number of accidents. Regarding accidents, DEKRA took great care in harmonizing accident data across countries as they might be different rules in the reporting of these events.

In order to execute the analysis we restricted the sample to those projects that have information on workplace accidents at least 2 years before and 3 years after the start of BAPP. The start of BAPP is measured by the month when observations start. This generated a sample of 88 sites. In the **Table 11** we compare the sample and the population and executed statistical test for several variables in our dataset. Except for year of start of BAPP, all the other variables are not statistically different between the population and the sample. Regarding year of start, in the sample we have newer projects (see appendix 3.7.1).

[45] The reason for this is that the data of BAPP projects (observations, sheets, accidents, etc.) is stored in a software that is local and proprietary to each site. In 2013 DEKRA decided to collect all the historical information by asking the collaboration of its clients to share and retrieve the data to DEKRA. A substantial portion of their clients collaborated in this effort.

Table 11. Comparison of population and sample of sites

	Population Average (S.D.)	Sample Average (S.D.)	Statistically different?
Workers	279 (223)	245 (160)	No
Accidents	1.59 (2.33)	1.22 (1.39)	No
Industry	(Categorical)		No
Country	(Categorical)		No
States within US	(Categorical)		No
Year of start BAPP	(Categorical)		Yes
Who trains observers	(Categorical)		No
Type of Implementation	(Categorical)		No
Number of critical behaviors	27.6 (7.2)	27.3 (6.6)	No

3.3.2. Evolution of the number of observers and the contact rate

In this section, we explore how cooperation evolves within a site as a BAPP implementation gets implemented. In particular, we explore the expansion in the number of observers, the expansion of the reach to workers (i.e., the contact rate), and the individual effort of observers.

To study this, we define three terms using the following equation:

"Contact rate" = observations / workers = observations / observers x observers / workers

$$= \text{“intensity”} \times \text{“participation”} \quad (6)$$

"Contact rate" is the number of observations per worker in a site in a given month. Workers excludes supervisors or managers, it only considers "frontline" employees which are eligible for observations and for becoming observers. Contact rate reflects the expected number of observations that worker could expect to receive in a month. As indicated, BAPP aims to achieve a contact rate of 1 over time.

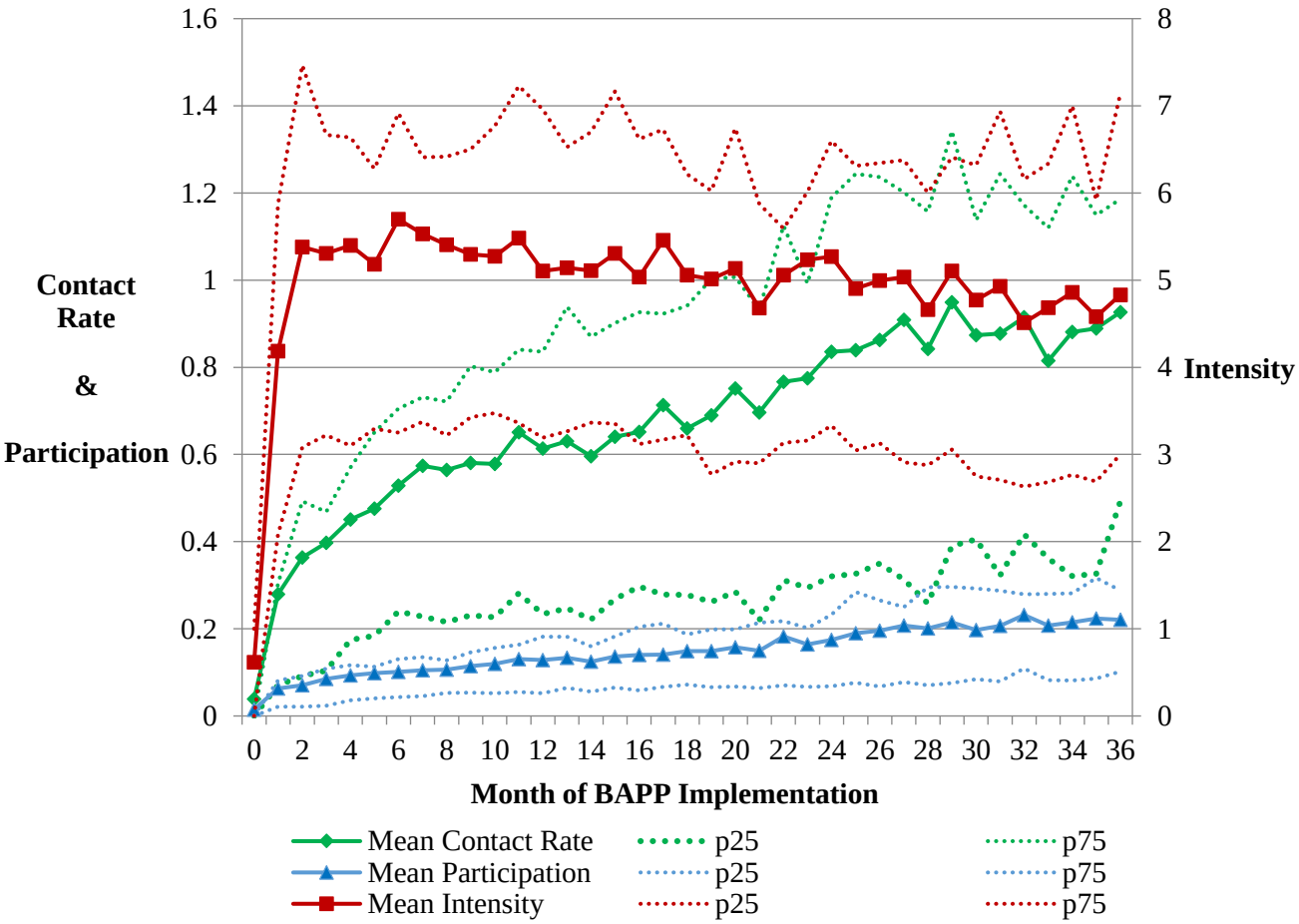
The contact rate can be broken down into two components: "Intensity" which captures the number of observations per observer per month, and "participation" which captures the share of workers that have become active observers. Active means that the observer has done 1 or more observations in a month. Intensity measure the magnitude of the cooperative effort by cooperators, and participation captures the penetration of cooperators in the site.

In the **Figure 7**, we display the average and percentiles 25 and 75 for these three variables over the 36 months of a BAPP implementation (considering the 88 sites of our sample). Contact rate (green line) approach the goal of 1 by the end of year 3, but there is considerable variation across sites (dotted green lines). This indicates that overtime, BAPP reaches approaches to its goal of 1 observations per worker. Intensity (red line) is very stable over time, displaying a very slight decrease from ~5.3 in the first year to ~4.8 in the third year. Variation is also high (red dotted lines): sites at the 25th percentile display around 3 observations per observer per month, while at the 75th percentile this increases to 6.5. This indicates that the average cooperative effort is stable over time. Participation has a steady and uniform increase from 6%-8% in first few months to 20% at the last months of the third year. Given the average number of workers of 245 in our sample, this translates into a change from ~10 observers to ~50 observers over the span of 36 months.

Overall, taken at face value, the evolution of these indicators suggest that overall cooperation does diffuse within the average BAPP implementation. Although participation is not high, and its expansion is slow, we see a steady increase over time. In addition, average intensity doesn't seem to significantly decrease as the number of observers expand. This is surprising given the lack of a defined "interaction structure" in BAPP.

However, these average values conceal a much richer underlying dynamics. As we will show in section 3.3.5, if we break down the analysis by cohort of observers, we find that the new additional observers, as compared to the observers of the committee, have: i) much lower intensity, and therefore, cooperative effort does suffers as observers expand; ii) have and much higher rotation, and therefore, and the adoption of cooperation becomes fragile as observers expand: for the same increase of active observer it is needed to train an increasing number of workers as BAPP expands. Below we will dive deeper into these dynamics by studying its impact on performance. For that, we have to turn first to the analysis of the impact of BAPP on performance.

Figure 7. Evolution of contact rate, intensity and participation over a BAPP implementation



3.3.3. Impact of BAPP

The previous section indicates that on average BAPP gets cooperation going (abstracting for now the interesting dynamics that are generated across different cohorts of observers). Therefore, studying is impact To study the impact of BAPP on accidents, we use the following model:

$$ACCIDENTS_{it} = b_1 + b_2 \times BAPP_{it} + b_3 \times TREND_{it} + b_4 \times (BAPP_{it} \times TREND_{it}) + b_5 \times \ln(WORKERS_{it}) + U_i + ERROR_{it} \quad (1)$$

In equation (2) we model the accidents of the site i in the month t. BAPP is a variable that takes the value of 1 in the month where the first observation is executed in the site. TREND equals to $(t - \theta_i)$ where t is the month and θ_i is the month when the BAPP started in the site. Given our sampling, this variable goes from -24 to +36. We added a site fixed effect U_i to the estimation in order to control for

time-invariant store unobservables. As a control, we added the natural logarithm of workers as more workers translate into more accidents[46].

The test we perform with this model is a within site before and after comparison, where we control for a common trend for all sites. Given that we include BAPP x TREND in the model, the coefficient b3 captures the average accidents trend without BAPP.

In **Table 12** we display the results. All models include site fixed effects. In column (1) we explore the simplest model, only with BAPP and Ln(workers) as control. In (2) we add the trend. In (3) we add the interaction between BAPP and TREND. In column (4) we display the POISSON fixed effect estimates as a robustness.

From columns (1) we that the impact of BAPP has a high statistical significance. Model 2 shows that the TREND is negative and statistically significant, but that BAPP looses its statistic significance. This is expected as there is collinearity between BAPP and TREND (BAPP equals o when TREND is lower than zero and 1 when TREND is higher than zero). This model not only introduces collinearity, it also generates the doubt of whether really the impact of BAPP in (1) is simply capturing a trend. But model (3) dispels this concern: it is easy to appreciate that the trend turns negative only after BAPP. The trend without BAPP (b3) is flat and non-significant.

In columns (3) and (4), we find a negative and statistically significant impact of BAPP on the accidents of the site. The p-value of the joint t-test for BAPP, TREND and TREND*BAPP is below 0.001 (a joint t-test for BAPP and BAPP*TREND is significant at 5%); this test is required because, as discussed, BAPP and TREND are highly collinear (in model 3 the VIF is above 6 for these variables).

In all the models, the impact of number of workers is well behaved, with a strong positive impact on accidents.

Using column (3), we find that BAPP is related to a decrease in the level of accidents of 0.2 accidents (b2) and then for each month to a decrease in the accidents by 0.011 per month of implementation

[46] The safety literature has documented a downward secular trend in accidents over time. This might require adding year fixed effects to control for this trend. However, there are two reasons that argue against this. First, in the appendix we show that the majority of the projects start between 2003 and 2007 but that we still have long tails before and after. In these tails, only a few projects –in many cases one or two– start in each year. Given this small number of projects in the tails, the year fixed effect might capture the effect of BAPP on not a secular trend. For the fixed effect to capture properly the secular trend, a high density of projects is required. Second, the site fixed effects can capture part of the potential impact of a secular trend. The 3 years over which the project was executed is part of the unobserved characteristics captured by U_i. That said, we ran several models adding year fixed effects, month fixed effects, year*industry fixed effects, and year*country fixed effects and the results did not change; instead, they became stronger.

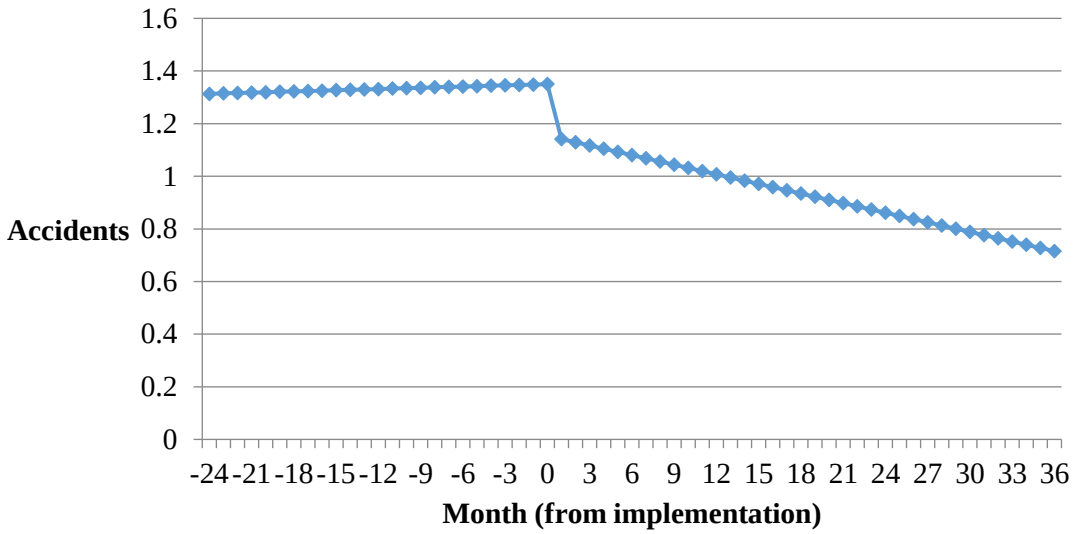
(b4), which after twelve months amount to 0.132 accidents. Given the mean of our dependent variable, these estimates are large, accounting for 15% each (or a total decrease of 30% at the end of the first year of BAPP). We graph this result in the **Figure 8**. We performed power calculations on these estimates. Assuming significance of 5% and a standard deviation conditional on year, number of workers and site dummies (equal to 1.66 accidents) we find that the power of the effect of BAPP after 1 year is 48%, after 2 years is 73%, and after three years is 92%. This suggests that, although the statistically significant effects may be false positives, it becomes more likely that we are uncovering a true effect as the implementation progresses (particularly at the end of the third year).

Table 12. Impact of BAPP on accidents

	Accidents – OLS (1)	Accidents – OLS (2)	Accidents - OLS (3)	Accidents - POISSON (4)
BAPP	-0.357*** (0.087)	-0.162† (0.104)	-0.198*† (0.115)	-0.156*† (0.085)
TREND		-0.007*† (0.004)	0.001† (0.007)	-0.001† (0.005)
BAPP x TREND			-0.011† (0.009)	-0.011† (0.007)
Ln(WORKERS)	1.030*** (0.300)	1.028*** (0.306)	1.028*** (0.302)	0.714*** (0.088)
Site fixed-effect?	Yes	Yes	Yes	Yes
Constant	-4.171** (1.61)	-4.241** (1.61)	-4.149** (1.60)	
R-square (Log Likelihood)	42.20%	42.28%	42.32%	(-5,390.16)
Observations	4,762		4,762	4,762
Mean of dependent variable before BAPP	1.338		1.338	1.338

Errors in parentheses are robust and clustered at the site level. * p<0.1, ** p<0.05, *** p<0.01 in two-tailed test. † indicates p<0.01 in a two-tailed joint t-test (this test is required as there is multicollinearity between BAPP, TREND and their interaction). A joint t-test on BAPP and BAPP x TREND in model (3) is also statistical significant at p<0.05.

Figure 8. Impact of BAPP over time



However, given our sample of project that only received a BAPP implementation, it is not possible to assess whether BAPP would be effective if it were to be randomly assigned to a population of sites. The absence of information on projects that don't have BAPP precludes us to say anything in this regard. Therefore, all of our estimates, including the power calculations, should be placed in reference to the population of projects that received or are prone to receive BAPP[47]. What we can do, notwithstanding, is to probe whether within our (biased) sample the effects are causal. Given that we already included site fixed effects, the main threat to identification is time-variant unobservables at the site level. For example, it could be that a year before the start of BAPP, the site changed its manager which happened to be safety-conscious, and thus it is this manager that is related to both to BAPP and a reduction in accidents.

To tackle this issue we executed two analyses. First, we do a flexible placebo test using the following model:

$$ACCIDENTS_{it} = b_1 + \sum_j (\pi_j \times YEAR\_BAPP\_P_j \times BAPP\_P_{it}) + b_3 \times TREND_{it} + \sum_j (\rho_j \times YEAR\_BAPP\_P_j \times BAPP\_P_{it} \times TREND_{it}) + b_5 \times \ln(WORKERS_{it}) + U_i + ERROR_{it} \quad (2)$$

In this model, BAPP_P is the "placebo BAPP" and takes the value of 1 after the 12th month preceding the real start of BAPP (i.e., BAPP starts in month -11). YEAR_BAPP is a dummy set that identifies

[47] In the parlance of the experimental paradigm, we estimate an intent-to-treat effect and not a treatment on the treated. It's as if: we had 200 BAPP programs to deliver; we randomly generated control and treatment groups of, say, 200 sites each; the treatment consisted on telling the treated sites that they could implement BAPP; in the end, only 88 sites self-selected to get the treatment; and data was collected only for the 88 sites. Given that we didn't collect baseline data on the control and those that self-selected, we cannot assess the representativeness of our findings beyond those with an "intent" to be treated.

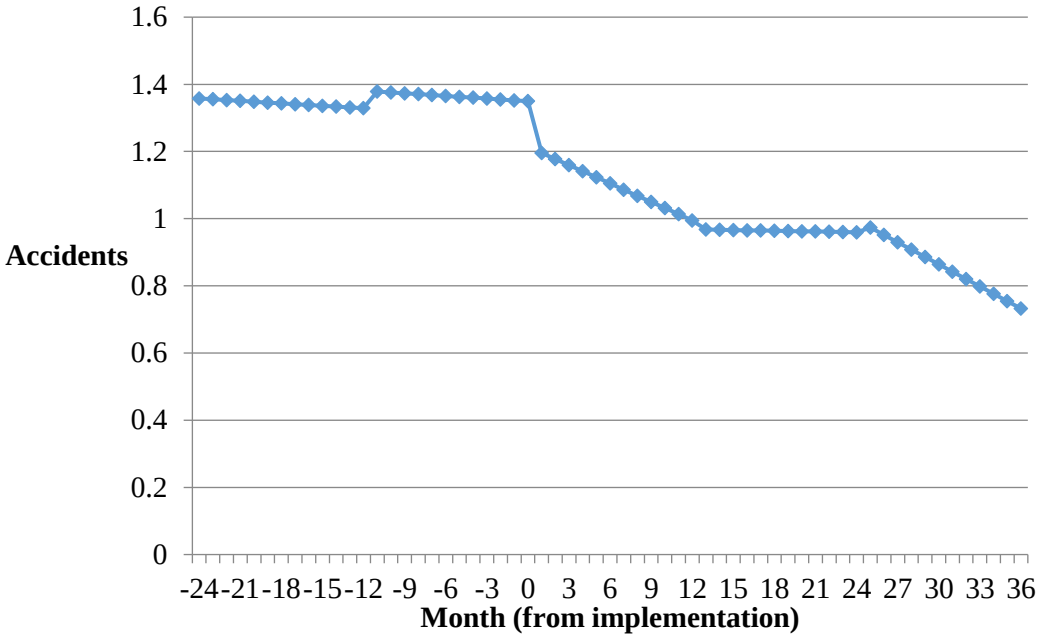
the year preceding the real start of BAPP (from -11 to 0, where 0 is the month preceding the start of observations), the first year of observations (from 1 to 12), the second year of observations (from 13 to 24) and the third year of observations (from 25 to 36). (Thus, J=4.) Essentially, this models breaks down the impact of BAPP on the level and slope into four parts, including one year before the actual start, the placebo year. If the sites were already experiencing a change in their safety due to an unobserved time-variant element, then we would expect to find movement in the placebo year. The coefficient b3 now identifies the trend in the months going from -24 to -12.

Table 13 presents the estimates of equation 2. Interpreting this table can be tricky, so we graph the result in **Figure 9**.This figure clearly shows that there is no effect in the year before BAPP, neither at the level or slope. The effect of BAPP is concentrated in the years 1 and 3 (and the year number 3 can be partly attributed to culture change; see analysis below).

Table 13. Placebo test on the impact of BAPP

	Accidents - OLS
BAPP_P x PLACEBO YEAR	0.049 (0.246)
BAPP_P x FIRST YEAR	-0.085 (0.246)
BAPP_P x SECOND YEAR	-0.323 (0.404)
BAPP_P x THIRD YEAR	0.220 (0.524)
TREND	-0.002 (0.014)
TREND x BAPP_P x PLACEBO YEAR	-0.000 (0.018)
TREND x BAPP_P x FIRST YEAR	-0.016 (0.023)
TREND x BAPP_P x SECOND YEAR	0.002 (0.020)
TREND x BAPP_P x THIRD YEAR	-0.019 (0.019)
Ln(WORKERS)	1.028*** (0.303)
Site fixed-effect?	Yes
Constant	-4.211** (1.610)
R-square (Log Likelihood)	42.34%
Observations	4,762
Mean of dependent variable before BAPP	1.338
Errors in parentheses are robust and clustered at the site level. * p<0.1, ** p<0.05, *** p<0.01 in two-tailed test. † indicates p<0.001 in a two-tailed joint t-test (this test is required as there is multicollinearity between BAPP, TREND and their interaction). The joint t-test on BAPP and BAPP x TREND is also statistical significant at p<0.05.	

Figure 9. Impact of BAPP in placebo year



The second analysis that we execute in order to check for time variant unobservables is a random trend model. This model fits an individual slope for each site:

$$ACCIDENTS_{it} = b_1 + b_2 \times BAPP_{it} + b_i \times TREND_{it} + b_4 \times (BAPP_{it} \times TREND_{it}) + b_5 \times \ln(WORKERS_{it}) + U_i + ERROR_{it} \quad (3)$$

In this model, the coefficient of TREND is indexed, meaning that each site has their own trend. To estimate this model we use first differences (to which we add a constant a_1) and a fixed effect technique on the new data:

$$\Delta ACCIDENTS_{it} = a_1 + b_2 \times \Delta BAPP_{it} + b_i + b_4 \times \Delta(BAPP_{it} \times TREND_{it}) + b_5 \times \Delta\ln(WORKERS_{it}) + \Delta\, ERROR_{it} \quad (4)$$

Controlling for a specific trend for each site allows to control for unobservables that vary over time within the site and that might generate a systematic change in the accidents. The results are displayed in the **Table 14**. In column 1, we find that BAPP decreases their coefficients, both at the level (from -0.198 to -0.056) and the slope (from -0.011 to -0.008). The result is graphed in the **Figure 10**. Statistical significance suffer in these models, as models in difference are noisier (see the r-square).

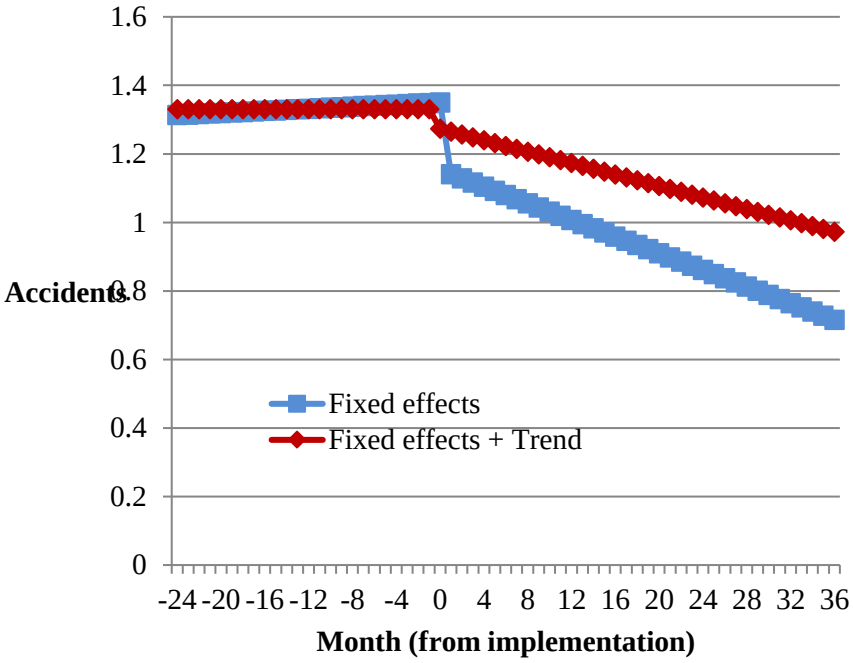
However, controlling for site-specific trend could also capture the quality of the BAPP implementation. The coefficients b_2 and b_4 are capturing the average impact of BAPP, thus b_i can be capturing the variation in the quality of the BAPP implementation. This implementation quality is a

time variant unobservable at the site level. Therefore, the estimates of 4 could be biased depending on the rarity of the different extremes of implementation quality. In the columns (2), (3) and (4) we attempt to accommodate for that possibility by eliminating the top and bottom 5%, 10% and 20% of the slopes b_i (eliminating the top and bottom 1% yields similar results to column 1). Here we find that the impact of BAPP increases and recovers its statistical significance. This is suggestive that the extreme values of time-variant unboservables are tilted toward the cases that are not favourable to safety; for example, more extreme cases of low implementation quality than high. This resonates with intuition and with the values displayed above in **Figure 7**.

Table 14. Impact of BAPP adding a site-specific trend as control

	ΔAccidents (1)	ΔAccidents (2)	ΔAccidents (3)	ΔAccidents (4)
Sample:	Full	Excluding top and bottom 5% of b_i	Excluding top and bottom 10% of b_i	Excluding top and bottom 20% of b_i
ΔBAPP	-0.056 (0.189)	0.066 (0.180)	0.197 (0.174)	0.065 (0.189)
Δ(BAPP x TREND)	-0.008 (0.013)	-0.017 (0.014)	-0.022* (0.013)	-0.025** (0.009)
ΔLn(WORKERS)	1.317** (0.609)	1.274* (0.719)	1.268 (0.799)	1.755* (0.971)
Site fixed-effect? (b_i)	Yes	Yes	Yes	Yes
Constant	-0.000 (0.008)	0.003 (0.008)	0.004 (0.007)	0.008 (0.006)
R-square	1.54%	1.44%	1.45%	5.9%
Observations	4,748	4,199	3,776	2,773
Errors in parentheses are robust and clustered at the site level. * p<0.1, ** p<0.05, *** p<0.01 in two-tailed test. All models are estimates using OLS panel fixed effect.				

Figure 10. Impact of BAPP adding a site-specific trend



The placebo and random-trend test, although not definitive, provide evidence that the impact of BAPP we document is likely to be causal (within the sample of sites prone to implement the methodology).

Another way to assess the credibility of the estimates is to assess mechanisms. Given how it is structured, we should expect BAPP to operate in some predictable ways. First, BAPP is a methodology that taps into the voluntary cooperation of workers. As the circle of cooperation enlarges we should expect BAPP to experiment difficulties as the number of observers expand. In the section 3.3.4 we exploit this in great detail and find strong evidence of this breakdown in cooperation.

The second mechanism we explore is how the pre-implementation culture of the site affects the impact of BAPP. As indicated above, DEKRA provides a service of surveying the culture of the site. This survey is always executed before any implementation of BAPP, typically around three months before the start of the implementation. This survey is used as an input at the planning stage. In the **Table 15** we describe the 10 cultural dimensions of this survey. DEKRA developed and refined this survey in the early 2000s in collaboration with organizational behaviour scholars. The survey is executed typically to all workers of a site using a web-based interface. Each one of the 10 factors/dimensions of the survey comes for a set of 5 points likert-scale items, whose psychometric properties (e.g.,

reliability) were assessed by previous scholars collaborating with DEKRA[48]. In non-reported regressions (available upon request), we executed a detailed analysis on how the impact of BAPP varies according to culture. We find that the overall score in the survey doesn't affect the impact of BAPP. Regarding the dimensions, we find that BAPP has a lower impact when the score for "Group relations" and "Approaching others" was high. Given that these dimensions are correlated themselves with a decrease in accidents, we interpret these findings as evidence of a substitution effect. BAPP operates by improving group relations (it teaches workers how to relate to one another in a good way) and teaching workers how to approach co-workers and provide feedback on their risky behaviour. If the pre-existing culture already displays these elements, then the impact of BAPP diminishes: the site are already doing what BAPP is supposed to do. The only additional dimension that is significant is "teamwork" which boosts the impact of BAPP. This dimension is not related to accidents, so it can't operate as a substitute. Instead, we believe this finding shows that BAPP is implemented in a better and more effective way when teamwork is high, which is intuitive. Finally, given that BAPP is a technology "by a for the workers" is not unexpected not to find heterogeneity of impact with respect to the factors grouped as "organizational factors".

[48] The underlying items and their scores was not provided by DEKRA due to copyright. We only obtained the value for each aggregate value for factor/dimension of the survey. More information about the survey can be found at https://dekra-insight.com/en/topic/organizational-culture-diagnostic-instrument-ocdi. In private conversations with DEKRA, I was reassured that the psychometric properties of the survey were closely scrutinized to secure its validity.

Table 15. Dimensions of culture survey

Area	Dimension	Definition by Dekra
Organizational factors	Procedural justice	The extent to which individual workers perceive fairness in the supervisor's decision-making process.
	Leader-member exchange	The relationship the employee has with his or her supervisor. In particular, this scale measures the employee's level of confidence that his supervisor will go to bat for him and look out for his interests.
	Perceived organizational support	The employee's perception of the employee that the organization cares about him, values him, and supports him.
	Management credibility	The employee's perception of the employee that what management says is consistent with what management does.
Team factors	Teamwork	The extent to which employees perceive that working with team members is an effective way to get things done.
	Group relations	The employee's perception they employee has of his relationship with co-workers. How well do they get along? To what degree do they treat each other with respect, listen to each other's ideas, help one another out, and follow through on commitments made?
Safety factors	Organizational value for safety (or Safety climate)	The safety climate scale measures the extent to which employees perceive the organization has a value for safety performance.
	Upward communication	The extent to which communication about safety flows upwards in the organization.
	Approaching others	The extent to which employees feel free to speak to one another about safety concerns.
	Injury reporting	The degree to which it is easy and secure to report safety incidents within the site

Impact on culture

Now we turn to the impact of BAPP on culture. In this subsection, we don't expand theoretically on the concept of culture. Instead, we take as face value the conceptualization of culture by DEKRA along with the instrument that they developed and used. Instead, the purpose of this subsection is to simply explore the impact of BAPP on culture thus understood. If BAPP operates as expanding cooperation in safety, then we should expect to find impact on culture. Particularly, we should find an impact on safety impact.

DEKRA collected data on before and after culture surveys for 78 sites. The first survey occurred between 2001 and 2006 as part of the survey development and refinement effort executed by DEKRA

with scholars. This effort included 94 sites and ended up in a proprietary report. The follow up occurred between 2005 and 2009, with some attrition occurring.

The model we use to assess the impact of BAPP on site culture is the following:

$$CULTURE_{iw} = b_1 + b_2 \times SECOND\_SURVEY_w + b_3 \times SECOND\_SURVEY_w \times YEARS_i + b_4 \times Ln(RESPONDENTS)_{iw} + MONTH_{iw} + U_i + ERROR_{iw} \quad (5)$$

Culture is the average of the items of the survey, whether as a specific area or the survey as a whole, for the site i on the wave of survey w, which can take the value of w=1 or w=2. The dummy SECOND_SURVEY takes the value of 1 if the survey is the follow-up survey (w=2). Given that the follow-up survey is not executed at the same time, the variable YEARS capture the number of years that separate the follow-up from the baseline survey (which goes from 1 to 6, quite uniformly distributed). The variable RESPONDENTS measures the number of workers that responded the survey (which is close to the number of workers). We add a site fixed effect (U_i) to control for time invariant unobservables and a month of survey fixed effect ($MONTH_{iw}$) to control for any seasonality that might affect responses[49].

In the **Table 16** we present the results for the survey as a whole, that is, the average for the 10 dimensions. Column 1 shows that average impact of BAPP on overall culture is 0.123, statistically significant at 99% and representing a bit more than half a standard deviation. In column 2 we find that this effect is increasing over time. At the course of 3 years, the impact on culture is 0.121, and over 6 years is 0.22, or 1 standard deviation.

Of course, these estimates are subject to endogeneity from omitted variables, specifically from time variant unobservables at the site level. For example, just like the case of accident, it could be that BAPP coincided with a change in top management that generates a change in culture while implementing BAPP. We provide four elements against this possibility. First, the gradual increase of culture coincides with the gradual impact that BAPP has on the sites, as shown in **Figure 8** and in **Figure 7**. A time variant unobservable would need to generate this pattern, which excludes, following our example, of a change in the management team, which is a one-off change. Second, in model 3 we add an interaction between the second survey with the number of respondents (or number of workers,

[49] We don't add a year of survey dummy set because the overlap between baseline and follow up very small (less than 15 sites). This generates a big bias as the year fixed effects would be capturing a large part of the SECOND_SURVEY variable.

as survey compliance is high). This variable can capture some of the time variant unobservables of the site. However, we find that the results are robust to the inclusion of this interaction.

As a third argument, we study the different areas of the survey. In the **Table 17** the results indicate that the increase in culture is exerted primarily on the "Safety Factors", the area where BAPP is poised to exert the greatest direct impact. The impact on safety culture is significant: after three years, an increase of 40% of a standard deviation; after six years, an increase of 80%. The impact of BAPP on the "organizational factors" and "team factors" areas is indirect. In columns (2) and (4) we see that safety factors is positively associated to the remaining two factors. Therefore, one can compute the indirect effect of BAPP on them. On its third year, BAPP exert an indirect impact on "organizational factors" that is equivalent to 24% of its standard deviation (using column (6) and (2) we compute [0.006 + 0.028*3] * 0.752 / 0.278). On the 6th year, the increase is 47%. Analogously, the indirect effect of BAPP on "team factors" is 21% and 40% of its standard deviation on the third and sixth year respectively.

The fourth and final argument is the result for Oster delta. For model (5) of **Table 17** we obtained a delta of 3.6 (assuming a maximum r-square of 1). This means that in order for unboservables to overthrow our results, the selection on unobservables would need to be 3.6 times the selection on observables. Overall, evidence suggest quite strongly that BAPP exerts a causal impact on culture. Of course, as discussed above, given our sample, this impact is circumscribed to the sites that received (or are prone to receive) the BAPP methodology[50].

Table 16. Impact of BAPP on site culture

	CULTURE (1)	CULTURE (2)	CULTURE (3)
SECOND_SURVEY	0.123*** (0.019)	0.022† (0.055)	0.105‡ (0.117)
SECOND_SURVEY x YEARS		0.033*† (0.017)	0.030*‡ (0.016)
Ln(RESPONDENTS)	-0.069 (0.051)	-0.068 (0.049)	-0.056 (0.054)
SECOND_SURVEY x Ln(RESPONDENTS)			-0.015 (0.020)
Site fixed-effect?	Yes	Yes	Yes
Constant	3.856 (0.008)	3.855*** (0.259)	3.808*** (0.274)

[50] We also did an additional model replicating column (6) of **Table 17**. We added an interaction between the second survey and the "organizational factors" and the "teamwork factors". These two interaction would capture time variant unobservables that changed these areas in the survey. The results remained the same as compared to column (6): the coefficient SECOND_SURVEY x YEARS was 0.027 significant at 10% and the joint t-test significant at 1%.

Adjusted R-square	81.43%	82.18%	%
Observations	156	156	156
Mean (st. dev.) of culture on first survey	3.526 (0.228)	3.526 (0.228)	3.526 (0.228)
Errors in parentheses are robust and clustered at the site level. * p<0.1, ** p<0.05, *** p<0.01 in two-tailed test. All models are estimated using an OLS panel fixed effect. † indicates p<0.01 in a two-tailed joint t-test. ‡ indicates p<0.05 in a two-tailed joint t-test.			

Overall, the impact that we document for BAPP indicates that cooperation is beneficial, as is it typically assumed in the literature. Here we provide much needed field evidence of the power of cooperation. BAPP allows to explore the impact of cooperation as it evolves within a site quite cleanly, without the confounds that other types of approaches face (for example, the organizational citizenship literature).

The large impact on accidents and culture is impressive. This begs the question of why companies haven't executed this type of cooperation-based-methodologies on their own, as a grass-root innovation that latter diffuses within and across companies? We believe there are four answers to this:

i. "They might not know". Just like evolution requires mutations, organizations require to come up with the idea of such a policy/practice (internally or from copying), and this might never occur.
ii. "They know, but the incentives are not there". As discussed above and documented in 3.3.8, becoming an observer and executing observations is a social dilemma, and therefore, any grass-root attempt a BAPP type of policy will have to fight against the resistance and friction generated by self-interested free-riding. Instead, BAPP provides a nice catalyzer to get the cooperative spark going. Whether these cooperation incentives get diluted as BAPP evolves within a site, is analyzed in the next section.
iii. "They know, they have the incentives, but peer-to-peer coordination is hard". Even if social preferences are prevalent in a site, workers would need to coordinate in order to execute such a policy. This is not easy to do, and if done poorly, it may generate unfair situations affecting the willingness to cooperate. This is related to the clarity problem in building relational contracts (Gibbons and Henderson, 2013).
iv. "They know, they have the incentives, but hierarchical coordination crowds-out the incentives". An alternative to peer-to-peer coordination is to have the company or managers to direct in a top-down fashion the collective effort. However, this can plausibly crowd out social preferences by the difficult in assessing and believing the true underlying motivations

of companies. This related to the credibility problem in building relational contracts (Gibbons and Henderson, 2013).

Table 17. Impact of BAPP on different dimensions of culture

	ORGANIZATIONAL FACTORS (1)	ORGANIZATIONAL FACTORS (2)	TEAMWORK FACTORS (3)	TEAMWORK FACTORS (4)	SAFETY FACTORS (5)	SAFETY FACTORS (6)
SECOND_SURVEY	0.037 (0.029)	-0.008 (0.065)	0.038 (0.246)	0.028 (0.054)	0.088*** (0.022)	0.006† (0.044)
SECOND_SURVEY x YEARS		0.016 (0.020)		0.004 (0.016)		0.028*† (0.016)
Ln(RESPONDENTS)	0.043 (0.069)	0.040 (0.070)	-0.096* (0.048)	-0.097** (0.048)	-0.067 (0.039)	-0.063* (0.037)
ORGANIZATIONAL FACTORS			-0.254 (0.365)	-0.256 (0.367)	0.324*** (0.081)	0.292*** (0.084)
TEAMWORK FACTORS	-0.226 (0.280)	-0.226 (0.278)			0.219*** (0.073)	0.203*** (0.073)
SAFETY FACTORS	0.793*** (0.196)	0.752*** (0.181)	0.599* (0.353)	0.591* (0.354)		
Site fixed-effect?	Yes	Yes	Yes	Yes	Yes	Yes
Constant	1.191 (1.041)	1.349 (1.039)	2.773*** (0.627)	2.808*** (0.615)	1.939*** (0.515)	2.115*** (0.519)
Adjusted R-square	77.12%	76.79%	68.85%	65.07%	85.32%	85.98%
Observations	156	156	156	156	156	156
Mean (st. dev.) on first survey	3.258 (0.278)	3.258 (0.278)	3.680 (0.259)	3.680 (0.259)	3.740 (0.221)	3.740 (0.221)

The dependent variables are the average of the factors within each area of the survey. Errors in parentheses are robust and clustered at the site level. * p<0.1, ** p<0.05, *** p<0.01 in two-tailed test. All models are estimated using an OLS panel fixed effect. † indicates p<0.01 in a joint two-tailed t-test.

3.3.4. Large groups decrease the impact of BAPP

Thus far, we have documented the beneficial impact of BAPP on accidents and culture. In this section, we dig deeper and unveil the dynamics underlying this average effect. In particular, we study how different elements components of cooperation affect the impact of BAPP. To do so, we go back to the three variables defined in section 3.3.2: the contact rate and its two components, intensity and participation. These variables capture the essence of cooperation in BAPP as it evolves: intensity captures cooperative effort, and participation captures the spread of the cooperative trait, and contact rate captures the amount of cooperative benefits with workers.

To explore the how the contact rate affect the impact of BAPP, we use the following regression model:

$$ACC_{it} = b_1 + b_2 \times BAPP_{it} + b_3 \times TREND_{it} + \sum_j b_{4j} \times BAPP_{it} \times QUINT\_CR_j + b_5 \times \ln(WORKERS_{it}) + U_i + ERROR_{it} \quad (7)$$

In this model, QUINT_CR capture the quintiles of the contact rate, thus J = 5. In the **Table 18** we present the results. The joint t-test indicates that BAPP as whole is significant[51]. In column (2) we add as a control the interaction between BAPP and TREND. However, the coefficients of the quintile of contact rate are stable across columns (1) and (2). (Contact rate grows as the time passes in the implementation (**Figure 7**). At the same time, BAPP changes the slope reducing accidents over time (**Figure 8**). And this could be driven by effects other than the increase in contact rate, such as the change in culture.)

Each coefficient b_{4j} captures that effect of BAPP at each quintile of contact rate. In the **Figure 11** we graph the result. Contract rate has a non-linear relation on the impact of BAPP on accidents. It increases in the first two quintiles, then drops slightly for the third quintile, and finally it drops quite sharply for the last two quintiles. These results begin to unveil the dynamics at play: the

[51] Similarly to table 12, the joint t-test is required because contact rate has a high collinearity with trend. Notice that in column (1), if the baseline coefficient BAPP is dropped and its interaction with the 5th quintile kept, then the interaction with the 2nd and 3rd quintile would display p-values of 0.014 and 0.034; the same for column (2).

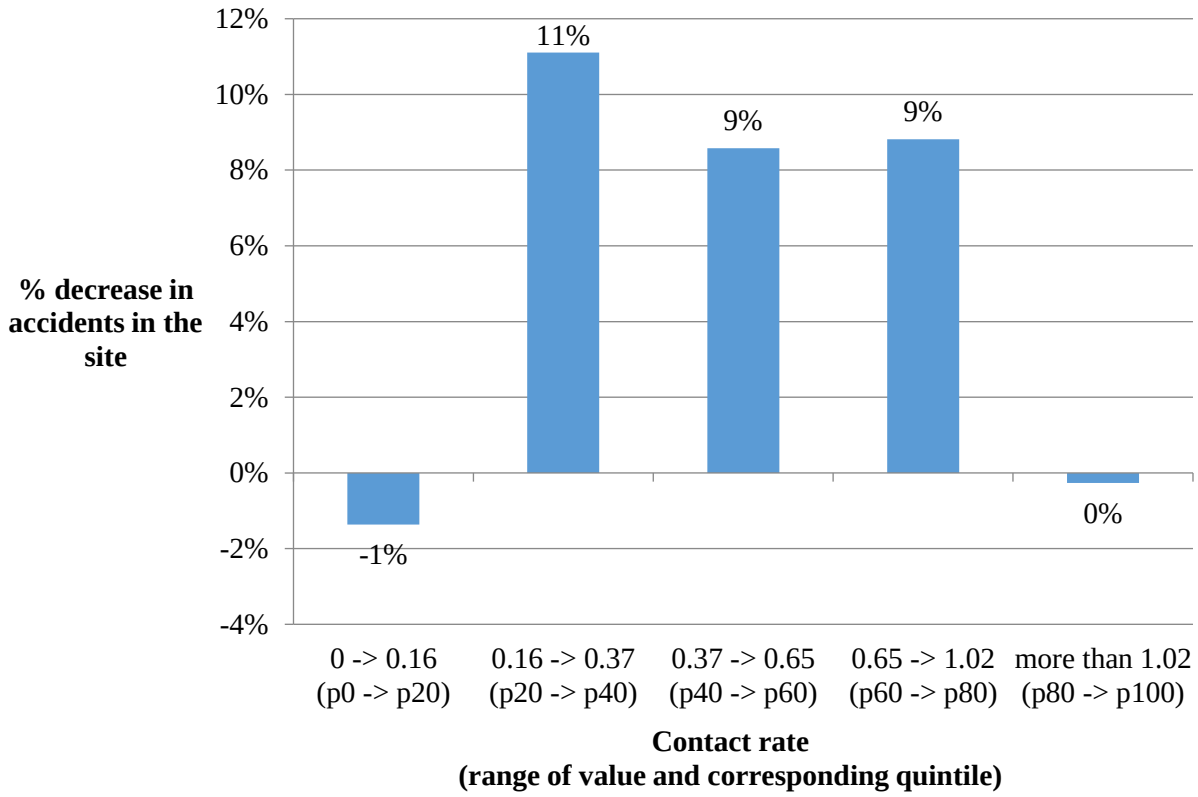
benefits of cooperation seem to be decreasing. However, to fully understand what is going on, we study the components of contact rate.

Table 18. Role of contact rate on the impact of BAPP

	Accidents (1)	Accidents (2)
BAPP	-0.124† (0.191)	-0.123‡ (0.192)
BAPP X 1ST QUINTILE OF CONTACT RATE	0.018† (0.145)	-0.029‡ (0.142)
BAPP X 2ND QUINTILE OF CONTACT RATE	-0.155† (0.175)	-0.188‡ (0.175)
BAPP X 3RD QUINTILE OF CONTACT RATE	-0.125† (0.125)	-0.148‡ (0.121)
BAPP X 4TH QUINTILE OF CONTACT RATE	-0.004† (0.109)	-0.020‡ (0.106)
BAPP X 5TH QUINTILE OF CONTACT RATE	(Omitted)	(Omitted)
TREND	-0.006† (0.004)	0.001‡ (0.007)
BAPP X TREND		-0.011‡ (0.009)
Ln(WORKERS)	1.082*** (0.318)	1.085*** (0.319)
Site fixed-effect?	Yes	Yes
Constant	-4.528*** (1.690)	-4.448*** (1.691)
Adjusted R-square	41.00%	41.00%
Observations	4,625	4,625
Mean of dependent variable before BAPP	1.338	1.338

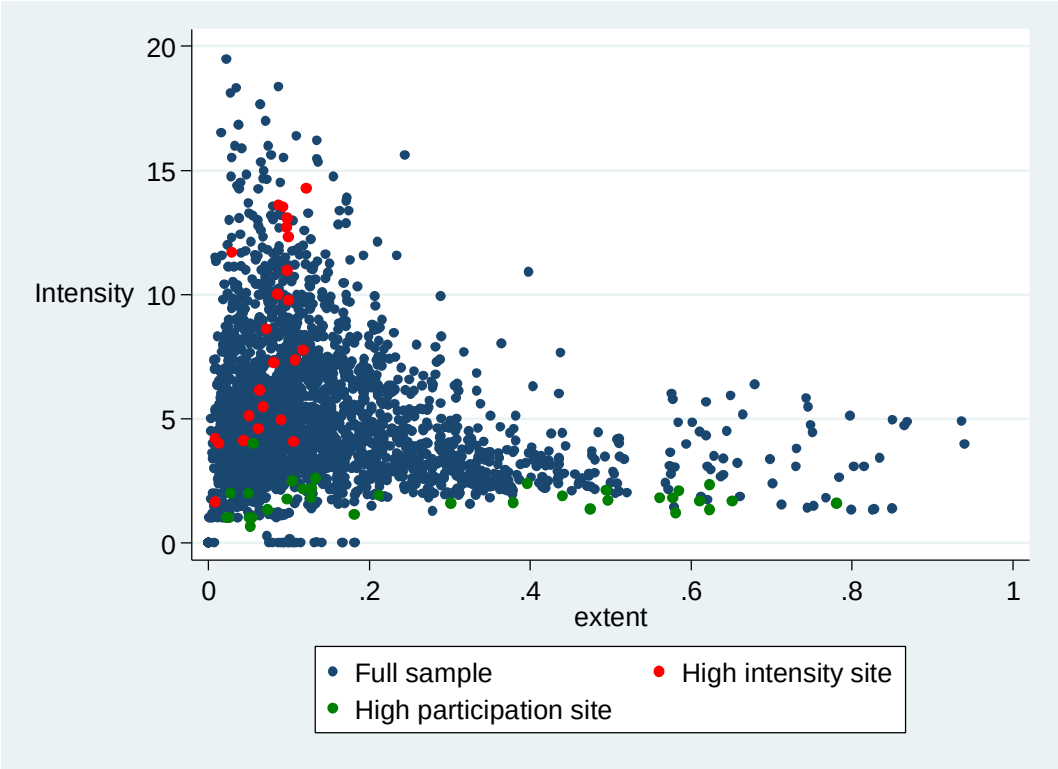
Errors in parentheses are robust and clustered at the site level. *** p<0.01 in two-tailed test. All models are estimated using an OLS panel fixed effect. † indicates p<0.001 in a two-tailed joint t-test. If TREND is dropped from the Joint test in column (1), the p-value is 0.063; if dropped from the Joint test in column (2), the p-value is 0.087. In column (1), if the baseline coefficient BAPP is dropped and its interaction with the 5th quintile kept, then the interaction with the 2nd and 3rd quintile would display p-values of 0.014 and 0.034; the same for column (2).

Figure 11. The impact of BAPP varies according to contact rate.



A high contact rate can be achieved using two generic strategies: high intensity and low participation, or low intensity and high participation. In **Figure 12** we display all the month-site combinations of participation and intensity for the three years of BAPP implementation. It can be readily seen that there is a trade-off between intensity and participation. For example, in red we display a site that achieved a contact rate of 1 by growing on intensity while keeping participation around 0.1. In green we display a site that achieved a contact rate of 1 by growing on participation while keeping its intensity around 2. We exploit this variation in strategies to estimate the isolate the impact of intensity and participation.

Figure 12. Two strategies to increase contact rate



In order to explore the separate impact that participation and intensity exert on accidents, we use the following model:

$$ACC_{it} = b_1 + b_2 \times BAPP_{it} + b_3 \times TREND_{it} + \sum_j b_{4j} \times BAPP_{it} \times QUINT\_INT_j + \sum_j b_{5j} \times BAPP_{it} \times QUINT\_PART_j + b_6 \times \ln(WORKERS_{it}) + U_i + ERROR_{it} \quad (8)$$

This model is the same as equation (7) with the difference that now we break the five quintiles of the contact rate into two sets of five quintiles of intensity and participation. The coefficients b_{4j} and b_{5j} capture the marginal impact of each quintile of intensity and participation respectively.

In **Table 19** we present the results. Considering that intensity and participation grow systematically during the implementation, in column (2) we add the control of BAPP times TREND. However, the results don't change compared to column (1) (this suggest that the individual term of TREND is a sufficient control). The results indicate that intensity decreases accidents, and that this impact

is increasing. On the contrary, participation decreases accidents at first, but then it increases them. Participation is not significant due to high collinearity with TREND. Participation increases with time. If the variables of participation and intensity are dichotomized into dummies –reducing the collinearity problem slightly, both dummies are statistical significant (see the table 23 below).

In **Figure 13** and **Figure 14** we display the impact of intensity and participation, respectively. Each figure mirrors the two sites highlighted in **Figure 12**: while keeping one dimension constant at its second quintile, we display the impact of changing quintiles in the remaining dimension. This figure display a clear inverted-U relationship between participation and accidents. This means that, conditional on intensity, participation is only beneficial up to approximately a participation of 0.08. Given the average site size of 245 employees, this means that after having approximately 20 observers, adding more observers is detrimental. This results provide supporting evidence for the prediction that cooperation would suffer as it expands: more observers mean that the free-riding temptation in providing cooperative effort increases.

This results provide supporting evidence for the prediction that cooperation would suffer as it expanded. However, one would expect that there could be decreasing returns from adding more observers (as in the contact rate), but not a negative impact. In order to answer this question, in the next section we dig deeper using observation and observer level data.

Table 19. The role of intensity and participation on the impact of BAPP

	Accidents (1)	Accidents (2)
BAPP	0.016 (0.149)	-0.039 (0.152)
BAPP X 1ST QUINTILE OF INTENSITY	(omitted)	(omitted)
BAPP X 2ND QUINTILE OF INTENSITY	-0.113 (0.089)	-0.118 (0.091)
BAPP X 3RD QUINTILE OF INTENSITY	-0.144 (0.101)	-0.147 (0.103)
BAPP X 4TH QUINTILE OF INTENSITY	-0.218* (0.126)	-0.226* (0.130)
BAPP X 5TH QUINTILE OF INTENSITY	-0.267** (0.117)	-0.266** (0.119)
BAPP X 1ST QUINTILE OF PARTICIPATION	(omitted)	(omitted)

BAPP X 2ND QUINTILE OF PARTICIPATION	-0.169† (0.119)	-0.144† (0.113)
BAPP X 3RD QUINTILE OF PARTICIPATION	-0.016 (0.110)	0.015 (0.116)
BAPP X 4TH QUINTILE OF PARTICIPATION	0.037 (0.096)	0.084 (0.094)
BAPP X 5TH QUINTILE OF PARTICIPATION	0.141† (0.158)	0.218† (0.166)
TREND	-0.008* (0.005)	0.007 (0.007)
BAPP X TREND		-0.013 (0.010)
Ln(WORKERS)	1.126*** (0.321)	1.132*** (0.323)
Site fixed-effect?	Yes	Yes
Constant	-4.782*** (1.712)	-4.713*** (1.172)
Adjusted R-square	41.07%	41.11%
Observations	4,625	4,625
Mean of dependent variable before BAPP	1.338	1.338
Errors in parentheses are robust and clustered at the site level. * p<0.1, ** p<0.05, *** p<0.01 in two-tailed test. All models are estimated using an OLS panel fixed effect. †A test of equality of BAPP X 5TH QUINTILE OF PARTICIPATION and BAPP X 2ND QUINTILE OF PARTICIPATION is rejected at the 20% and 10% significance in column (1) and (2), respectively.		

Figure 13. The impact of BAPP varies according to Intensity.

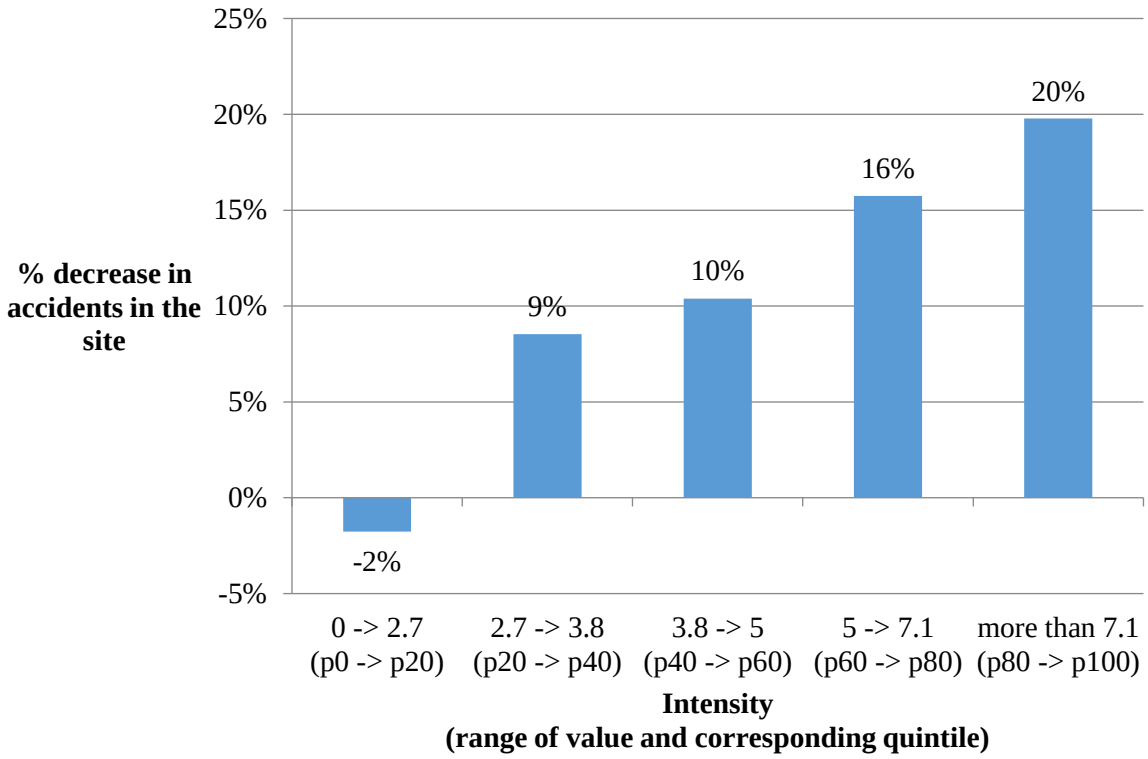
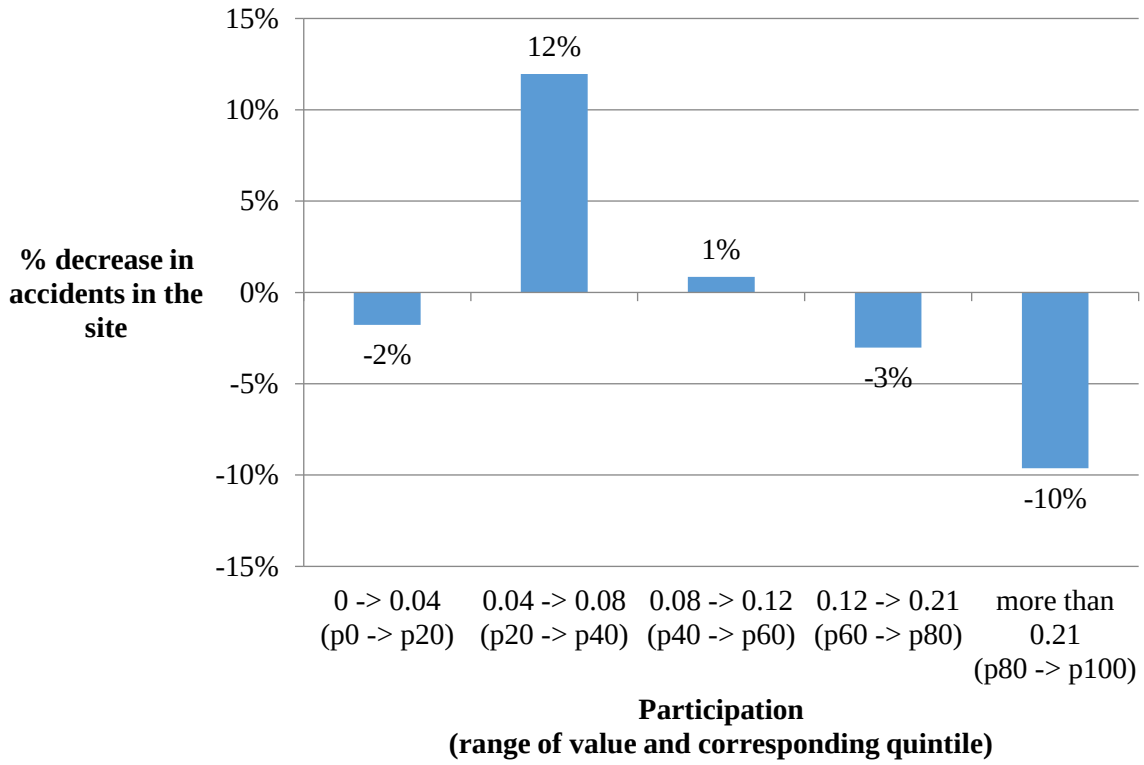


Figure 14. The impact of BAPP varies according to participation.



3.3.5. Why do large groups decrease the impact of BAPP?

To understand what was driving the negative impact of adding more observers, we collected observation level data for the 88 projects in our sample. The dataset contains 1,265,176 observations in total. For each observations we observe: site, date, name of observer, area of the site, presence of a coach, critical behavior being tracked, risky or safe behavior, and the rest of information in the sheets of observation (e.g., comments).

Using this dataset we will show that newer observers i) exert a lower cooperative effort, that is, they execute substantially less observations per month, and ii) have lower tenure as observer, making cooperation fragile as it expands. First, in **Figure 8** we display the average number of active observers per site. The number starts at the 10 observers of the committee, to end up with 50 observers by the end of the third year of BAPP. Second, in the **Figure 16** we break down the

number of active observers into five quintiles of entry order, that is, into five cohorts of observers. For all observers that participated in BAPP, we recorded the "date of entry" as the date of their first observation and using that, an "order of entry" for each observer within their site. To generate the cut-offs of the quintiles/cohorts, we use the information at the observer-month level. There are many observers that participated over the 36 months, and plenty that participated in only a handful of periods. The cut-offs were computed to separate all the observer-months entries into equal sized groups according to "order of entry". Thus, the cohorts are "weighted" by the number of months the observers were present or active. This allows to generate meaningful cutoffs that acknowledge the "importance/relevance" of the resulting cohorts[52]. The quintiles separate the total presence over time of observers according to their order entry.

Figure 16 decomposes the total number of active observers by cohort. For example, the **Figure 15** indicates that at the period 12 there are, on average, 30 active observers per site. **Figure 16** shows that these come from the following cohorts:

i. 7 observers from the 1st cohort (observers that with an entry order between 1 and 13),
ii. 6.7 observers from the 2nd cohort (observers that with an entry order between 14 and 36),
iii. 7.8 observers from the 3rd cohort (observers that with an entry order between 37 and 78),
iv. 6.3 observers from the 4th cohort (observers that with an entry order between 79 and 168),
v. 2.2 observers from the 5th cohort (observers that with an entry order between 169 or more),

The weights of the cohorts in the total number of observers of course varies according to the period of the implementation. Early in the implementation the first cohort is more prevalent, and as the implement progresses, newer cohorts take more importance in the composition of the active observers. In the last year, the largest number of belongs to the last cohort.

[52] The results we display below do not change if different criteria are used to generate the quintiles such as not weighting by active months, or weighting by the number of observations.

From **Figure 8** we also learn that the amount of rotation of observers is not small. For example, four of the nine observers from the first cohort leave the project by the end of the third year. Notably, rotation increases as we move into higher quintiles. The second quintile has roughly 7 active observers on average but the pool of observers where these active observers is drawn equals 23 observers (36-14+1). This implies a larger rotation for the 2nd quintile compared to the first. The third quintile has roughly 8 active observers drawn from a pool 42 observers, while the fourth draws roughly 9 observers from a pool of 90 observers.

Figure 15. Average number of active observers per site

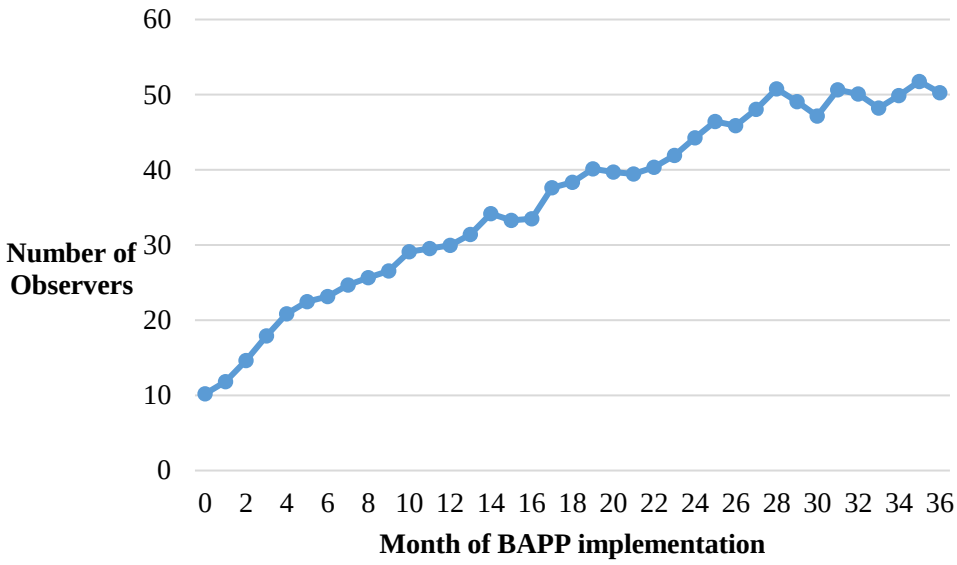
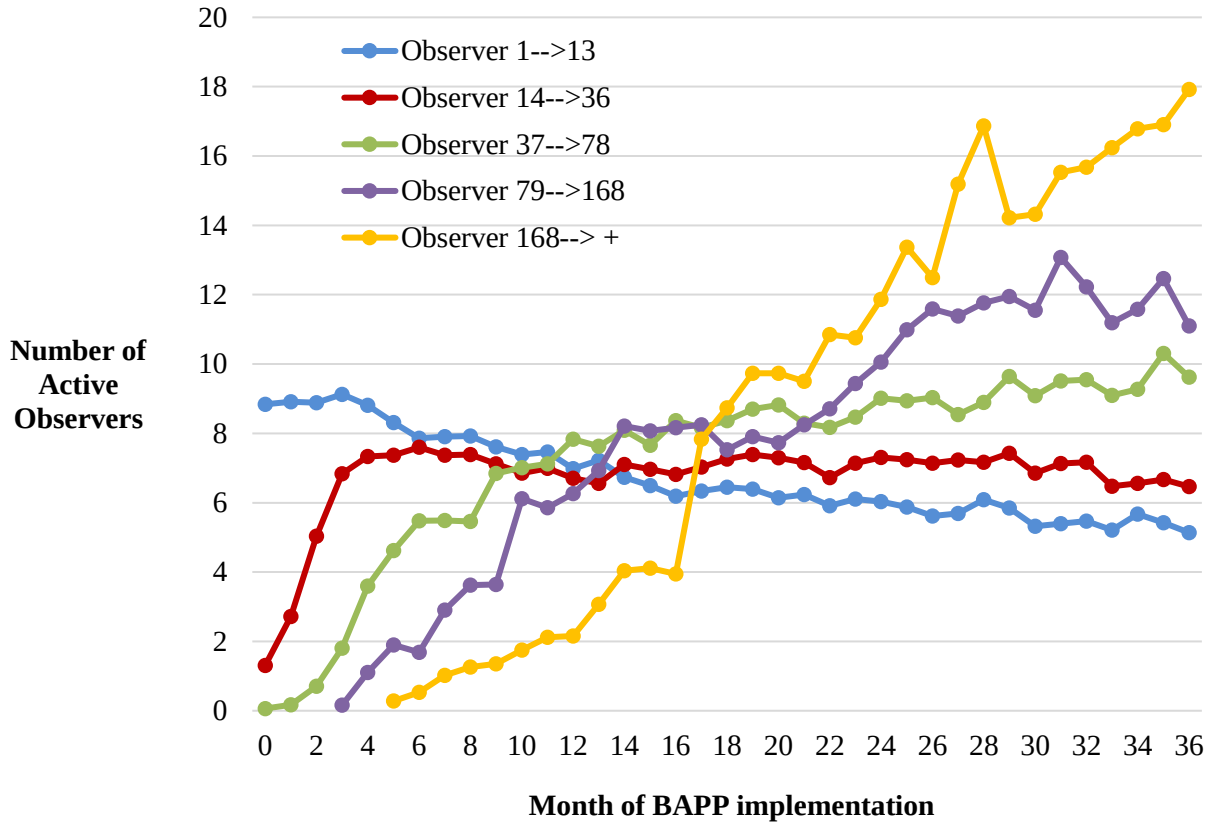


Figure 16. Number of observers per quintile of entry (or cohort)

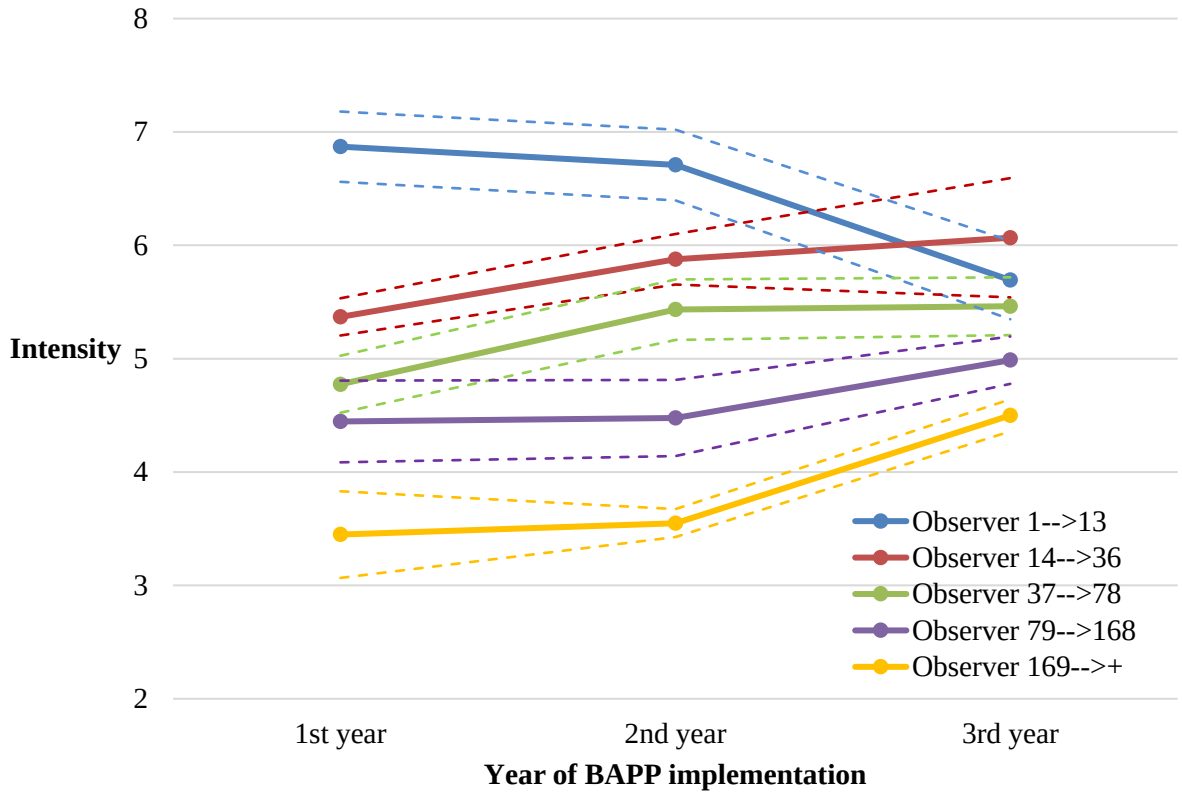


Then, for each quintile, we computed the total number of observations executed for each month and we divided that by the number of observers displayed in **Figure 16**. This yields the intensity for each quintile. The results, for which we summarize at the year level, are displayed in the Figure 17. Here we can clearly see that the intensity experiences an important drop as we move up in the quintiles. In the first year, the first quintile execute 7 observations per month, while the fifth quintile only executes 3.5. The data also shows some convergence over time. The first executes 5.7 in the third year, while the fifth quintile executes 4.5. In the graphs, the dotted lines display a 95% confidence interval. We see that the differences are statistically significant.

How does these results explain the inverted U documented in **Figure 14**? These results strongly suggests that the increase in accidents from high participation comes from a decrease in the cooperative effort of the newer observers. The newer observers display higher rotation, and thus,

less time as observers; and while they are observers, they execute a substantively lower amount of observations. This lower intensity is not captured by the model estimated in the **Table 19**. The intensity used in that model is aggregate, and its level and the variance is quite stable over the 36 months of implementation (see **Figure 7**). Instead, the effect that the inverted U is capturing is the lower intensity of the newer cohorts. Simple calculations can confirm this. Going from a participation of 10% --the peak of the inverted U-- to a participation of 20%, means going from 25 observers in month 8, to 50 observers in the second semester of the third year. From **Figure 16** we know that these 25 additional observers come in the most part from quintiles 4 and 5. At the eight period the weighted intensity is 5.9, while the intensity of the observers that will enter is no higher than 5, closer than 4 and 4.5 for the last two years. According to Figure 13 and Table 19, this decrease of (roughly) 1.5 in intensity translates into a 15% to 20% increase in accidents, which is the magnitude we observe in the inverted U.

Figure 17. Newer observers execute fewer observations



However, the analysis executed so far in this section is subject to confounds. The most important is that the quintiles are generated for the whole sample, and as such, they do not consider any site-specific elements such as size, the implementation strategy or the amount of coaching. For example, it could be that the lower intensity of higher quintiles is due to a higher participation rate: in order to achieve the aimed contact rate of 1, low intensity might be needed. In order to check whether the findings are robust to these confounding elements, we use the following regression model:

$$INTENSITY_{ijt} = b_1 + \sum_j b_{2j} \times OBS\_QUINT_{ij} + b_3 \times TOT\_OBS_{jt} + b_4 \times TENURE_{ijt} + T_t + U_j + ERROR_{ijt} \quad (9)$$

In this model we regress the number of observations of the observer i in the site j in the month of implementation t (from 1 to 36) on the quintile of the observer (the cohorts defined above), the number of observers in the site (which captures participation), the tenure of the worker (measured as the months elapsed between the month of first observations and the focal month) which control for the impact of rotation (higher quintiles have higher rotation), and fixed effects of site and month of implementation (we also used the product of site and month as a fixed effect and results remained unchanged). Sadly, we could not add observer fixed effects as the cohort of the observer is time invariant. The results are displayed in the **Table 20**. The column (1) show that the detrimental impact of higher cohorts of entry is robust to the control variables we used.

However, using cohorts has the downside that sites have different number of workers, and therefore quintiles that are defined by the distribution across sites (and not within) can be affected by the a size confound. For example, larger sites will have more observers in higher quintiles, but display lower intensity for another reason (e.g., they pursue a high participation). To accommodate this, in columns (2) and (3) we use the entry of order of the observer, and this variable, conditional on site (column 2) or site-month fixed effects (column 3) will not be affected by such concerns. Using column (3) estimates we find that the 50^{th} observer in entry order within a site displays 0.95 less observations, whereas the 100^{th} observer displays 1.8 less observations.

Table 20. Regression of intensity on entry order

	Intensity (1)	Intensity (2)	Intensity (3)
1ST QUINTILE OF ENTRY ORDER	3.056*** (0.255)		
2ND QUINTILE OF ENTRY ORDER	1.993*** (0.253)		
3RD QUINTILE OF ENTRY ORDER	1.336*** (0.184)		
4TH QUINTILE OF ENTRY ORDER	1.085*** (0.127)		
5TH QUINTILE OF ENTRY ORDER	(Omitted)		
ORDER OF ENTRY		-0.016*** (0.002)	-0.02***(0.001)
ORDER OF ENTRY ^2		0.00002*** (2.09e-06)	0.00002*** (2.34e-06)
TENURE	0.022*** (0.007)	0.036***(0.007)	0.006 (0.006)
NUMBER OF OBSERVERS	-0.006*** (0.001)	-0.005*** (0.001)	(omitted)
Month of implementation fixed-effects?	Yes	Yes	No
Site fixed-effects?	Yes	Yes	No
Site # Month of implementation fixed effects?	No	No	Yes
Constant	1.912*** (0.367)	4.965*** (0.268)	1.052
R-square	8.51%	8.46%	27.99%
Observations	91,145	91,145	91,145
Mean of dependent variable	5.28	5.28	5.28

Errors in parentheses are robust and clustered at the observer level. *** p<0.01 in two-tailed test. All models are estimated using OLS.

The implications on worker rotation extracted from **Figure 16** is also subject to confounds. We execute the same analysis as for intensity. The result is displayed in **Table 21**. For this model, we use observer tenure as the dependent variable. As we use this variable, it is crucial to include the "time implementation X site" dummies (model 2): both tenure and order of entry increase as the implementation elapses. The test that this regression performs is to assess whether the order of entry takes away (or adds) from to the "automatic" relationship between time of implementation and tenure. The results indicate a very robust and large negative relationship between the ranking

of entry and tenure. The 50th observer in entering BAPP has 5.7 months of lower tenure, equivalent to 60% of the mean tenure. Tenure is relevant: in the appendix 3.7.2 we show that higher tenure in new/additional observers improves the impact of BAPP on accidents. These results strongly suggest that cooperation becomes shakier as the number of observers expands and that this diminishes the impact of BAPP.

Table 21. Regression of tenure as observer on order of entry

	Tenure as observer (1)	Tenure as observer (2)
ORDER OF ENTRY	-0.119*** (0.0006)	-0.119*** (0.0005)
ORDER OF ENTRY ^2	0.0001*** (1.33e-06)	0.0001*** (1.19e-06)
NUMBER OF OBSERVERS	0.013*** (5.48e-04)	(omitted)
Month of implementation fixed-effects?	Yes	No
Site fixed-effects?	Yes	No
Site # Month of implementation fixed effects?	No	Yes
Constant	1.153*** (0.148)	0.415*** (0.084)
R-square	75.12%	79.90%
Observations	91,145	91,145
Mean of dependent variable	9.33	9.33
Errors in parentheses are robust and clustered at the observer level. *** p<0.01 in two-tailed test. All models are estimated using OLS.		

With these results, the picture is more complete: BAPP generates a positive average impact, but this impact is limited as the number of observers expands. The new additional observers provide a lower cooperative effort and stay less time, diminishing the capacity of cooperation to deliver results as it expands.

3.3.6. Evidence on the benefits of an "interaction structure"

The implementation of BAPP provides freedom to the site to try different tactics and strategies. In Appendix 3.7.2 we exploit this variance and execute a thorough analysis on the heterogeneity of

impact of BAPP. In this section we will focus on a specific aspect of the BAPP implementation that provides initial evidence on the importance of "interaction structures" for cooperation.

Drawing for our conversations with DEKRA and ACHS we realized that some sites specialize their observers on different areas of a site, executing its observations mainly in that area. Also, some observers naturally specialize, even if the site doesn't define a policy in that direction. The area specialization of observers has two main effects:

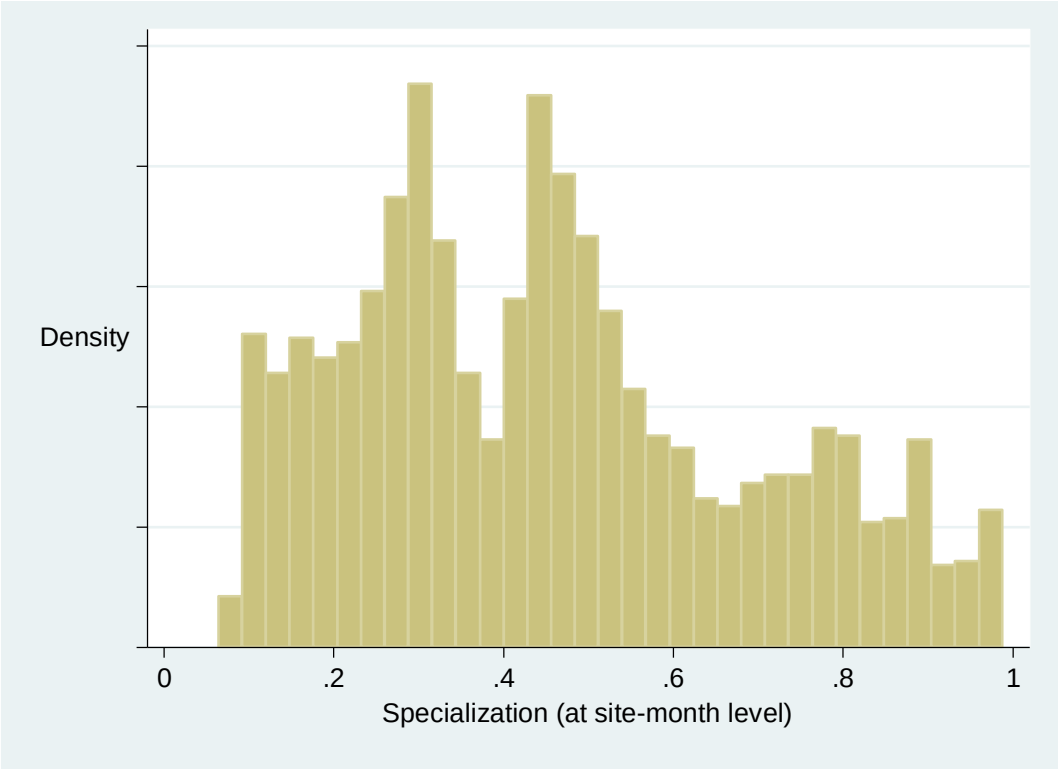
i) A "learning effect": the observer learns about the tasks being performed in the area and therefore can provide better and deeper feedback to workers,

ii) An "interaction structure" effects: the observer now interacts with a reduced set of workers and this increases the frequency of interaction with observed workers and therefore the capacity of direct reciprocity to sustain cooperation. If the area where the observer specializes is 1/5 of the site, then the frequency of interaction can increase by a factor of 5.

The observation sheet displays the different areas of the site where the observer can execute a particular the observation. The set of areas is pre-defined by the team of observers that set ups the implementation and stays fixed throughout the implementation. We measured specialization as an HHI index: the sum of the squares of the share of total observations by the observer in each area of the site[53]. Then we averaged this for a site for every month (this generated some variation over time as the pool of observers changes in the site). In the **Figure 18** it can be appreciated that specialization at the site-month level has a wide distribution: some sites hardly specialize their observers while many do.

[53] We also used a measure that computes the HHI monthly and the results did not change. We prefer to use the HHI across the whole tenure of the observer because HHI monthly is by construction higher, as only a handful of observations are executed each month.

Figure 18. Distribution of specialization



To assess the impact of specialization we executed the following model:

$$ACC_{it} = b_1 + b_2 \times BAPP_{it} + b_3 \times TREND_{it} + b_4 \times (BAPP_{it} \times TREND_{it}) + b_5 \times BAPP_{it} \times Z_{it}$$

$$+ b_6 \times \ln(WORKERS_{it}) + U_i + ERROR_{it} \quad (10)$$

This is the same model as equation (1) with the addition of the interaction of BAPP and the vector Z. In Z we add our variable of interest, specialization, but also control variables. We control for the participation and intensity using dummies of high/low participation (based on the median), average observer tenure on the site, and accumulated experience. Accumulated experience is measured as the accumulated number of observations per observer (prior to the focal one), and then averaged it at the month-site level. This variable captures the degree to which learning can improve the impact of BAPP (Argote and Miron-Spektor, 2011) and can help to disentangle the "learning" from the "interaction structure" mechanisms of specialization. Tenure allows to

distinguish between simply the passage of time and actual experience (although positively correlated at .5, there is plenty of variance in experience for different level of tenure).

In Table 22 we present the results. In column (1) we find that specialization can dramatically boost the impact of BAPP. If we move from the 10^{th} percentile in specialization (0.17) to the 90^{th} percentile (0.82), the decrease in accidents produced by BAPP is higher by 0.42 accidents a month which is roughly a third of the dependent variable and more than half of the average impact of BAPP documented in **Figure 8**. Regarding experience, it doesn't affect the impact of BAPP, suggesting that learning is not crucial in BAPP[54]. Our conversations and interviews with DEKRA and ACHS suggest that this is the case. Learning how to execute observations properly is not tricky. Getting the worker to change its behaviour is the difficult part; and although accumulated experience might help to engage better with workers, the incentives that the worker faces might play a larger role in social dilemmas.

Column (2) helps in disentangling the mechanisms of specialization. We fail to find any interaction between specialization and accumulated experience. If specialization were to be operating through the "learning effect", one would have expected to find some movement with this variable. Instead, the lack of significance provides suggestive evidence that specialization operates through the "interaction structure" effect: specializing means that the interactions become much more frequent between the observer and workers, providing room for direct reciprocity to take hold. In section 3.4 we will test this correlational result using a field experiment.

[54] The result of tenure is not expected. In the appendix 3.7.2 we show that this effect is driven by tenure of earlier cohorts of observers, which according to interviews, if they stay too long, they might get demotivated from exerting cooperative effort too long, while many co-workers don't do their share. For new observers (or latter cohorts), tenure is marginally beneficial.

Table 22. The role of specialization on the impact of BAPP

	Accidents - OLS (1)	Accidents – OLS (2)
BAPP	0.210 (0.134)	0.212 (0.166)
TREND	-0.033** (0.014)	-0.033** (0.014)
BAPP x TREND		
BAPP x SPECIALIZATION	-0.649** (0.212)	-0.655** (0.291)
BAPP x HIGH_INT	-0.283*** (0.106)	-0.283*** (0.106)
BAPP x HIGH_PART	0.226** (0.098)	0.226** (0.097)
BAPP x TENURE	0.034** (0.014)	0.034** (0.014)
BAPP x EXPERIENCE	0.001 (0.001)	0.001 (0.002)
BAPP x SPECIALIZATION x EXPERIENCE		0.000 (0.005)
Ln(WORKERS)	1.230*** (0.331)	1.230*** (0.330)
Site fixed-effect?	Yes	Yes
Constant	-5.247*** (1.757)	-5.246*** (1.755)
R-square	43.30%	43.30%
Observations	4,447	4,447
Mean of dependent variable before BAPP	1.338	1.338

Errors in parentheses are robust and clustered at the site level. * p<0.1, ** p<0.05, *** p<0.01 in two-tailed test. † indicates p<0.01 in a two-tailed joint t-test (this test is required as there is multicollinearity between BAPP, TREND and their interaction). The joint t-test on BAPP and BAPP x TREND is also statistical significant at p<0.05.

3.3.7. Independent evaluation by DEKRA

The research project coincided, and ran in parallel, with an internal assessment and redesign of the BAPP methodology by DEKRA. This assessment effort was independent of this research project and a redesign team spearheaded it. They gathered data from several sources, with an important share coming from the experience of consultants and clients.

The assessment and new vision of the redesign team was very closely aligned to the assessment coming out if this research, particularly the picture drawn in the last sections. The fact that cooperation –that is, effort by observers– breaks down once participation increases too much, played a prominent role in their own assessment. Similarly, there was close convergence on the role of specialization, tenure and the other levers discussed in the appendix 3.7.2. We view this convergence as increasing the credibility of the findings presented so far.

3.3.8. Alternative framework: Snowdrift game

We interpreted the detrimental effect of high participation under the light of a standard public goods framework. Under this framework defection is dominant and thus, generating an increase in frequency in cooperation is always fighting against self-interested defection. In this case, as the expansion of observers becomes too large, BAPP lacks the mechanisms (or interaction structures) to support the cooperative effort of additional observers. If this is a correct framework, then BAPP is always in peril if a mechanism is not in place. And the more effective the mechanisms are, the larger the penetration of cooperation would be in the population. We took this approach in the design of the experiment present in section 3.4.

However, there is another social dilemma framework that could apply to our setting, and if so, it might imply different interpretations of our findings as well as different policy prescriptions and experimental design. This alternative social dilemma is the Snowdrift Game (see Hauert et al, 2006) (also known as the Hawk-Dove game in the context of aggression and conflict). To illustrate this game we borrow from Hauert et al (2006, p. 196): "consider two drivers on their way home that are caught in a blizzard and trapped on either side of a snowdrift. Each driver has the option to remove the snowdrift and start shoveling or to remain in the car. In contrast to the Prisoner's Dilemma, the best choice now clearly depends on the other driver: if the other cooperates and starts shoveling, it pays to defect and remain in the car but if the other defects, it is better to shovel and get home than to wait for spring. Similar situations may occur whenever the act of cooperation creates a common good that can be exploited by others, i.e. whenever the benefits of cooperation accrue not only to the partner but also to the cooperator itself." Hauert et al (2006) explain than in a group of N-persons, the snowdrift situation can be generated when: i) the cooperator benefits

himself as part of his cooperative act and this amount is larger than his cost C (where B is the benefit provided to the whole group and B/N is the per-capita benefit that the co-operator provides other players and himself), and ii) the benefits that an additional cooperator generates is lower than the benefit the previous cooperator generated (i.e., benefits are decreasing, or there is a "discount"). As a consequence of i) and ii), there is a threshold in the number of observers below which the net benefit of cooperating is higher than defecting. Therefore, result of an N-person snowdrift game is that there is an internal equilibrium with a stable mix of cooperators and defectors. This is in contrast with a public good setting, where defection always dominates, that is B/N is always larger than C (unless a mechanism for cooperating is provided)[55].

If a snowdrift situation is present in our setting, then a stable share of observers (i.e., co-operators) would expected to be a natural outcome of BAPP. Then, the findings of the successful diffusion of BAPP within sites, of the reduction in accidents and the change in culture becomes, perhaps, a bit "less impressive". And therefore, the problem of solving the cooperation breakdown in high participation becomes less challenging. Furthermore, it could be argued that keeping participation low in BAPP is a natural solution, one that is better not to fight against.

How can we test whether our setting is a snowdrift situation? In the following we consider the first game we introduced in section 3.2.1, that is, the game played among workers in becoming observer.

First, we check the first condition stated by Hauert et al (2006). As an initial step, we will check whether collective effort is beneficial (i.e., the benefits when everyone cooperates surpasses the cost, namely B > C). This is a necessary condition for a public goods game. The second condition,

[55] A closely related game to the Snowdrift game is the "Volunteer game" (Diekman, 1985; Archetti, 2009). In this social dilemma, a benefit is provided to the group by only one or a few of the group members. In the two person case, the snowdrift and volunteer game are the same. In the N-person case, the volunteer game is very similar to the snowdrift game, yielding the same conclusion: in equilibrium, there is a stable group of co-operators. However, the volunteer game has a framework that is simultaneous; a game that admits a sequential interpretation, like the snowdrift game, fits much better our setting. In the latter, the additional player cooperating provides a lower benefit than the previous one while; in contrast, in the former the benefit for the group is fixed and all the players decide at once whether to volunteer or not to produce it.

in contrast to the N-person snowdrift, is that B/N is always lower than C (namely defection is always individually dominant).

Let's consider the benefits assuming that cooperation is extensive. We assume extensive to mean 50 active observers, the average that BAPP achieves after three years (on an average site of 250 workers and 50 active observers, this means that approximately 2 or 3 times as many additional observers were in place at some point.) From **Figure 8** we derive that over the first three years of BAPP, the reduction in accidents is 25% across the site (including the observers)[56]. Assuming a yearly rate of accidents of 4% for developed countries (Hämäläinen et al, 2006), this means that they get a 1 percentage point reduction in the likelihood of experiencing an accident. The economic value of an accident has been documented to be in the range of US$20,000 to $70,000 by Viscusi and Aldy (2003)[57]. This means that the value of the benefit to the observers is in the range of US$400 to $1400 (used in Viscusi and Aldy, 2003, year 2000 US$). Considering inflation of 2% a year, this leads to a range of US$549 to $1,922 in 2016. On top of this, there are direct costs of accidents (cost of accident-indexed insurance rates, lost production, material lost, cost of replacing workers) which amount, approximately, to 50% of the value of an accident (Brahm and Singer, 2015). Thus, the benefit coming from a reduction in expected accidents amount to the range of US$824 to $2,883. Productivity also improves if safety prevention policies improve (Brahm and Singer, 2015). Although the elasticity is hard to establish, a 0.3 has been found in some studies (Fernández-Muñiz et al, 2009). If we assume that one third of the productivity improvements flow to workers as increased wages (as documented in the vocational training literature; see Conti, 2005; Dearden et al. (2006); Colombo & Stanca, 2014), then the 25% in accidents translates into 2.25% increase in wage. Given a 2016 median wage of $30,000 USD for the workers in the US (where the majority of BAPP implementations have been executed), this translates into an additional $450 USD. This puts the benefits on the range of $1,273 USD to $3,333 USD. This is the benefit that

[56] If one assumes that 100% cooperation (everyone is an active observer) and accidents are fully eliminated, the conclusions do not change. As we'll see, these would continue to hinge upon crucially on the size of the private "reputational" benefits (see footnote 58 below).

[57] This is correlational evidence. Experimental evidence from Brahm et al (2018) suggest correlational studies underestimate the true value of accidents.

each worker receives from having 50 active observers cooperating (with many more having been observers at some point in time).

Let's consider the cost of cooperation. The workers devote on average 5% of their time to BAPP (see section 3.2). Given the wage of $30,000 USD, the cost they experience is close to 1,500 USD. From this, one needs to deduct the extra benefit that observers receive by being an observer in terms of lower accident rate: given that they learn by observing, we estimate in our experiment that observers obtain 15% additional reduction in likelihood of having an accident (see section 3.4.8). This translates to $165 USD and $576 USD of additional benefit, and therefore to a cost in the range of $924 USD to $1,335 USD. In addition, one needs to deduct the private reputational benefits that the observer receives from becoming an observer. If one assumes that these benefits amount to 2% of wage, then the final cost ranges between $324 USD and $735 USD. This is the cost that each observer experiences as he cooperates.

Therefore, comparing benefits from collective effort versus its costs, we see that it is beneficial (B > C). We can also see that individually defection might always be dominant. If the benefit is divided by 250 workers, we have the benefit ranging between $5 USD and $13 USD (B/250)[58], much lower than cost which ranges between $324 USD and $735 USD. These calculations fulfil the conditions for a public goods game: collective cooperation is beneficial, but the individual incentive is to defect. In order to meet the first condition of Hauert et al (2006), the private reputational benefit would have to offer at least an additional 2% to 2.5% of wages on top of the 2% we already assigned. This would bring down cost to zero, or negative, and therefore fulfil Hauert's condition i).

[58] B/N is the value of the expected reduction in accidents that a single observer generates on co-workers from executing his observations. This includes himself, as he benefits from being observed by other observers (if he is the only observer, a "self-observation" would be needed as assumption). For computing B, we assumed a reduction of 25% based on "extensive cooperation" of 50 active observers. If one assumes 100% cooperation, and that accidents would disappear as consequence, then the benefits of $5 USD to $13 USD would need to be multiplied by a factor of 4, obtaining $20 USD and $52 USD. This different assumption doesn't alter the consequence regarding the type of game being and the crucial role of the reputational benefits have on that regard.

This leads us to the crucial second condition. There are two values that are subject to changes as more observers participate: the private reputational benefits and the impact on accidents of the additional observers. First, the private reputational benefits (set at 2% above for all observers equally), are clearly higher for earlier observers and then they diminish as more observers participate in BAPP, potentially being eliminated altogether for the Nth observer. This would suggest that, if the private reputational benefits are higher than 4.5% for the first observers, then a snowdrift game might be in place. The second variable is whether the impact of observations on accidents diminishes as more observers are within the site. This is not to be confused with lower intensity by newer observers, which is a behavioural outcome: what needs to be assessed is the impact of the effort put by the additional observer on accidents. To test this we exploit the change in participation across months. The test we execute is to interact participation and intensity. The idea is to see whether the impact of the execution of observations (of "effort") changes if you have a few or many observers. If there is a negative interaction between intensity and participation, it means that the marginal benefit of executing observations is lower when you increase the number of observers. This would be a sign of decreasing benefits of the additional cooperative behaviour by new observers.

Table 23 displays the result of the interaction between intensity and participation. Instead of quintiles, we simplify the analysis by using dummies of low-high intensity and participation using their median as a cut-off. In the column (1) we find the same result as in **Table 19**. In column (2) we fail to find any statistical significance for the interaction term. The impact of changing intensity is the same whether you have a few or many observers. This strongly indicates that there is no decreasing return of the cooperative act of executing observations for the additional observer.

Thus, considering only the benefits on private reputation, the "first game" in BAPP would be a snowdrift. In contrast, if one only considers the benefits on accidents, the game would be a public good. In weighting these two elements, reputation and accidents, we believe the second one is more important. First, accidents represent the direct benefit of BAPP, while reputation is an indirect benefit. Second, in our estimations above, the benefits from accidents are between 1.5 and 3 times larger than the baseline benefits from reputation (2%). Third, reputational benefits of more

than 4.5% of wage for the initial observer in the joint committee might be plausible. But for the new observers, this represents a stretch. Therefore, we conclude that it is more likely that we are placed in a public goods setting. And if a snowdrift is in place, the threshold is probably the committee; after that, it is very likely that B/N is lower than C. Just like in biology, it is very difficult to definitively tell which type of game is being played in a social interactions (Hauert et al, 2006).

We are not arguing that workers and observers actually perform these calculations. These calculations simply estimate that the costs and benefits are within the range of social dilemmas and that these costs will eventually play out over time in any given company that it is implementing BAPP. Consequently, behaviour should slowly follow. Evolutionary theory (which is the broad framework we are applying) indicates that a replicator dynamic is required to change behaviour over time in a population. Replicator dynamic is a process by which every period –every month in our case– a small percentage of workers changes strategy (i.e., defection or cooperation) towards the strategy that is faring better (either defection if interaction structures are absent, or cooperation if they are in place). In our case, we can envision several processes that make the benefits and costs gradually observable (even quantifiable) for workers/observers. From previous implementations and the execution of the focal one, workers may learn from managerial behaviour (and from peers) the level of reputational benefits attached to being an observer. In the same way, given the large impact of BAPP in accidents, the workers will eventually realize the improvement in safety (we have witnessed this on the ground!); and of course, workers do attach a (subjective) value to improved safety. Time away from everyday tasks will also be costly to the company and the worker. Many companies have tight and detailed budgets, typically broken down to specific areas in a site. In this case, someone (the worker, his/her supervisor or the area manager) will feel the pressure of time not spent according to budget. This will put pressure on observers as they "take time away" to execute observations. In our interviews, observers frequently feel this cost. In sum, we argue that cost and benefits will surface over time in various ways, this affects how different workers are feeling/assessing their position (or strategy), and if a position is advantageous, workers will gradually adopt it.

Table 23. Interaction of intensity and participation

	Accidents (1)	Accidents (2)
BAPP	-0.163 (0.117)	-0.160 (0.119)
BAPP X HIGH INTEN.	0.186** (0.089)	0.194* (0.103)
BAPP X HIGH PART.	-0.169* (0.094)	-0.160 (0.121)
BAPP X HIGH INTENS. X HIGH PART.		0.015 (0.101)
TREND	-0.001 (0.007)	0.001 (0.007)
BAPP X TREND	-0.013 (0.010)	-0.013 (0.010)
Ln(WORKERS)	1.105*** (0.319)	1.105*** (0.319)
Site fixed-effect?	Yes	Yes
Constant	-4.569*** (1.690)	-4.569*** (1.692)
Adjusted R-square	41.13%	41.12%
Observations	4,625	4,625
Mean of dependent variable before BAPP	1.338	1.338
Errors in parentheses are robust and clustered at the site level. * p<0.1, ** p<0.05, *** p<0.01 in two-tailed test. All models are estimated using an OLS panel fixed effect. mn (1) and (2), respectively. A joint t-test on BAPP and its interaction is significant at 10%.		

The second public good game that is played in BAPP is among observers. Here the choice is between consummate (high) and perfunctory (low) effort. Very similar calculations can be made to the first game (for example the benefits are the same). We reached stronger conclusion: the game 2 it is more likely a public good (instead than a snowdrift) than game 2. The main argument is that consummate effort provides a less strong and clear signal for reputation building than the fact of becoming observer (e.g., effort is not public, low effort can have "many" justifications).

3.4. Evidence from experiment

In the previous section, we document that the beneficial impact of BAPP is affected by the breakdown of cooperation when the group of observers grows large. Following this finding, we set out to conduct a field experiment with an *intervention* that might restore cooperation as groups

grew large. First, we introduce the setting where we carried out the experiment. Then we describe the three experimental treatments, explaining their implementation and logic. Then, we present the take-up and analyze the impact of the experiment on several outcomes: number of observations (intensity), coaching, risky behaviour, accidents, and likelihood of becoming an observer.

3.4.1. Setting

We executed the experiment in the years 2017 and 2018 in Chile. We collaborated with the Chilean Safety Association (ACHS) and one of their clients, SODIMAC. ACHS is a private non-profit organization that provides services in occupational safety and health (OSH). These services include prevention, medical treatments when accidents occur, and pensions and subsidies when a worker becomes handicapped. ACHS is one of the three non-profit organization providing OSH services in Chile (there is a fourth that is state-owned). These firms are regulated and monitored by the labour and health branches of the Chilean government. Half of the Chilean firms and workforce are affiliated with ACHS, making ACHS the market leader. Firms are mandated to contribute a percentage of their payroll (on average 1.2 percent) to ACHS or one of its competitors as an insurance premium. This percentage has a common sector component and a firm-specific component, both of which are tied to historical safety performance. ACHS is mandated to devote at least 12% of its budget to prevention services, of which the bulk is spent on training (at companies, in open classrooms, and using e-learning).

ACHS partnered with DEKRA in 2012 (the partner was the company BST at that time, which was then acquired by DEKRA) in order to implement BAPP in their affiliated firms. DEKRA provided deep training to ACHS personnel for several years, generating the capability to deliver BAPP as designed and executed elsewhere. This included the training and mentoring of a cadre of BAPP consultants within ACHS, sharing handbooks, guidelines, IP and software. To secure the success

of the partnership and the integrity of the BAPP delivery, DEKRA relocated some permanent staff within ACHS[59].

With the support of DEKRA, the researchers approached ACHS, which agreed to support the execution of an experiment, modifying the BAPP methodology within one of their clients. The natural candidate among their clients was SODIMAC[60], which subsequently also agreed to the project. SODIMAC is a home improvement stores company that has operation across South-America. In Chile, they employ 20,000 employees and own about 75 stores scattered across the country. A SODIMAC store typically employs between 200 and 350 workers. SODIMAC had already implemented BAPP in 5 stores and a distribution centre, all of which started in 2014. In 2017, they announced the implementation of BAPP in 4 new stores, which we were allowed to influence in an experiment from June-2017 through June-2018 in a staggered way (multiple stores don't start BAPP implementation at the same time)[61]. BAPP implementations are coordinated by the SODIMAC OSH department at the company's headquarters in Santiago. ACHS consultants work in tandem with this department and the store managers to plan and execute BAPP. The store manager becomes the sponsor of BAPP. The enabler and the site committee report directly to the store manager.

3.4.2. Experiment design

Stores, consultants and guidelines

We executed the experiment in four stores, two located in Santiago, the "La Reina" and "Huechuraba" stores, one located in the south of Chile, the "Temuco" store, and one in the north

[59] One big difference between BAPP implementations in ACHS and those normally executed by DEKRA, is that firms affiliated to ACHS do not pay the cost of the BAPP implementation (which is very costly). Just like other prevention services, ACHS finances BAPP with the insurance prime paid by firms. We believe that this, if anything, can play against the success of BAPP, as payment typically provides an extra motivation by top management to justify their investment. In this sense, BAPP in Chile –and our experiment– provides a better setting to test the "for the workers by the workers" spirit of BAPP (or using our theoretical parlance, the condition of voluntary cooperation).

[60] One of the authors had previously conducted research within SODIMAC, analyzing the relationship between productivity and safety. This research project was sponsored by ACHS's research foundation, FUCYT.

[61] In 2017 they also embarked into 3 additional implementations in distribution centres. In mid-2018, they added 4 additional stores. We did not intervene these BAPP implementations.

of Chile, the "Antofagasta" store. These stores have average BAPP-relevant workforces of 258, 268, 334, and 234 workers, respectively, and each has a store manager who is the sponsor of BAPP[62]. The stores where selected by SODIMAC in the third quarter of 2016 with a plan to implement in the first semester of 2017 and to start observations mid-year.

The stores were assigned dedicated BAPP consultants from ACHS executing the implementation. There were three consultants in total (the two Santiago stores shared the same consultant). We worked closely with these consultants, who executed the experimental treatments on the ground as part of the implementation. First, we sent and discussed with them the experimental guidelines. These guidelines included the context of the research, the design of each treatment, a detailed implementation protocol (which provided details on how to execute the treatments), a communication protocol (which provided details on how to communicate the project and how to reply to questions that might arise), and materials (e.g, the slides to use).

We describe the implementation as we introduce each treatment below. Regarding the communication of the research project, communication activities were precisely marked. In the 1st month, the consultant informed the store manager that, as part of the delivery of BAPP, some small changes would be introduced in the methodology in order to support a research project carried out by the University of Cambridge, which was sponsored by all three partners DEKRA, ACHS and SODIMAC. The same message was delivered to the enabler and the starting team of observers in the 3rd month, after each was constituted. In the 3rd month, the enabler and the team were also asked to answer a short and voluntary personality and social preferences survey (explained below). In the 4th month, treatments 1 and 3 were explained to them (the latter only to the two stores that received it). Importantly, for all these communications instances, the three consultants used the same powerpoint slides carrying the exact same message. We emphasized the importance of following the guidelines and the scripted messages.

[62] BAPP excludes supervisors, area/line managers, senior managers, etc. from observations. As indicated above, BAPP is "by the workers, for the workers".

Treatments

The three treatments were designed during the last quarter of 2016. Treatment 1 is the baseline and is applied to all four stores. Treatments 2 and 3 aimed to explore conditions that can boost (or hinder) the impact of treatment 1 and were applied to only two stores each. **Table 24** displays which store received which treatment. Each treatment profile was randomly assigned to the stores (i.e., the assignment of the columns of **Table 24**). Treatments 1 and 2 are a within-store, while treatment 3 applies to the whole store. Below we describe the treatments, first describing the details of their design and their execution, and then explaining their logic and intended effect.

Table 24. Distribution of treatments across sites

	Antofagasta Store	Temuco Store	Huechuraba Store	La Reina Store
T1: Group structure	X	X	X	X
T2: Identity		X		X
T3: Reputation			X	X

Treatment 1: "Group structure". This baseline treatment is central to the experiment. With this treatment, in each of the four sites, we created a group structure for half of the observers in the starting team. This structure was designed as follows: suppose the starting team had n observers plus the enabler. Then half of the observers (n/2) were randomly chosen, and each of those chosen was assigned a share of 1/n of the total number of workers in the store as a "group" [63]. In case of an odd number of observers, we selected the even number below the mid-range (max minus min divided by 2) (e.g., in a case of 9 observers, we only selected 4 at random). A selected observer was allowed to execute observations only among the workers assigned to him ("his group"). Equally, workers assigned to an observer were allowed to be observed only by their assigned observer. Workers were informed that they were part of a group but not about who the other

[63] The enabler also executes observations. This rule allowed to have the same number of workers per-observer (which include the enabler) both in treatment and control.

members of the group were. The identity of the workers in a group was known only to its assigned observer. Thus, Treatment 1 generated small anonymous groups around each of one half of the observers. What about later added new observers? If a worker was enrolled from one of these groups, then he would also be bound to observe the workers only within his group. The consultants implementing the treatment were instructed not to tell the committee observer originally assigned to a group to act as a guide or leader for the new observers in its group (however, this might, and was allowed to, arise naturally).

The remaining observers that were not selected, plus the enabler, were allowed to execute BAPP observations freely across all remaining workers that had not been assigned. Thus, in the control group, BAPP was implemented in its regular form, without any pre-defined structure on who could observe whom. Note that the enabler was allocated to the control group within the site. As the enabler had a different status and role, we decided not to include him in the randomization. However, enablers do execute many observations (typically more than the rest). Below, we provide details on how we dealt with this complication in the econometric analysis.

How were the groups implemented? First, in the 4th month of implementation, when the starting team was being trained to execute observations (see section 3.2), the BAPP consultant communicated that, as part of the research, some randomly chosen observers would be focusing their observations on a subset of the workers of the site (also randomly chosen). The consultant used a lottery box with folded small pieces of paper, of which only half were marked as "selected for the treatment". All the observers had to select one piece from the box, after which they had to open it and read it aloud. Then, we paired the selected observers with their respective groups (of workers to be observed). To do so, the consultant placed several lists on the table – one for each group – that contained the names of the workers included in the group. These lists were numbered and put on the table facing down. The selected observer had to pick one list, which would be his/her assigned group. These lists were prepared by the research team beforehand and sent to the consultant prior to his/her visit to the site. The workers were allocated to the lists randomly, stratifying by gender, age, tenure and task (we coded the workers' tasks, e.g., "cashier"). To produce the lists, we used the site's most recent worker rosters as provided by SODIMAC

(typically one or two months before the month of the assignment). As part of the communication protocol, the consultant motivated randomization by explaining that it assured that no one would be penalized by or benefit from having a special set of workers to observe (i.e., groups were not biased)[64]. In order to communicate to the workers in a group that they had a specific observer assigned to them, a set of letters was printed and handed out to the selected observers. The observers were instructed to introduce themselves and hand out the letters to all the workers in their group within a month or at the first observations (whichever came first). This letter is reproduced in Appendix 3.7.3. The message of the letter was the following: a brief introduction to BAPP; an introduction of the role and name of the assigned observer; a notice to only accept observations from this assigned observer; and an invitation that the worker him/herself could become an observer in the future. (In Treatment 2, we added extra elements to this letter.) This message of the letter also played a role in enforcing the compliance of the groups as the implementation progressed (Section 3.4.4 below addresses treatment take-up in detail). Each observers in the control group was also given a list; it contained all the workers that were not assigned to a group. The observers in the control group could observe workers only from this list.

Stores experience a non-negligible rotation in their workforce (about 5% per month). This required frequent updates to the lists and letters. On average, we updated the lists every two months (see the details in **Table 25**). In these updates, the newly joining workers were randomly assigned to the groups or the control (again stratifying the assignment). The lists were updated accordingly and sent via the consultant to the store enabler, who distributed them to the observers in treatment. Also, the letters were printed and delivered to the new workers that were assigned to a group.

The consultants monitored the execution of the treatments as they visited the sites over subsequent months. The consultants did this mainly by asking the enabler and selected committee observers

[64] Also, the communication protocol of the treatments stated that if workers asked why this treatment was being generated, the consultant had a specific answer to provide (which occurred once), which indicated that DEKRA and ACHS wanted to study whether having small groups or a large one was better, and that a-priori there were good arguments for both: small provides high focus but low flexibility, but large provides low focus but high flexibility.

whether the groups and lists have been used (and reminding them if compliance was low). Overall, the compliance observed during these monitoring instances was positive.

The group sizes are summarized as follows. A store had on average 10 observers in the committee and 250 workers. Thus, roughly 5 observers and 125 workers were randomly matched in groups of 25 workers. The remaining observers and workers experienced a standard implementation, without group structuring. Across four sites, we had about 20 observers in treatment and 20 observers in control (before the addition of new observers), as well as 500 workers in treatment and 500 workers in control. **Table 25** summarizes the groups and their size for each store (as well as additional stats regarding the implementation). **Figure 19** and **Figure 20** show the evolution of the number of observers per store and per treatment condition (excluding enabler).

Table 25. Implementation details of each store

	Antofagasta Store	Temuco Store	Huechuraba Store	La Reina Store
Workers subject to BAPP observation	233.5	333.6	257.7	268.3
Number of observers in starting team (including the enabler)*	10	10	12	11
Number of active observers May-18 (including the enabler)	22	27	24	19
Number of groups*	4	4	5	5
Average number of observers per group ‡	3.2	2.8	2.5	2.6
Average number of observers per group in May-18 ‡	4.7	2.7	3	3
Average number of workers in groups	28.0	41.9	24.7	25.9
Number of workers in control	121.5	166	134.2	138.8
Month of 1st observation	Jul-17	Jun-17	Oct-17	Aug-17
Months of lists and letter update**	Aug-17, Oct-17, Dec-17, Jan-18, Mar-18, Apr-18	Aug-17, Oct-17, Dec-17, Jan-18, Mar-18, Apr-18	Oct-17, Dec-17, Jan-18, Mar-18, Apr-18	Aug-17, Oct-17, Dec-17, Jan-18, Mar-18, Apr-18,
Month of entry and number of new observers enrolled	Oct-17 (9 obs.), Feb-18 (8 obs.), May-18 (5 obs.)	Oct-17 (9 obs.), Jan-18 (8 obs.), Feb-18 (9 obs.), Abr-18 (6 obs.)	March-18 (7 obs.), May-18 (8 obs.)	March-18 (6 obs.), May (6 obs.)

Notes: (1) for the number of workers and observers we display are the averages all the lists that were handed out on the implementation and they include the observers in each group/control. (2)* After the starting team of observers was trained and assigned to treatment they had to go out and execute observations. However, some observers might not execute them and quit BAPP in the first or second month. This happened in three stores. In Antofagasta, Temuco and Huechuraba, one observer assigned to a group quitted (we probed whether it was the treatment that caused this, but this it wasn't clear as other elements were present as well in their decision). After it was clear who wasn't quitting, we corrected the lists as follows: if the observer that quitted was part of a group, their workers were randomly assigned to the other groups; if the worker was part of control, the control list wouldn't be changed. We did this in order to avoid excessive changes in list and, given the enabler as a default in control (who doesn't quit), to be conservative on the sizing of groups (i.e., not to favor treatment 1 with smaller groups). One example: Temuco. Originally we had 5 groups and control and thus 11 observers (including enabler). We had 33.4 workers per observer. However, we lost one observer assigned to a group. Thus, the new number of workers per observer in treatment changed to 33.4 * 5 / 4 = 41.9 (3) ** if the updated was in, for example October, that meant the workers in the store we used in the update were those present at the end of that month. We then sent the update around the 10th day of the next month, in the example 10th of November. (4) ‡ we compute the average without considering the months where the groups was constituted by only one member (i.e., the committee observer appointed to it). The average includes the committee observer.

Figure 19. Evolution of the number of observers in each store

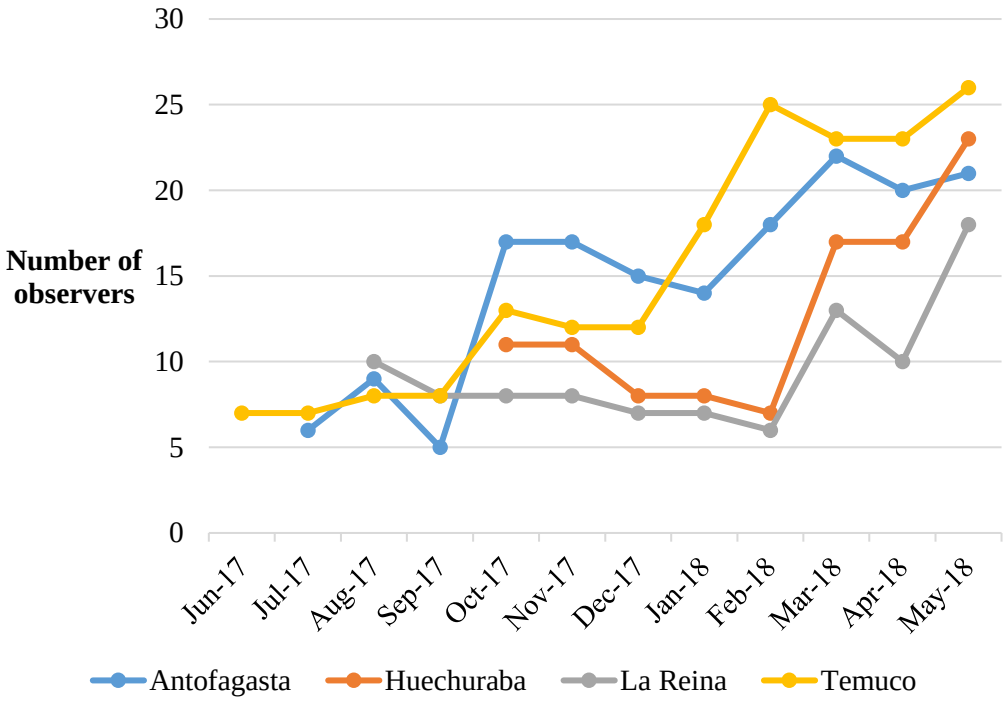
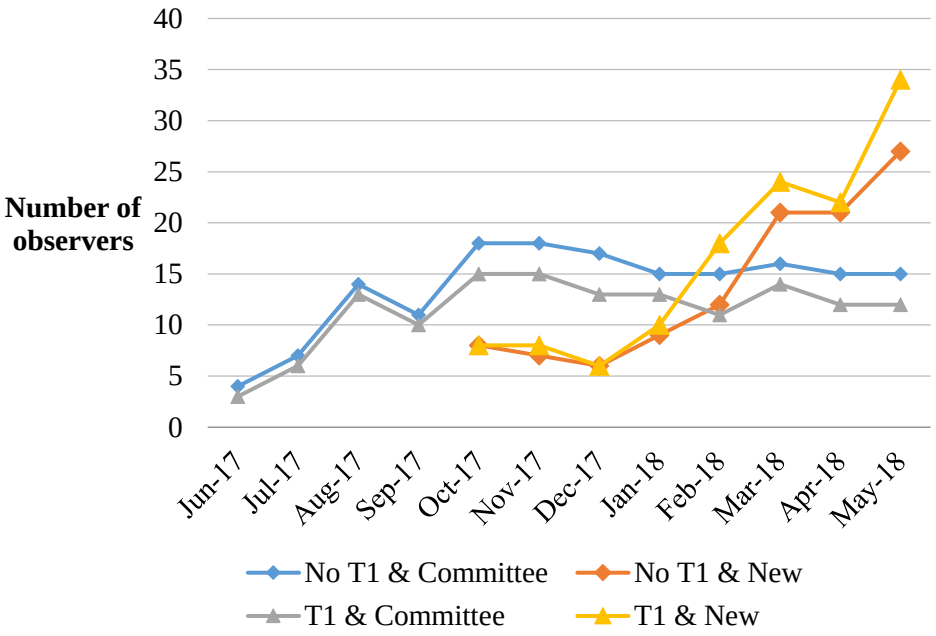


Figure 20. Evolution of the number of observers in each condition



Rationale of Treatment 1.

In order to predict the impact of this treatment, we analyze how it changes the three social dilemmas from section 3.2.1: a public good game among workers in becoming observers, a public good game among observers in providing level and sustainment of effort, a prisoner's dilemma among an observer and a worker in the event of an observation. In particular, we explore how the treatment can facilitate one mechanism that is regularly used to sustain cooperation in such dilemmas: direct reciprocity.

We start with the second game. Treatment 1 limits the detrimental effects that an increasing number of observers places on the incentives to exert effort. Consummate effort among a group of observers could be sustained by direct reciprocity, that is, the threat of effort withdrawal by a partner (another observer). However, it is well known that reciprocity as a mechanism to sustain cooperation breaks down quickly as size grows (Boyd and Richerson, 1988). Absent formal targeted punishment (as BAPP is voluntary and therefore does not include "firing" or "disciplining" of low effort observers), reciprocity in public goods is limited to punishment by all

the other members via withdrawing effort. And this threat becomes less potent as the group grows (Boyd and Richerson, 1988).

Creating groups in Treatment 1 changes the set of observers that plays the game, and thus the incentives. Specifically, direct reciprocity recovers potency, as threats of not exerting consummate effort now occur within the group. In a normal implementation, there are several tens of observers, while the number of observers within a group is just a handful (in the experiment, it's 3 on average), and therefore direct reciprocity can still work.

In the third game, Treatment 1 dramatically increases the extent to which a single worker interacts with a specific observer. Imagine a site of 200 workers and 20 observers (10 in the starting committee and 10 new), where each observer executes 5 observations a month, and suppose an observation selects a worker randomly each time. Then the likelihood that a worker repeats observations with a single observer in the next month is P(Being observed) * P(Same observer) = (5*20)/200 * 1/20 = 1/40 = 2.5%. In Treatment 1, in contrast, we have 5 treatment groups of 20 workers each and 1 control group of 100 workers. We assume that the 10 new observers are equally distributed to treatment and control. The likelihood of repeating an interaction with one observer in the next month in the control group is equal to P(Being observed) * P(Same observer) = 5*10/100 * 1/10 = 1/20 = 5%. In the treatment group, this likelihood becomes P(Being observed) * P(Same observer) = 2*5/20 * 1/2 = 1/4 = 25%. Thus, the likelihood of repeated interactions increases by a factor of 5 under treatment as opposed to control. This factor would further increase if observations per observer are higher under treatment (prediction for the 2nd game, see above), and a higher number of new observers enter treatment rather than control (as observing becomes more attractive, see below). This increase in frequency of interaction will boost the incentives for cooperation: the worker having a higher incentive to accept observations, be engaged and be willing to change his/her behaviour, and the observer having a higher incentive to provide frequent observations of high quality.

We now turn to the first game, the rewards of which are influenced by our analysis concluding that the second and third game would increase the incentives to become an observer. This is particularly the case for the third game. Higher repeated interactions between the observer and the worker

might also lead the worker, as a reciprocal response to consummate effort by the observer, to be more willing to become an observer if asked. In game 2, the observer may be very sensitive to high cooperation among a few observers within the group, whereas without groups, the observer may be very sensitive to free-riding among the many (new) observers. Note that calculations might not be necessary here; it could just be the operation of our evolved capacity to prefer small groups when asked to cooperate (see Bowles, 2009.)

Across the three games, Treatment 1 is likely to affect new observers more intensely than the observers of the initial committee. In section 3.3.8 we documented the role of private "reputational" benefits of becoming an observer in the benefits and costs experienced by different observers. Given that these benefits are much higher for the observers in the committee, we estimated that for these observers cooperation may well be dominant, that is, $B/N > C$ (i.e., they might be below the critical threshold in the snowdrift game). In contrast, new observers would face incentives to defect, $B/N < C$. As the members of the initial committee are more likely to cooperate than the new observers, the new observers are likely to be much more influenced by the treatment.

There is a second theoretical argument that also supports the expectation that the groups of the Treatment 1 will exhibit higher cooperation, without appealing to strategic behavior (i.e., cooperation is conditioned on past behavior). The idea of "spatial selection" proposes that even if players do not have contingent strategies but only one fixed behaviour (cooperation or defection), a population structure that generates assortment of cooperative types can make cooperators reap higher benefits than defectors. In well mixed populations (everyone has the same likelihood of playing with everyone else), cooperation without strategies cannot evolve. However, Nowak and May (1992) simulated n-agents playing a prisoner's dilemma with their immediate neighbours (as in the squares of a chessboard) and showed that pure cooperation (playing against pure defection) is always sustained in a population (but oscillates): cooperation can prevail within local clusters of cooperators. Treatment 1 may have the same effect, particularly for game 1, where contingent strategies such as reciprocity are less clear in how they operate. In Treatment 1, the workers are limited to interacting with a reduced set of neighbours (their own group). This could increase the

level of cooperation by assortment of cooperatives types within a local group. (Spatial selection may possibly also operate in game 2, but here it is less clear that the theoretical requirements are met.[65])

Treatment 2: "Identity". In the stores "La Reina" and "Temuco", we modified the letters that were given to the workers in Treatment 1, by adding two elements. First, we added a simple name to each group: "Group 1", "Group 2", etc. Second, at the end of the letter, we added a list with the names of all the workers that were part of the group (and their area/task). We display the letters in Appendix 3.7.3[66].

Rationale of Treatment 2.

In this treatment we attempted to generate identification of observers and workers with the group, which in turn could increase the incentives to cooperate. The addition of a name follows the minimal group paradigm that has its roots in social psychology (Tajfel, 1970). This line of research assigns subjects in experiments to different groups based on arbitrary elements, showing that this generates more help to in-group members (Tajfel, 1982) to view them as more trustworthy and cooperative (e.g., Kramer, 1991), and to be more cooperative (Loch and Wu, 2008). In our case, this would correspond to a group name possibly motivating workers and observers to muster more cooperation across all three games.

[65] Specifically, each observer now interacts with fellow observer of its group, instead with all on the site (particularly for new observers). We can calculate the condition that favours cooperation: if the benefit to cost ratio is higher than the average number of interactors in the group (or neighbours in the network) then cooperation can survive (Ohtsuki et al, 2006). Using the figures from section3.3.8, we can estimate that the benefit (B/N) to cost ratio (C) within a group is roughly 0.3, with an average number of interactors in the group of roughly 3. However, if the private reputational benefits increase from 2% to 4%, then the ratio can easily surpass 3 (the ratio increases exponentially with increases in reputation benefits).

[66] On the design stage, as can be seen in the preregistration document, treatment 2 also included participation of a portion of new observers in the joint committee monthly meetings and meetings of the observers within the groups. However, these treatments proved infeasible to be executed on the ground (mainly, due to time restrictions for new observers). On hindsight, this change was positive as we can now know better about a single mechanism, i.e., adding names, instead of a combination of mechanisms, i.e., names and meetings.

However, in the social psychology tradition subjects allocate resources without having to face a social dilemma (Bernhard et al, 2006). Recent research within behavioral economics suggests that the minimal group paradigm doesn't necessarily hold when subjects face conflicting incentives between individual and group welfare (e.g., Buchan et al, 2006; Charness et al, 2007). Two studies show evidence that groups require a joint history: Bernhard et al (2006) use a lab-in-the-field third party punishment game (TPPG) in tribes in Papua New Guinea and show that altruistic punishment (a cooperative act) of norm violations is much higher if the victim of the violation belongs to the punisher's group; Goette et al (2006) use lab-in-the-field prisoner's dilemma and TPPG with subjects drawn from platoon members that were randomly assigned to platoons and had just a 3 week history together: they find that cooperation increases towards with platoon membership, and that altruistic punishment is higher if the victim is a fellow platoon member. Consistent with these findings, it has been shown in the lab that identity from minimal groups has a positive effect on cooperation in social dilemmas only when there is common knowledge of group affiliation (Guala et al, 2013; Yamagishi and Mifune, 2008)[67]. Following these findings, we decided to add the names of the group members in order to generate common knowledge and allow for a history to spontaneously emerge as group members might relate to one another and recognize themselves a peers in their groups. Loch and Wu (2008) suggest that joint history can also be minimal: in a non-cooperative supply chain setting, group names significantly increase cooperation if paired with as little as a short introduction.

In consistency with these previous theoretical and empirical findings, this treatment should change the cooperative tendencies of observers and workers, and therefore, it should affect the cooperation decision in all three games discussed in section 3.2.1.

However, there is a specific context at SODIMAC and within BAPP that might cause this treatment to generate a negative reaction in the workers. In the third game we described, workers have to cooperate with BAPP in being willing to receive observations (it is not infrequent that they refuse

[67] In a more general note, this research relates to the work on the economics of identity by Akerlof and Kranton (2000, 2005, and 2010).

or are disengaged); and if they receive one, they have to engage and strive to change their behaviour. A particular worry of workers within the company is the feeling of being "spied on" and "denounced" ("ratted out") by BAPP observers. BAPP explicitly addresses such fears by emphasizing anonymity and by using the motto "no spying, no naming, no blaming". Thus, providing the names of the group members may, in the mind of workers, compromise this motto (although though the lists do not lift the anonymity condition of observations) and cause a negative reaction. This issue came up at the design stage with DEKRA and ACHS[68].

Treatment 3: "Reputation". In the stores "Huechuraba" and "La Reina", we published on the site bulletin board the number of observations carried out by all the observers in the site. At the start of each month, the research team would access the data on observations and generate a report that included: the name of the observer, his /her starting date, the accumulated number of observations until the previous month, and the monthly average of observations. This list was ranked by the average number of observations per month from highest to lowest. This list was sent, via the consultant, to the enabler of the site who would print it and publish it on the bulletin board of the site. The enabler would send, via the consultant, photographic evidence of the publication of the report. In Appendix 3.7.4, we display the report and two sample photographs.

Rationale of Treatment 3.

This treatment builds on the idea of indirect reciprocity (Nowak and Sigmund, 1998 and 2005), which suggests that cooperation can evolve if the behavior displayed in a round of a cooperative dilemma (cooperation or defection) generates reputation (good or bad) that affects how other players behave toward this player in subsequent rounds: cooperate if reputation is good, and punishment (for example, withholding cooperation) if reputation is bad. This is a powerful mechanism, particularly when interactions are anonymous so that direct reciprocity cannot operate.

[68] In SODIMAC, there was a strike that covered 30% to 40% of workers between November and December 2016. Labor relations within the company became quite tense after this strike. This may well have contributed to the feeling of being "spied on" or "ratted out" by treatment 2.

Lab experiments provide supporting evidence for this mechanism (Rand and Nowak, 2013), and recent work has shown a large impact in the field (Kraft-Todd et al, 2015).

Indirect reputation is uniquely effective in scenarios where encounters are anonymous. Anonymous encounters are rare in organizations, and yet, reputation is widely observed in practice (e.g., gossip is frequent). Therefore, it is interesting to understand how indirect reciprocity might operate in a scenario of repeated interaction, where direct reciprocity is at play. This is exactly how our treatment operates: Treatment 3, just like Treatment 2, is an interaction effect for Treatment 1. That is, we study whether Treatment 1 is boosted (or not) by the presence of indirect reciprocity. (Given that Treatment 3 is at the store level, the only meaningful analysis is with the interaction with Treatment 1, which is a within store treatment).

There is an absence of studies analyzing this interaction, and in general on the interplay between the mechanisms that sustain cooperation (Rand and Nowak, 2013) (see Van Veelen et al, 2012, for a notable exception). Roberts (2008) analyzes which of these strategies might dominate. When a player decides his/her next move contingent on past behavior of the other player, s/he can either use the "reputational" information (i.e., interaction with others) or the information coming from their mutual previous interactions. The main finding in Roberts (2008) is that direct reciprocity tends to dominate indirect reciprocity if the expectation of future interactions is high (personal information is prioritized). This finding suggests that Treatment 3 will not generate a strong interaction with Treatment 1, as it might be dominated by the improved direct reciprocity under Treatment 1.

However, an important aspect to consider for evaluating this treatment is the fact that sites typically have a report that is displayed and shared on committee meeting where the historical number of observations is displayed. This would specially affect the behavior of committee members, and less the behavior of new members. This would imply, that the treatment 3 would simply make public the private enforcement mechanisms happening at stores informally (via peer pressure for example). Then the question changes: how is it that treatment 3 might change the incentives to exert this private enforcement? For example, one might think that public enforcement of treatment 3 might substitute the private enforcement.

Given that the terrain of interactions between direct and indirect reciprocity is largely unchartered, we view this treatment as mainly exploratory, perhaps with the potential to provide evidence for new theoretical developments.

Identifying the treatment effects

While Treatment 1, our central treatment, is the only treatment in the "Antofagasta" store (Table 10), it is present and contrasted with the control, as well as examined for how it is modified by being combined with Treatments 2 and 3, across the four sites. In Appendix 3.7.5, we show calculations in support of this. The identification of Treatment 1 across all sites is important for the statistical power of the experiment.

Pre-experiment power calculations

On observations. Assuming power of 80% and significance of 5%, and using data on previous BAPP implementations in the DEKRA dataset, we calculated the number of observers in the treatment plus control conditions in a balanced experiment in order to be able to detect different effect sizes. In Appendix 3.7.6, we present the resulting curve. In our experiment, we expected to have 50 to 60 observers (on average) which would allow us to detect an effect with a minimum size of 2 or 1.8 observations (equivalent to 90% and 80% of a standard deviation, respectively). If we find a significant result below that, power would be an issue in need of exploration.

On accidents. Assuming power of 80% and significance of 5%, and using data on previous workplace accidents at the worker level from SODIMAC (we had access to that data for 2014), we calculated the number of workers in treatment plus control conditions in a balanced experiment in order to be able to detect different effect sizes in observations. In the appendix 3.7.6 we present the resulting curve. In our experiment, we expected to have 1000 workers which would allow us to detect an effect with a minimum size of roughly 20% of one standard deviation in accidents. Given that the occurrence of accidents is infrequent, 20% of the variance is equivalent to two and a half times the monthly mean of accidents in SODIMAC (~1.5 percentage points, considering all different type of accidents). Therefore, if one finds statistical significant results below this level, then power needs to be explored.

If one assumes gains from having panel data, then one could reduce the size of the detectable effect by approximately 30% to 50% depending on assumptions (Mckenzie, 2012). That is, the threshold would be 1.4 to 1 observations and ~1 percentage points in the monthly accident rate (a bit more than two means).

End of the experiment

At the outset of the experiment, it was agreed with ACHS and SODIMAC to execute the experiment until "mid 2018". Stores or enablers were not informed about this approximate date. Consultants were informed but requested not to tell any party in the stores about it. Around February 2018, it was agreed with the senior manager sponsoring the experiment, to run it the experiment until June 2018. Often, it is hard to keep end dates secret, and they might slowly percolate to enablers and the store. Thus, in order to avoid a "last-period" drop in the collaboration of the sites (e.g., compliance with treatment), we decided to communicate to the consultants in early May that the experiment would end in June 2018, but we internally committed to execute the analysis of the experiment with the data only until end of May 2018[69].

End of experiment interviews

In the month after the end of the experiment, we executed three meetings on site with the enabler, a group of 3 observers and 3 workers in Treatment 1, and a group 3 observers and 3 workers from the control. We executed a structured interview format, avoiding leading questions. The objective of these meetings was to understand qualitatively the mechanisms that generated the results we found, and whether they align with our theoretical interpretation. We also executed similar format one-on-one interviews with the consultants executing the treatments. In these interviews, in addition to probing into mechanisms, we also assessed the integrity of the implementations of the treatment.

[69] We actually experienced some signs of a drop in collaboration. For example, Treatment 3 took longer to get implemented on June-18, and the photo evidence came back more slowly.

Survey to observers

We sent an online survey to the starting team of observers and to the subsequently enrolled observer. This survey was sent immediately after the observer entered BAPP. The survey was voluntary and confidential. The survey was sent by the research team of the University of Cambridge and it included a terse explanation about the research project (i.e., not revealing the topic nor the purpose of the research). The survey contained a Big Five personality, a Dictator Game (DG), Third Party Punishment Game (TPPG) and a list of its social network within the store. The DG and TPPG were played with fictitious money, indicating to the respondent to imagine they had an endowment. The purpose of this survey is threefold: 1) check the balance of the randomization, 2) study any heterogeneity of impact that might be driven by these traits, 3) study possible differences between the observers in the starting team and those enrolled subsequently.

Ethics review board and Preregistration

This project was revised and approved by the internal review board of the Cambridge Judge Business School. The approval date of the project was 12 of December 2016 and the approval number is #16-035.

This project was preregistered in the American Economic Association's registry for randomized controlled trials. The ID of the registration is AEARCTR-0002350 and the registration was executed on July 25, 2017.

3.4.3. Datasets

We built two panel datasets to execute the analysis. The first is a dataset of observers for every month in the BAPP implementation from its start (which varied across the four stores) to May 2018. We recorded the name of the observer, the number of observations, information encoded in these observations (number of coached observations, number of CBI behaviors observed/reported, number of risky/safe behaviors), whether he/she was a member of starting committee or a new observer, and the treatment(s) he was allocated to (or control).

In the second dataset, we built a monthly panel of workers, and accidents, from January 2016 to May 2018 with data obtained from various sources. For each month, we have information about

each worker in each of the four participating stores (ID, age, tenure, gender, and job title) from SODIMAC's personnel registers. Using our randomization of workers, we assigned to each worker a treatment or control condition after BAPP started in the respective site. From the other dataset, we assigned the status of active observer (executing observations or not) to each worker in each month. This allowed us to study the impact of treatments on the likelihood of becoming an observer. To study impact on accidents, ACHS provided information about all accidents at SODIMAC, each accident indexed by the time of the accident, the ID of the injured worker, the type of the accident, and the number of lost days due to the accident. The breakdown of accident type is the following: work accidents (with or without loss of time), commuting accidents (accidents that occurred between home and work), and quasi-accidents (incidents, typically minor that did not meet the conditions to be attended by ACHS; e.g., not a workplace incident or not a real/meaningful incident). Accidents were also labelled according to whether they were first time accidents or repeat accidents (e.g., the worker injured a foot on a given day, it was treated, but two weeks later the same injury came back without a new independent event). Using this information, we indicated in our data, for each worker and each month, if he/she had an accident. We only considered first time accidents, using repeat accidents only to accurately establish the total number of lost workdays that a specific accident had produced.

3.4.4. Balance of covariates, self-selection of observers and take-up

As stated in the last section, we randomly assigned half of the observers and workers of each store to either the control or Treatment 1 condition, through the stratification of their age, gender, tenure, and job title. These entail two randomizations: workers to groups or control (executed by researchers), and selection of observers in the committee (executed by the consultant on the ground).

Table 16 demonstrates the balance of the first randomization of workers. We present the statistics of the covariates used to stratify the sample, to check for balance. Using a t-test, we find that there are no statistical differences between treatment and control groups for any of the four stores in the study, suggesting that the randomization was effectively executed. **Table 27** demonstrates the

randomization of observers. Again, we do not find any statistically significant differences between control and treatment, indicating that the consultants also executed the randomization effectively[70].

[70] Here we have 24 and 15 observers in control and treatment, respectively whereas **Table 25** shows 24 and 18. The reason for the difference is that in La Reina and Temuco, in the 4th month we lost 2 and 1 observers, respectively, that belonged to a group (two of them because of transfer to another city and two for personal motives). These observers were replaced by new observers in their groups (and in the case of La Reina, the enabler covered the group for a few months). If we add these 3 observers to the table, the results remain unchanged.

Table 26. Balance check of worker randomization, for each store in the study.

	Antofagasta Store			Temuco Store		
	Control	Treatment	Diff (p-value)	Control	Treatment	Diff (p-value)
N	153	153		110	109	
Average age	35.7	34	1.6 (0.35)	36.3	36.2	0.1 (0.91)
Share of women	49%	48%	1% (0.84)	32%	31%	1% (0.90)
Average tenure	4.9	4.7	0.2 (0.76)	8	7.7	0.3 (0.65)
Distribution of job titles						
Full-time seller	25%	30%	-5% (0.43)	35%	32%	3% (0.63)
Part-time seller	27%	23%	4% (0.46)	24%	28%	-4% (0.44)
Operator	14%	11%	3% (0.56)	13%	8%	5% (0.20)
Replenisher	9%	7%	2% (0.64)	10%	9%	1% (0.85)
Other	25%	28%	-4% (0.52)	18%	22%	-4% (0.40)
	Huechuraba Store			La Reina Store		
	Control	Treatment	Diff (p-value)	Control	Treatment	Diff (p-value)
N	122	123		126	126	
Average age	38.3	37.2	1.0 (0.53)	34.8	34.8	0.0 (0.98)
Share of women	52%	54%	-2% (0.80)	43%	43%	0% (0.96)
Average tenure	5.9	5.7	1.8 (0.78)	6	5.7	0.2 (0.75)
Distribution of job titles						
Full-time seller	22%	23%	-1% (0.88)	26%	24%	2% (0.74)
Part-time seller	33%	32%	2% (0.79)	30%	33%	-2% (0.71)
Operator	12%	14%	-2% (0.58)	12%	11%	1% (0.83)
Replenisher	10%	10%	1% (0.83)	7%	10%	-2% (0.51)
Other	23%	21%	2% (0.65)	24%	22%	2% (0.74)

Table 27. Balance check of observer randomization

	Committee members - All Stores			Committee members - All Stores (not considering enablers)		
	Control	Treatment	Diff (p-value)	Control	Treatment	Diff (p-value)
N	28	15		24	15	
Average age	40.5	44.1	-3.53 (0.29)	41.6	44.1	-2.48 (0.48)
Share of women	54%	47%	7% (0.67)	54%	47%	8% (0.66)
Average tenure	7.9	10.1	-2.2 (0.20)	8.0	10.1	-2.1 (0.25)
Distribution of job titles						
Full-time seller	46%	40%	6% (0.69)	42%	40%	2% (0.92)
Part-time seller	11%	7%	4% (0.67)	13%	7%	6% (0.57)
Operator	7%	13%	-6% (0.52)	8%	13%	-5% (0.63)
Replenisher	11%	7%	4% (0.67)	8%	7%	2% (0.85)
Other	25%	33%	-8% (0.57)	29%	33%	-4% (0.79)

The workers that become an observer might not be the same as the rest of the workers. This could affect the generalizability of our study: the impact of BAPP might be simply because the workers that become observers are different, and not because BAPP has the capacity of generating a change in behaviour in the "average" worker. This could also be the case for our treatments: it might be that they only work on the type of worker that becomes an observer. To evaluate the extent of this problem, in the **Table 28** we explore the difference in sex, age, tenure, and type of job which we had available with administrative data. In panel a) we find that observers are older and have a higher tenure than the rest of the workers of the site. However, from panel b) and c) we find that this difference is generated exclusively by the observers that are part of the committee. In panel b) we also learn that the committee has a higher share of women. The new observers are not different to the workers of the site. This pattern makes administrative sense as, in general, site managers and BAPP consultants might prefer more experienced workers to embark on a new project.

The difference between committee observers and new observers does not seems to pan out in terms of personality traits, altruism and social network. In the panels a) and b) of **Table 29**, we used the

answers of the observers that replied to the survey we do not find any difference in terms in terms of the big 5 personality traits, of the amount given in the dictator game and the size of their social network. This suggest that the criteria for selection of committee members is experience, and not personality, behavioural or social traits.

Table 28. Difference between observers and workers

	Observers Mean (standard deviation)	Workers Mean (standard deviation)	t-test (p-value) {Wilcoxon Rank sum test}
Panel a). All observers vs workers			
Share of women	0.415 (0.494)	0.404 (0.491)	0.804
Age	37.61 (11.9)	33.74 (12.21)	0.001***
Tenure	6.64 (5.46)	5.17 (1.63)	0.011**
Distribution of Job titles			{0.738}
Number	118	1,343	
Panel b). Committee observers vs. workers			
Share of women	0.55 (0.50)	0.404 (0.491)	0.065*
Age	44.39 (9.76)	33.74 (12.21)	0.000***
Tenure	10.28 (5.35)	5.17 (1.63)	0.000***
Distribution of Job titles			{0.971}
Number	38	1,343	
Panel c). New observers vs. workers			
Share of women	0.35 (0.49)	0.404 (0.491)	0.343
Age	34.38 (11.5)	33.74 (12.21)	0.644
Tenure	4.91 (4.62)	5.17 (1.63)	0.701
Distribution of Job titles			{0.699}
Number	80	1,343	

Notes: *** p-value <0.01, *** p-value <0.05, *** p-value <0.1. We used all the workers that were employed while the experiment was being conducted. We lose three observers in committee given that we filtered by the type of workers that were eligible for BAPP

observations and to become new observers (not supervisor or manager). To make an apples to apples comparison we dropped the cases of committee members that were supervisors. The result do not change if we include these back.

Table 29. Difference between committee members and new observers

	Observers members of the committee Mean (standard deviation)	New observers Mean (standard deviation)	t-test (p-value) {Wilcoxon Rank sum test (p-value)}
Panel A: Differences in administrative data			
Share of women	0.55 (0.08)	0.35 (0.05)	0.039 **
Age	43.5 (1.63)	34.22 (1.24)	0.000 ***
Tenure	9.98 (0.86)	5.02 (0.52)	0.000 ***
Distribution of Job titles			{0.990}
Number	40	81	
Panel B: Differences in the survey			
Big 5: Neuroticism	2.33 (0.07)	2.39 (0.12)	0.607
Big 5: Openness	3.91 (0.07)	3.98 (0.12)	0.584
Big 5: Extraversion	3.69 (0.07)	3.68 (0.14)	0.938
Big 5: Agreeableness	3.94 (0.05)	4.01 (0.11)	0.426
Big 5: Conscientiousness	4.23 (0.07)	4.10 (0.14)	0.369
Dictator game	5.03 (0.55)	4.29 (0.52)	0.375
Social network	6.9 (0.93)	4.70 (1.12)	0.149
Number	30	17	

Notes: *** p-value <0.01, *** p-value <0.05, *** p-value <0.1. Big 5, Dictator game and Social network were collected using a qualtrics survey. Big 5 questions are measured using a 1 to 5 likert scale. For the dictator game, we asked employees to imagine they receive an endowment of 10,000 CLP, and asked them to decide how much to give to an stranger (0, 1,000, 2,000, … , 10,000). For the social network, we asked workers to state with how many co-workers in the site they have a social relation (i.e., acquaintance, friend).

Our treatment is an Intention to Treat (ITT). The lists of workers that we distributed to observers (plus the letters to workers) might not be sufficient to secure compliance with the groups. As a

consequence, it is necessary to explore the degree to which observers complied with executing observations only in their groups. This would allow to estimate the "real" effect of the treatment, that is, its impact in case of full compliance. For this, we implemented a short survey to gather information about the take-up in each store. The survey was conducted on a tablet by the enabler in each store, and the goal was for each enabler to survey 60 workers from a random subsample of workers assigned to Treatment 1. We sent the tablets to the store once the accumulated contact rate reached 1, surveying workers between January 2018 and May 2018.

Table 18 presents the results of the survey. Averaging across stores, 92% of those surveyed indicated knowing about the implementation of BAPP in their store, and of those knowing BAPP was being implemented, 92% knew they had assigned observers to observe them. Of those who knew they had assigned observers, 78% acknowledged having received the letter from their respective observer. The second part of the survey asked how many times they had been observed and how many of those observations were made by their observers, showing that on average 85% of the observations were realized by their assigned observer. This number was driven down by Huechuraba with a low compliance of 52%, which we confirmed in our interviews (this low compliance did not arise during the monitoring executed by the consultant). For the remaining stores, the share of observations realized by assigned observers was over 91%. These numbers suggest that treatments were effectively implemented in stores, allowing the estimation of an ITT impact. This estimate will be a lower bound of the "actual" effect of the treatments.

Table 30. Survey results for take-up check, for each store in the study.

	Antofagasta Store	Temuco Store	Huechuraba Store	La Reina Store	**Total**
Total surveys	38	26	46	37	**147**
Knows BAPP is implemented in store	32	26	42	35	**135 (92%)**
Knows he has assigned observers	29	24	39	32	**124 (92%)**
Received the letter	21	19	37	20	**97 (78%)**
Mean of times observed*	2.5 (2.6)	2 (2.2)	1.8 (1.8)	1.8 (1.8)	**2 (2)**

Mean of times observed by observers*	2.1 (2.1)	1.8 (1.9)	0.8 (0.8)	1.5 (1.6)	**1.5 (1.6)**
Mean of share of obs. realized by observers*	91% (89%)	92% (90%)	52% (52%)	93% (97%)	**85% (83%)**
* Numbers in parenthesis restrict the count to surveyors who acknowledge having received the letter.					

3.4.5. Impact on observations

To study the impact of the treatments on the number of observations per observer, we use the following model:

$$SHEETS_{ijt} = b_1 + b_2 \times TREAT1_{ij} + b_4 \times TREAT1_{ij} \times TREAT2_{ij} + b_5 \times TREAT1_{ij} \times TREAT3_{ij}$$

$$+ b_6 \times NEW_{ijt} + b_6 \times ENA_{ijt} + b_7 \times TEN_{ijt} + b_7 \times TEN_{ijt} \times NEW_{ij} + v_{jt} + u_{ijt} \quad (12)$$

In this model we regress the number of observations by observer i in store j in the month t ("SHEETS") on the treatment dummies. Treatment 2 and Treatment 3 enter as interaction effects on Treatment 1. We control by the number of months that the observer has been active (TEN) in order to capture the ramp up in observations that naturally occurs when observers enter BAPP. The binary control variable NEW captures whether the observer is not a part of the starting committee. Figure 17 showed that new observers conduct systematically fewer observations. We also control for the interaction between TEN and NEW, as the dynamics can be different according to Figure 17. We also control for store and month with dummies (v_{jt}), which is needed because the stores with Treatments 2 and 3 started their BAPP implementations later and thus, given the ramp up in observations in the first two months, their exclusion would introduce a negative bias on these treatments. We also control for the enablers by identifying them with the dummy ENA. Enablers were instructed to execute observations in the control group, introducing downward bias in b_2 (enablers typically execute more observations than the rest of observers). To control for this, we decided to keep enabler in the sample while adding ENA; the alternative of excluding them from the analysis would generate sample selection bias (Heckman, 1979).

Given that Treatment 1 is predicted to generate a larger impact on new observers, we also extend the previous model in order to study this heterogeneity of impact:

$$SHEETS_{ijt} = b_1 + b_2 \times TREAT1_{ij} \times NEW_{ij} + b_3 \times TREAT1_{ij} \times COM_{ij}$$
$$+ b_4 \times TREAT1_{ij} \times TREAT2_{ij} + b_5 \times TREAT1_{ij} \times TREAT3_{ij} + b_6 \times NEW_{ij}$$
$$+ b_7 \times ENA_{ijt} + b_8 \times TEN_{ijt} + b_9 \times TEN_{ijt} \times NEW_{ij} + v_{jt} + u_{ijt} \qquad (13)$$

Model (13) splits the impact of Treatment 1 into two components, the impact on new observers and the impact on observers that are part of the committee (COM, which is equal to 1 minus NEW).

Table 19 and **Figure 21** displays the results. Column (1) indicates that Treatment 1 generates an increase of 0.97 observations, significant at 90%. This effect is equivalent to 20% and 34% of the mean and standard deviation of the dependent variable, respectively. Column (2) shows that this impact is driven by the impact of the treatment on the new observers. These observers conduct 1.38 more observations per month, a result that is statistically significant at 95%. Observers that are members of the committee display 0.58 additional observations under Treatment 1 but this is not statistically significant. New observers that do not receive Treatment 1 execute 1.60 fewer observations than a committee member, an effect size that is very similar to the difference estimated for the Dekra dataset between the first and second quintiles of observers (depicted in Figure 17). This result indicates that Treatment 1 operated as intended: it reduced the breakdown of cooperative effort as the number of observers increased[71]. Column (3) breaks down the impact of Treatment 1 on the committee member assigned to a group on two elements, when they were alone in the group and when a new observer enters the group. We see an increase in the amount of observations from having a companion in the group, however, the difference is not statistically significant. Our exit observer interviewees indicated that repetition of contact played an important role in the groups, particularly on whom other observers are "kept track off". Workers indicated

[71] The condition of being a new or a committee observer is not randomly generated. Thus, the right interpretation of our results of Treatment 1 on new observers is on those new workers that would be (self) selected to be observers under a group structure, which could be different than those (self) selected to become observers in the control. Indeed, comparing the two groups on the variables of Table 27 yielded two statistical differences: new observers are younger and have a lower tenure. **Table 11**

that having the same person observing them repeatedly made them more "committed" to follow their advice. These qualitative findings are consistent with direct reciprocity.

Adding Treatment 2 to Treatment 1 reduces the number of observations by roughly 1.5 per month, statistically significant at 95%. This means that the benefit that is obtained by creating groups and "structured growth" is eliminated if the names of the group members are revealed in the letter.

We do not find any statistically significant impact of Treatment 3 on observations. Below (in the section on "peer monitoring"), we analyse this treatment deeper in order to probe whether this seeming non-result has a plausible interpretation.

This result of Treatment 2 suggests, in the terminology of our theory discussion, that the distaste for the violation of anonymity was stronger than any identity effects that might have been generated by the Treatment 2, and this effect completely overwhelmed the impact of direct reciprocity. Our exit interviews strongly pointed at this interpretation: (partially) lifting the anonymity condition of BAPP generated a backlash from the workers. A strong argument being made was that the anonymity of observations is a cornerstone of BAPP. Providing the names of workers indeed jeopardized the BAPP promise of "no spying, naming, no blaming". This resistance was then enacted by a lower willingness to collaborate with observers by either rejecting being observed (as they are allowed to) or if observed, frowning about it and not being engaged. In terms of the social dilemmas introduced above, the distaste for being "listed" translates into a higher cost of cooperating to workers in game 3, which increases their likelihood of defecting.

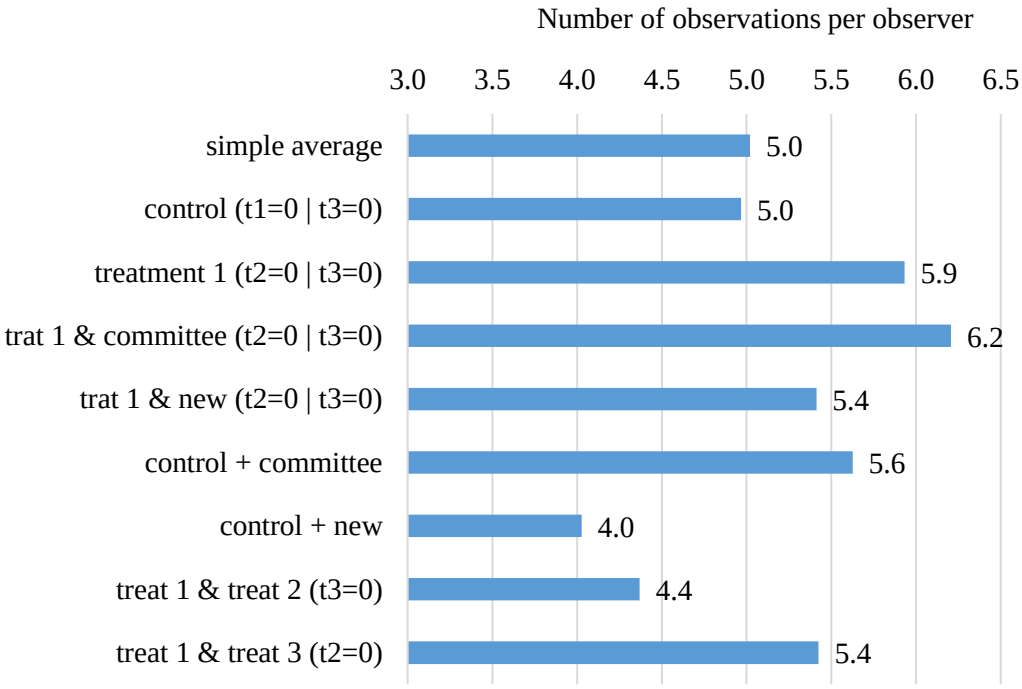
This results is novel for the literature, where transparency (broadly defined) is in general advocated because it fosters identity building or indirect reciprocity, leading to cooperation. We believe that this outcome is natural when the cooperative act entails providing a "positive" benefit to the third party, that is, it carries a neutral or positive signal for the recipient. However, in our case the recipient is told to change an *erroneous behaviour*, one which he/she has probably been advised not to incur in the past. This can generate a negative signal on the recipient if his name is not anonymous. Simply put, being "pointed at" on a mistake generates a cost, even if the "pointing" comes when receiving help. The plausibility and generalizability of this idea is an avenue for future research.

Table 31. Impact of treatments on number of observations

	Observations (1)	Observations (2)	Observations (3)
Treat 1	0.97* (0.53)		
Treat 1 x Committee observer		0.58 (0.66)	
Treat 1 x Committee observer alone			0.23 (0.61)
Treat 1 x Committee observer with group			0.88 (0.89)
Treat 1 x New observer		1.38** (0.57)	1.32* (0.57)
Treat 1 x Treat 2	-1.52** (0.67)	-1.56** (0.68)	-1.50** (0.59)
Treat 1 x Treat 3	-0.74 (0.61)	-0.51 (0.64)	-0.35 (0.68)
Enabler	3.40** (1.37)	3.28** (1.34)	3.26** (1.34)
Tenure	0.12 (0.14)	0.12 (0.14)	0.10 (0.15)
Tenure x New observer	-0.04 (0.16)	-0.04 (0.16)	-0.02 (0.16)
New	-1.17 (0.88)	-1.60* (0.91)	-1.64** (0.91)
Store-month fixed effects	Yes	Yes	Yes
Observations	585	585	585
R-square	38.95%	39.33%	39.54%
Mean (Standard deviation)	5.02 (2.82)	5.02 (2.82)	5.02 (2.82)

All regressions are estimated with OLS. Errors in parentheses: Robust and clustered at the observer level. * p<0.1,** p<0.05, *** p<0.01

Figure 21. Impact of treatments on the number of observations



Alternative explanations

Instead of direct reciprocity, an alternative mechanism that might lie behind the positive impact of Treatment 1 is the "guide" or "leadership" role that committee observers provide to incoming new observers. Without treatment, there is not a clear 1 to 1 relationship between a committee observer and a new observer, so the latter is left to their "own devices". Although our treatment protocol avoided tagging any role of "guide" and "leadership" to the committee observer under Treatment 1 (and explicitly instructed the consultants not to emphasize it), these roles could have emerged spontaneously. To test this alternative explanation, we executed a two-stage model where in the first stage we obtain a proxy for the quality of the observers in the committee before the entry of new observers, and then we plug these estimates into our basic regression of observations in **Table 31**. In the first stage, we executed a fixed effect model with the tenure controls and the store-month fixed effects and restricting the series of the committee observers under treatment 1 to the months before the entry of new observers into their group. Then, for the second stage, we created a variable "observer fixed effect" which was constructed as follows: i) for the each new observer in

Treatment 1 we assigned the fixed effect we estimated for the corresponding committee observer (i.e., the one assigned to the group of the new observer), ii) for the new entrants in the control group, we added the average of the fixed effects of the committee observers in control[72], and iii) For the committee observers we added their own fixed effect estimated in the first stage. The results show including this control does not alter the conclusions obtained in **Table 31**, if anything the results become stronger for the new observers with an increase of 50% in the coefficient (as expected from the addition of fixed effect, the coefficient of committee observers is close to zero). In particular, we found that a higher fixed effect of the committee members increases the number of observations executed by new observers both in the control and in the Treatment 1, but that the impact is higher in the former. These results indicate that quality of committee members (or their "leadership") has a role, but that this is not driving the impact we document for Treatment 1.

Another alternative explanation for the impact of treatment 1 is the idea of "Diffusion of responsibility" (Latane and Darley, 1968). This idea indicates that the willingness to volunteer effort decreases with the size of the group. (This idea is closely related to the idea that free-riding temptations increases with the group size.) The treatment 1 reduces the size of the relevant group that is responsible for helping: now the few observers of a group in treatment 1 are "alone" in observing workers. This would make the observers more willing to provide effort. However, there are three reasons that suggests that this mechanism is not driving our results. First, diffusion of responsibility is not strategic, in the sense of behaviour contingent on other's behaviour; but we do document strategic behaviour in our results (column 3 of table 29; table 31; interviews). Second, diffusion of responsibility will predict higher effort but *lower* entry of new observers; this is contradicted by our results in section 3.4.7 below. Third, we executed an additional analysis (available upon request) that splits the impact of treatment 1 according to the number of observes in the group: only the committee observer (1 observer), the committee observer plus one new observer (2 observers), the committee observer plus two new observers (3 observers), and the

[72] The results did not vary if we used another statistics such as the median or percentiles 25 and 75, or if we estimated the fixed of committee observers in control restricting their series before any new observers enter into the control condition.

committee observer plus three or more new observers (4 or more observers). We find that the impact doesn't decrease with the number of observers in the group. This contradicts the diffusion of responsibility explanation, which would predict a sharp decrease in effort.

Regarding the negative impact of Treatment 2, an important alternative mechanism could lie in the behaviour of the consultants. Given that Treatment 2 is basically an addition to Treatment 1, it could be that the two consultants that executed it – one consultant in Temuco and one consultant in La Reina –executed Treatment 1 in such a way that led to a negative outcome, and this "consultant effect" is being picked up by Treatment 2. In simple, idiosyncratic implementation by consultant by drive the results. However, several arguments and tests indicate that this did not driving the results. We describe these in turn.

First, the consultant in La Reina also executed BAPP in Huechuraba, a store that had Treatment 1 but not Treatment 2. Thus, if the execution of consultants was the issue, we would find a negative impact of Treatment 1, because in 3 out of 4 stores it would have been implemented in a "negative" way. However, we don't find this.

Second, following the previous point, we executed two regressions, one restricting the sample to the consultant in La Reina and Huechuraba, and the other by excluding the Temuco store (these are available upon request). We found the same pattern in these regressions: Treatment 1 reduces accidents and Treatment 2 increases them. This suggests that the result of Treatment 2 is not caused by an idiosyncratic implementation of Treatment 1.

Third, we executed a regression interacting Treatment 2 with the condition of being a new observer. If Treatment 2 is the effect of an idiosyncratic implementation of Treatment 1, we expect the negative effect of Treatment 2 to be smaller for new observers, as the Treatment 1 is particularly powerful on them and some of its effect should be maintained. In contrast, if the effect is coming from the workers (as it should if the effect is a worker backlash on "being listed"), there shouldn't be any difference between committee or new observers in the negative coefficient of Treatment 2. The latter is what we find: we fail to find a significant interaction effect between Treatment 2 and being a new observer (regression available upon request).

Fourth, below we present results that show that Treatment 2 has a negative impact on dependent variables that capture observed workers' outcomes (i.e., risky behaviour, accidents, and the likelihood of becoming an observer) and a null impact on the dependent variable that affect exclusively observer behaviour, namely coaching. This is consistent with workers as the driving force behind the negative effect of Treatment 2, and therefore, closer to our proposed mechanism of a "workers' backlash". On the contrary, if the influence of Treatment 2 came from idiosyncrasies of the consultant (or enabler), the impact would also be felt in the observers' coaching behaviours.

Fifth, we explored the effect of time on with the impact of Treatment 2. We find that Treatment 2 is particularly detrimental at the start of the BAPP implementation, generating approximately a backlash of two and half observations in the first couple of months. After that, the negative effect is gradually reduced so that by the end of the experiment it is small and close to zero. We also document an effect of time on Treatment 1, but of a smaller magnitude. These patterns are consistent with a backlash at the start of the Treatment 2, and then, as workers realize that the list of names was not ill-intended, they restore effort. These results indicate that the impact of Treatment 1 are cleanly felt at the end of the experiment, without a reversion effect from Treatment 2.

Clustering and power

The impact of consultants, or any other store-specific variable that affects the implementation of the treatment, is not limited to a possible bias in the coefficient, it could also affect the precision of the estimates. Common shocks within a store can generate correlations in the standard errors, and if it is not accounted for, standard errors can be too small. To accommodate for this, we executed additional regressions clustering the standard errors by store. Given that we had only four clusters, we used the correction proposed by Cameron and Miller (2015). For T1 x NEW, we obtain a p-value of 0.165 in a one-sided test, and for T1 x COM, we obtain 0.065. Thus, we lose significance in one parameter but gain significance in the other.

Abadie, Athey, Imbens and Wooldridge (2017) indicate when clustering is required. Based on an experimental design criteria clustering by store wouldn't be necessary: our Treatments 1 and 2 are

not correlated with the clusters as they are executed within stores. Therefore, inference executed at the observer or worker level wouldn't require clustering. However, according to Abadie et al (2017) clustering by stores would be necessary on grounds of sampling design. This is so because we have clusters in the population (all the stores of SODIMAC or even all the sites of BAPP) that we don't observe in the sample we have. In simple terms, to identify the findings within our sample clustering would be necessary, but to generalize beyond it, it would.

As we have discussed earlier, if we find a statistical significance of below 1.8, it is necessary to explore the power of the estimates. The ex-post power of "Treat 1 x New observer" in column 2 of Table 19 is 0.51 (assuming significance of 5%). For the interaction between Treatment 1 and Treatment 2, ex-post power is 0.62[73]. This means that in our sample, it is 51%/62% likely to detect the effect we observe (at 5%) if we assume that it is there to be found. Thus, power is not high (nor small), indicating that false negatives can happen not infrequently in a study as ours. Moreover, Ioannidis (2005) showed that insufficient power can also cause high rates of false positives, the identification of a statistically significant effect where there in reality is none. Ioannides (2005) recommends calculating the PPV statistics which reflects the likelihood that a statistically significant finding actually reflects a true effect. In our case, the PPV for "Treat 1 x New observer" equals [0.51*R/(0.51*R+0.016)], where 0.51 is the power, 0.016 the statistical significance, and R is the ratio of "true relationships" to "no relationships" in studies similar to this one (R can be very low in fully empirical and a-theoretical fields such as genome-disease association studies). If we assume R=0.25, then PPV is 89%. Arguably, R equal to 0.25 might be conservative, given the theory and the evidence from the DEKRA dataset that support our experiment. If R=0.5, then the PPV equals 94%. Compared to the values provided by Ioannidis (2005) our PPV is encouraging.

Impact on coaching

An additional cooperative behavior that observers can execute is "coaching" (see definition of the second game in section 3.2.1). This refers to observations where an observer is accompanied by

[73] We used an N of treatment equal to 16 (the average number of observers we had in "new x treatment 1") and 44 for the N of the control group (i.e., the rest of the observers). Roughly, we had 60 observers on average in our experiment.

and gets advice from a more experienced observer, typically the enabler or committee members. We explored the impact of the treatments on the amount of coaching that the observers received. In columns (1) and (2) of **Table 32**, we replicate equations (12) and (13) but now with the number of coached observations as the dependent variable. We use a Poisson regression because this dependent variable behaves as a count variable (results are consistent if we use OLS). Column (1) shows that Treatment 1 increases the amount of coaching that the observers receive. Column (2) shows that new observers receive substantially more coaching, particularly if the new observer is under Treatment 1. If we assume tenure, t2 and t3 equal to zero, the impact of being a new observer without treatment is exp(1.02) = 2.77 additional coached observations, whereas the impact of Treatment 1 on new observers is exp(1.02+0.4) = 4.13 additional coached observations. In contrast, for the committee members, Treatment 1 generates only exp(0.44) = 1.55 additional observations. Figure 17 provides the expected number of coached observations setting the covariates at their mean. This figure clearly shows how the observers under Treatment 1, particularly the new ones, receive disproportionately more coaching than the rest. Overall, these results are consistent with direct reciprocity generating more consummate cooperation among observers of Treatment 1.

Opposite to what we find for the number of observations, we don't observe a negative impact of Treatment 2 on coaching. This is another piece of evidence that the mechanism behind the negative impact of Treatment 2 is related to workers' reaction rather than to an idiosyncratic implementation of Treatment 1 in Temuco and La Reina. Coaching is a phenomenon that doesn't involve the workers, it is internal to observers. If the impact of Treatment 2 we observed in Table 17 is driven by a "workers' reaction", we should not find an effect when studying coaching. If, in contrast, consultants or enablers implemented Treatment 1 badly (generating an effect that would be captured by T1 x T2), then we should find an effect on coaching (because observers are readily influenced by the idiosyncratic implementation).

Returning to the leadership question regarding the committee members, columns (3) and (4) illuminate whether coaching mediates the impact of Treatment 1 on the amount of observations: do new observers under Treatment 1 conduct more observations because of repeated interactions

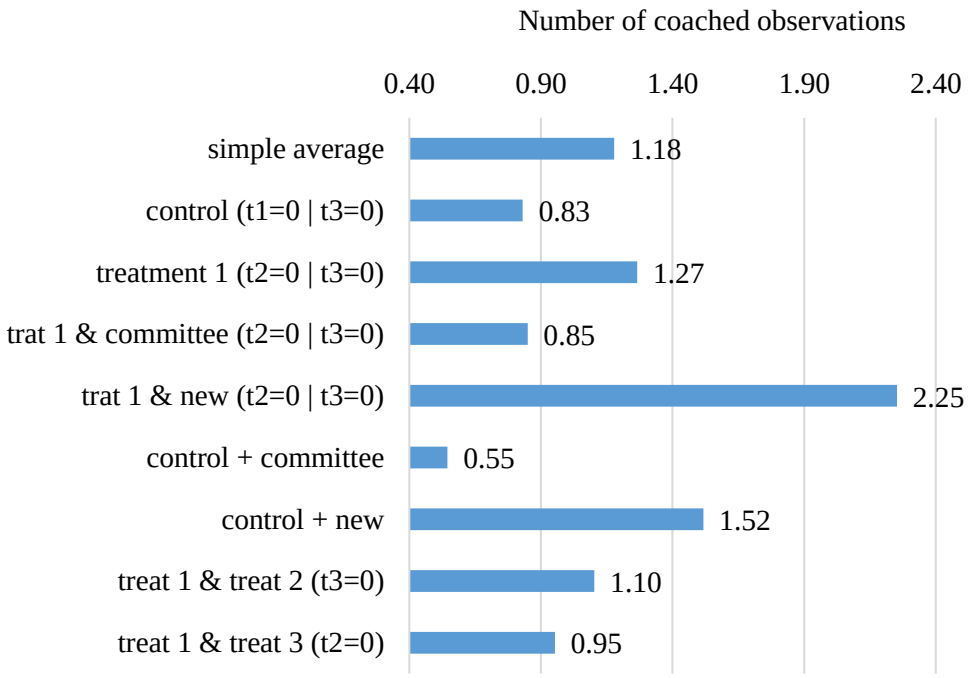
with a small group of workers and observers, or because they get more help via coaching from the enabler or the assigned committee observer? The results suggest that coaching captures only a marginal share of the impact of Treatment 1 on observations. First, column (3) establishes a positive and strong impact of coaching on the number of observations, adding observer fixed effects to improve identification, as coaching is not randomized. In column (4) we replicate the second column of **Table 31**, but adding coached observations as a control. We see that the coefficient of "Treatment 1" drops from 0.58 to 0.41 and that the coefficient of "treatment 1 x new observer" drops from 1.32 to 1.22. These drops indicate that the driving mechanism behind Treatment 1 is not help received as coaching.

Table 32. Impact of treatments on coaching

	Coached observations (1)	Coached observations (2)	Observations (3)	Observations (4)
Treat 1	0.42** (0.19)			
Treat 1 x Committee observer		0.44*** (0.22)		0.41 (0.64)
Treat 1 x New observer		0.40** (0.21)		1.22** (0.52)
Treat 1 x Treat 2	-0.14 (0.22)	-0.14 (0.22)		-1.52** (0.63)
Treat 1 x Treat 3	-0.27 (0.20)	-0.28 (0.21)		-0.43 (0.61)
Enabler	0.49*** (0.16)	0.49*** (0.16)		2.87** (1.19)
Tenure	0.02 (0.05)	0.02 (0.05)	0.69* (0.36)	0.11 (0.13)
Tenure x New observer	-0.38*** (0.09)	-0.39*** (0.09)	0.25* (0.14)	0.11 (0.15)
New	1.00*** (0.38)	1.02*** (0.39)		-2.17** (0.84)
Coached observations			0.56*** (0.09)	0.59*** (0.11)
Observer fixed effects	No	No	Yes	No
Store-month fixed effects	Yes	Yes	Yes	Yes
Observations	585	585	585	585
R-square	21.69%	21.69%	65.70%	44.49%
Mean	1.15	1.15	5.02	5.02

(1) and (2) are Poisson regressions. (3) and (4) are OLS regressions. Errors in parentheses are robust and clustered at the observer level. * p<0.1, ** p<0.05,*** p<0.01.
Results are robust to also including sheet as a control. We preferred not to add sheets as control since coached observations are an independent decision, and thus not fully nested in observations as flags or risks. For example, the mean of coached observations is 1.2, the 50th percentile is 1 and the 90th percentile is 3. Given that the mean of observations is 5 (and the 25th percentile is 3), for the majority of observers there are enough observations in order to execute coaching (and thus becoming an independent decision).

Figure 22. Impact of treatments on coaching



The Effect of Treatment 3 and Peer Monitoring

Thus far we have not found any impact of Treatment 3. In the assessment of behavior in the next section, we also fail to find a significant impact. This is surprising given that prior research is very supportive of the indirect reciprocity/reputation mechanism as a driver of cooperation. One possible explanation of this lack of support is that we don't test a pure reputation effect, instead, we test an interaction effect between reputation (Treatment 3) and reciprocity (Treatment 1: repeated interactions between observers of group, and between observer and workers of the group). Theoretical work suggests that the interaction of mechanisms supporting cooperation is by no means obvious (Nowak and Rand, 2013). For example, direct reciprocity and network/spatial selection may interact either synergistically or antagonistically, depending on the levels of repetition and assortment (Van Veelen et al, 2013).

A second possible explanation is that the real nature of treatment 3 wasn't simply about generating a reputation-based enforcement where no enforcement was present. Instead, Treatment 3 occurred in a context where there was a reputation based mechanism among committee observers. As

discussed above, the number of observations executed by each observer was frequently displayed and discussed at the monthly committee meetings. Our discussions with enablers and observers over the course of the experiment suggested that these committee-level discussions already generated real peer pressure on those observers that did not execute their share. In this light, Treatment 3 might well have simply made public to the whole store what was already visible in private (in the committee), and thus Treatment 3 in effect failed to generate any additional peer pressure --- the peer pressure effect of committee discussions was not realized at the outset of the experiment and therefore not taken into account.

In order to test this idea, we executed the regressions displayed in **Table 33**. The variable "Low ranked in the last month" captures whether the observer is below the median of the cumulative number of observations per observer up to the previous month[74]. 42% of the variance in this variable is within observer, which allowed to plug observer fixed effects into the regression in order to improve identification (low rank is not randomized). The model analyses how Treatment 1 and Treatment 3 changed the way that "Low rank" operated in motivating observations. Column (1) indicates that a low rank in the previous month doesn't generated a significant change in observations. However, this average effect hides plenty of heterogeneity. Column (2) shows that low rank does incentivize observers to increase observations but only in the case of no treatment 1 and no treatment 3. Under Treatment 1, the positive of effect low rank disappears. We speculate that this is driven by the fact that Treatment 1 means that the observer is responsible for its own group of workers (in effect being "less comparable" to the other observers), and therefore, the threat of peer monitoring goes away. In general, peer punishment operates much better if there is a joint effort or a common pool where agents contribute. The same occurs with Treatment 3, it reduces the positive impact of low rank but with smaller magnitude and lower statistical significance. This result is consistent with "private" enforcement (display of rank acted upon in the committee) and "public" enforcement (reputation effect of a public bulletin board display) being substitutes. For example, the public list might be (implicitly) interpreted by the committee

[74] Results are consistent if we use the continuous ranking. We stick to the dummy variables as it is easier to interpret.

as a signal that the "list" now executes monitoring (for all to see), reducing the incentives of people in the committee to monitor and enforce effort levels. In our exit interviews confirmed that private enforcement was indeed a powerful force driving behaviour. However, evidence in favour of its interactions with Treatment 1 and Treatment 3 was present but not definitive.

Column (3) explores the triple interaction between Treatment 1, Treatment 3 and low rank. Here we find that the negative effect of Treatment1 is partially reverted when Treatment 3 is present. This is compatible with the "substitutes" interpretation that we have just arrived at: the observer under T1 is "independent" or "not accountable to" (not comparable with) other observers in the committee, but with Treatment 3, s/he is accountable to his/her group of workers (to whom s/he gave a personal letter), which restores (some of the) accountability.

Table 33. Impact of observation ranking and its interaction with treatment 1 and 3

	Observations (1)	Observations (2)	Observations (3)
Low rank in previous month	0.56 (0.54)	2.11*** (0.78)	2.42*** (0.90)
Treat 1 x Low rank in last month		-2.19** (0.76)	-2.73** (1.18) #
Treat 3 x Low rank in last month		-1.30† (0.86)	-1.89** (0.94)
Treat 1 x Treat 3 x Low rank in last month			1.34 (1.36) #
Tenure	1.35* (0.79)	1.04 (0.75)	0.98 (0.73)
Tenure x New observer	-0.03 (0.20)	-0.02 (0.20)	-0.01 (0.20)
Observer fixed effects	Yes	Yes	Yes
Store-month fixed effects	Yes	Yes	Yes
Observations	427	427	427
R-square (adjusted)	63.98% (47.98%)	65.51% (49.69%)	65.51% (49.69%)

Errors in parentheses: Robust and clustered at the observer level. † p<0.15 / * p<0.1 / ** p<0.05 / *** p<0.01 / # p<0.05 in a joint t-test. The results are robust to adding lagged observations as a control (this controls for a possible "reversion-to-the-mean" effect).

Moreover, **Table 34** demonstrates that these dynamics among low rank, Treatment 1 and Treatment 3 are much weaker or non-existent for new observers. We split all the coefficients by multiplying them with dummies for new and committee observers. We find that the impact of low rank on observations by new observers is half of the impact on committee observers. Only committee observers observe the ranking of observations in the monthly committee meeting. New observers are not in these meetings, so they receive less peer enforcement; they might still receive some direct monitoring, from peers in conversations and particularly from the enabler, which explains the significant but smaller coefficient we observe. In sum, lower exposure to private enforcement (or peer pressure) weakens the motivation impact of low rank, both in its direct and in its interaction effects.

Although we cannot rule out all possible alternative interpretations of these results, the effect sizes are large and nuanced, suggesting that enforcement, whether it is private (via peer pressure) or public (via a reputation mechanism) is highly sensitive to the context, such as the social structure (Treatment 1) and information availability (Treatment 3).

These results are robust to adding Treatment 2 and its interactions to the models of Table 22: we did not find any significant results. Just like the results on coaching, this lack of impact of Treatment 2 on the observers' peer pressure and public reputation again supports the interpretation that Treatment 2's impact (the negative effect of revealing names) is driven by the workers' reactions. The results are also robust to adding lagged observations, which control for the persistence of effort or a "reversion-to-the-mean" dynamic.

Table 34. Impact of observation ranking and treatments for committee versus new observers

		Observations (1)
Committee observers x	Low rank in previous month	2.21** (1.08)
	Treat 1 x Low rank in last month	-2.82*** (0.96)
	Treat 3 x Low rank in last month	-1.28 (1.13)
	Treat 1 x Treat 3 x Low rank in last month	
New observer x	Low rank in previous month	1.38*** (0.42)
	Treat 1 x Low rank in last month	-0.06 (0.62)
	Treat 3 x Low rank in last month	-0.51 (0.63)
	Treat 1 x Treat 3 x Low rank in last month	
Tenure		Yes
Tenure x New observer		Yes
Observer fixed effects		Yes
Store-month fixed effects		Yes
Observations		427
Errors in parentheses: Robust and clustered at the observer level. † p<0.15 / * p<0.1 / ** p<0.05 / *** p<0.01 / # p<0.05 in a joint t-test. The results are robust to adding lagged observations as a control.		

3.4.6. Impact on safety relevant worker behaviors

The observer has to record on the observations sheet whether the behaviors in the CBI that she/he focused on were executed in a safe or a risky manner[75]. One or more risky behaviors provide the ground for providing feedback to the worker.

[75] It could be argued that the number of CBI behaviours recorded by the observer (whether safe or risky) is a measure of observer effort on the execution of observations. We analysed the impact of the treatments on the total number of recorded CBI behaviours, conditional on the number of observations, but did not find any significant impact (see Appendix3.7.7).

Table 35 presents the impact of the treatment on the number of risky behaviors that were recorded by the observer. The models are the same as the ones in **Table 31** with the addition of the number of observations and the total number of recorded CBI items. These control are necessary because risky behaviors are spread among all workers --- the more sheets an observer completes and the more items she/he focuses on, the more risky behaviors she/he would finds.

Column (1) shows that the amount of risky behaviors is significantly lower for by workers in Treatment 1. The effect size of -0.99 behaviors is sizeable, representing a 28.5% decrease from the mean of 3.47 --- BAPP matters, in other words, observations do translate into a reduction of risky behaviors.. Columns (2) and (3) indicate that this impact is stable across new and committee observers. Again, Treatment 2 reverses the beneficial impact of Treatment 1. Again, as above, this suggests that Treatment 2 has a negative impact not because it is badly implemented (as observations still translate into reduced risky behaviors), because of a negative worker reaction to revealing the names, in the form of refusing to engage[76]. Finally, Treatment 3 is again non significant.

As an aside, it is interesting to note that the control variable "tenure" is related to a lower level of risky behaviors: this is a sign that observer experience does play a role in reducing risks.

[76] However, it could also be argued that the lower observations documented for treatment 2 in **Table 31**are the driving force for finding an increase in risky behaviour (and thus it wouldn't be the workers' reaction driving the effect). However, the inclusion of observations as a control refutes this concern: the estimates of **Table 35Table 35** is conditional on the impact of the treatments on the number of observations.

Table 35. Impact of treatments on worker behaviour

	Risky behaviors (1)	Risky behaviors (2)	Risky behaviors (3)
Treat 1	-0.99* (0.52)		
Treat 1 x Committee observer		-1.09 (0.70)	
Treat 1 x Committee observer alone			-1.12* (0.67)
Treat 1 x Committee observer with group			-1.07 (0.89)
Treat 1 x New observer		-0.89* (0.53)	-0.89* (0.58)
Treat 1 x Treat 2	1.15* (0.68)	1.14* (0.68)	1.14* (0.68)
Treat 1 x Treat 3	0.14 (0.70)	0.20 (0.75)	0.21 (0.72)
Enabler	0.76 (0.71)	0.74 (0.73)	0.74 (0.73)
Tenure	-0.08# (0.13)	-0.07# (0.13)	-0.08# (0.15)
Tenure x New observer	-0.16# (0.16)	-0.15# (0.16)	-0.15# (0.17)
New	0.62 (1.06)	0.51 (1.10)	0.51 (1.12)
CBI Items	0.02 (0.01)	0.02 (0.02)	0.02 (0.01)
Observations	0.48*** (0.15)	0.48*** (0.15)	0.48*** (0.15)
Store-month fixed effects	Yes	Yes	Yes
Observations	585	585	585
R-square	49.73%	49.75%	49.75%
Mean (per observation)	3.47 (0.69)	3.47 (0.69)	3.47 (0.69)

OLS. Errors in parentheses: Robust and clustered at the observer level. * p<0.1, ** p<0.05, *** p<0.01. # denotes p<0.1 in a joint t-test.

3.4.7. Impact on the likelihood of becoming observer

If the group interaction structure works as it appears so far, it might influence not only the number of observations that observers conduct, it might also increase the attractiveness for workers of becoming an observer. This is indeed what we demonstrate in this section.

Figure 23 and **Figure 24** examine this graphically and show that the effect is subtle at first glance. In the former, we present the evolution of the number of observers, separating in four groups

considering whether observers are from the treated group, and whether the observers are in the initial committee or whether they are new observers. In the later, we divide the new observers by the eligible workers, generating a "probability of being observer" for Treatment 1 and control. We see that as BAPP penetrates each store, we see an increase in new observers over time and that this increase is higher, albeit not much, in Treatment 1 compared to control. The committees slightly shrink over time because of natural rotation in observers. This results, which are not controlling for a potential backlash of treatment 2, suggests that Treatment 1 is affecting positively the likelihood of becoming observer and therefore, speeding up the diffusion of cooperation.

Figure 23. Evolution of the recruitment of observers in time

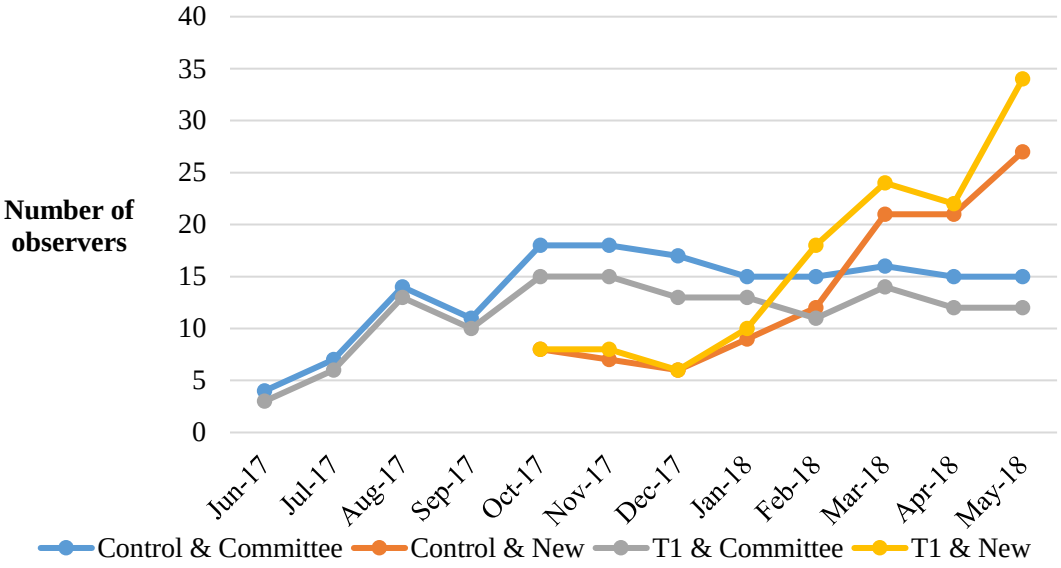
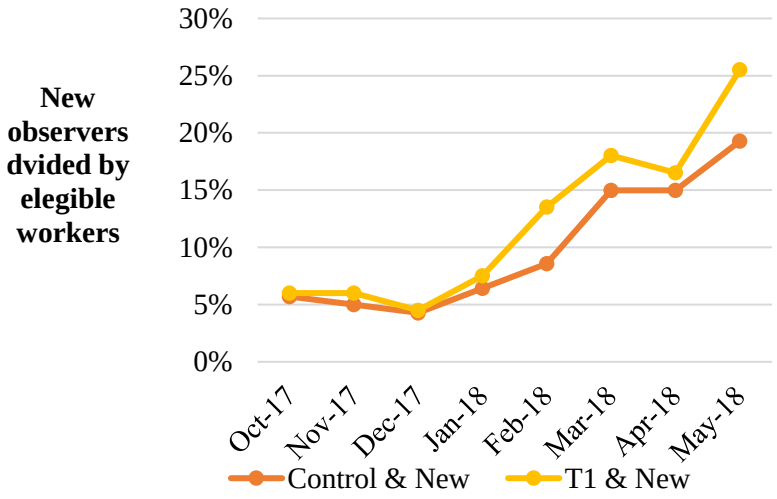


Figure 24. Evolution of the ratio of new observers to committee members in time.



These figures are merely suggestive and are subject to confounds. In order to statistically test the impact of the different treatments on the probability of becoming an observer, we run the following OLS model:

$$OBSERVER_{ijt} = b_1 + b_2 \times TRAT1_{ij} + b_3 \times TRAT1_{ij} \times TRAT2_{ij} + b_4 \times TRAT1_{ij} \times TRAT3_{ij} + X_{it} + \tau_{tj} + u_{ijt} \quad (14)$$

In equation (14) we model the likelihood of becoming an observer for all workers, excluding those who are part of the committee. $OBSERVER_{ijt}$ is a dummy variable that takes the value of 1 if a specific worker i in a stores j is an active observer in month t and zero otherwise. $TRAT1_{ij}$ is a dummy that takes the value of 1 if that worker is under Treatment 1 and zero otherwise. $TRAT2_{ij}$ and TRAT3 are defined in the same way, respectively. X_{it} is a vector of controls at the individual level for each period (age, tenure, gender and job title), while τ_{tj} are fixed effects at the store and the calendar-month level.

Table 36 presents the estimates of the model. Columns (1) and (2) uses all the months of implementation as the sample. The first column indicates that there seems to be no effect of Treatment 1 on the likelihood of becoming an observer. However, the picture changes when we

disaggregate the treatments: Column (2) indicates that the null effect in column (1) is due to Treatments 1 and 2 having opposite effects of comparable sizes. Treatment 1 increases the likelihood of becoming and observer by 1.9 percentages points over the timeframe of our experiment, which is almost equivalent to the mean likelihood of 2.2 percent. In contrast, Treatment 2 decreases the likelihood by 2.1 percentages points, reversing the positive impact of Treatment 1. Treatment 3 does not play a relevant role in the likelihood of becoming an observer. These results are consistent with the previous results on observations and risky behavior. By considering all the months of implementation, columns (1) and (2) include many early months where recruiting was non-existent or irrelevant, thus decreasing the power of the estimates. In column (3) we skip that problem by consider only the workers in May-18. The downside is that we don't consider attrition of observer, which is captured in column (2). The results are consistent and statistically stronger: Treatment 1 increases the likelihood of becoming observer by 5 percentages points (equal to the mean) while Treatment 2 reverts this, with a decrease of 7.2 percentages points.

Table 36. Impact of the treatments on the probability of becoming an observer.

	P(observer) (1)	P(observer) (2)	P(observer) by May-18 (3)
Trat 1	0.003 (0.006)	0.019# (0.013)	0.054** (0.025)
Trat 1 x Trat 2		-0.021* (0.012)	-0.072** (0.028)
Trat 1 x Trat 3		-0.009 (0.010)	-0.006 (0.027)
Individual Controls	Yes	Yes	Yes
Store-month fixed effects	Yes	Yes	No
Store fixed effects	No	No	Yes
Observations	10,879	10,879	1,072
R-squared	0.02	0.027	0.011
Mean	0.022	0.022	0.052

OLS. Errors in parentheses: Robust and clustered at the worker level. # p<0.15, * p<0.1,** p<0.05, *** p<0.01. All regressions exclude committee members. Sample restricted to months and stores with BAPP already implemented.

3.4.8. Estimating the economic effect: Impact on accidents

In this section we verify that the overall impact of the BAPP methodology on accidents is influenced by our experiment, in other words, we verify that the experiment is economically meaningful.

We first study the impact of BAPP as a whole, replicating the type of test executed in section 3.3.3. This allows us to confirm earlier findings, but more importantly, to weigh the impact of the experiment against the overall impact that BAPP has on the stores. In order to study the impact of the different treatments on the probability of having an accident, we run the following model:

$$ACCIDENT_{ijt} = b_1 + b_2 \times BAPP_{ij} + b_3 \times BAPP_{ij} \times TIME\_ELAPSED_{ij} + b_4 \times OBS_{ijt} + X_{it} + \tau_t + \gamma_j + u_{ijt} \quad (15)$$

Accidents is a dummy that takes the value of one if the worker i in the store j experienced an accident in the month t, and zero otherwise. The variables BAPP takes the value of one in the

month where observations start, and zero before that. The variable TIME_ELAPSED is a count variable that takes zero before BAPP and then 1, 2, 3, etc. for each month elapsed in the BAPP implementation of a site. Coefficient b2 capture the impact on the level at time 0, while b3 captures whether the impact of BAPP builds up over time. X is the same vector of controls as above. We control for month and store fixed effects to control for the common trend in accidents and store unobservables (in this model we can't add joint fixed effects as τ_{tj}) because it would be collinear with BAPP x TIME-ELAPSED). OBS_{ijt} is a dummy identifying that a worker is an observer after it becomes one: this variable captures the indirect impact of BAPP through the behaviour of observers (which might change beyond what is caused by being observed). In this sample, we include only the four sites of our experiment and a time period from January 2016 to May-2018. Thus, our test of BAPP's impact is a simple first-difference test, whose identification hinges on invariant store and worker unobservables.[77] In addition, we only consider workers that are subject of BAPP, that is, we exclude supervisors and managerial positions.

Table 37 presents the results. We study six different measures of accidents registered by ACHS, total accidents which we analyse in panel a), and its breakdown into work accidents (panel b), commuting accidents (panel e), and quasi-accidents (panel f). We further break down work accidents into two sub-groups: without lost working days (panel c) and with lost working days (panel d). Final, in the case of lost days, we also study the length of leave (panel g). In the first two columns of panel a) we find that BAPP reduces accidents over time, a result that is statistically significant at 90%. During its first 12 months, BAPP is correlated with an average decrease of 1.1 percentage points in the likelihood of any type of accident for that period [(0.0022 + 0.0016 x 12)/2], which is equivalent to a decrease of 9.5% in the yearly rate of accident (of any type) [0.011 / (0.0094 x 12)] (at the 12th month is twice that). By exploring the other panels, we find that the impact of BAPP is concentrated on work accidents, with no impact on commuting or quasi

[77] We left for further research the execution of more complex impact assessment methods such as i) matching the treatment stores with other similar untreated stores and execute a difference-in-differences (or a triple difference test using the supervisors and managerial workers as well), or ii) using the workers that are transferred to a new site after being treated by BAPP.

accidents. If we repeat the exercise of panel a) now for panel b) we find that BAPP is correlated with a reduction of 0.9 percentages points in work accidents in the first year, which is equivalent to 17.4% of the mean of work accidents for that period (at the 12^{th} month is twice that). These results are of a similar in magnitude as the results presented in section 3.3.3 above for the DEKRA archival dataset. However, different from the DEKRA dataset, we can show that within work accidents, BAPP is only correlated with a decrease in accidents that do not cause lost worktime. This suggest that more severe accidents might have a different data generating process, less related to worker behaviour and more to inherent workplace conditions. This is consistent with panel g) where we find no impact on length of leave (conditional on the presence of accidents with lost time). We also do not find an impact on commuting accidents, which is re-assuring as BAPP provides specific advice on workplace behaviour. Regarding quasi-accidents, they are also expected to be independent of BAPP, as these incidents are mostly not workplace accidents (e.g., an injury from weekend sport), and those that are workplace related are deemed not meaningful by ACHS.

In columns 3 and 4 of **Table 37**, we explore the indirect effect of BAPP through observers. In column 4 we add worker fixed effects to the estimation. Although this improves the estimation of column 3 by controlling for time invariant worker unobservables, it generates an additional problem of attrition bias. Since the rotation rate in SODIMAC is high (approximately 5%), the estimates in column 4 use the information of the subset of workers that have information before and after BAPP, which could change the population of workers (notwithstanding the fact that the estimates of column 3 and 4 are consistent). The results across these two columns indicate that if a worker becomes an observer, s/he enjoys an additional benefit in terms of safety in accidents that cause lost time. These workers receive the baseline benefit of BAPP –which is concentrated in accident without lost time– but then are able to reduce the incidence of severe accidents as well (those with lost time). This pattern is consistent with the safety literature, which has accumulated plenty of evidence suggesting that the reduction of severe accidents requires heavier investments. The effect is considerable: while other workers experience a 0.2% likelihood of experiencing a severe accident in any given month, column 3 indicates that observers only experience a 0.06%

likelihood. This reduction of 0.14 percentage points is equivalent to 15% of the expected likelihood of having any type of accident (0.14 / 0.94).

Table 37. Results of the impact of BAPP in accidents

Panel a)	Accidents (1)	Accidents (2)	Accidents (3)	Accidents (4)
BAPP	-0.0036 (0.0034)	-0.0022 (0.0036)	-0.0022 (0.0036)	-0.0013 (0.0036)
BAPP x Time elapsed		-0.0016* (0.0008)	-0.0016* (0.0008)	-0.0008 (0.0008)
Observer			-0.0007 (0.0031)	-0.003 (0.0039)
Ind. level Controls	Yes	Yes	Yes	Yes
Store FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Worker FE	No	No	No	Yes
Observations	30,193	30,193	30,193	30,193
R-squared (Adj. R^2)	0.0041	0.0042	0.0042	0.091 (0.024)
Mean	0.0094	0.0094	0.0094	0.0094
Panel b)	Work accidents (1)	Work accidents (2)	Work accidents (3)	Work accidents (4)
BAPP	-0.0013 (0.0022)	0.0000 (0.0023)	-0.0000 (0.0023)	0.0003 (0.0023)
BAPP x Time elapsed		-0.0015*** (0.0006)	-0.0015*** (0.0006)	-0.0006* (0.0004)
Observer			-0.0004 (0.002)	-0.0024 (0.0027)
Ind. level Controls	Yes	Yes	Yes	Yes
Store FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Worker FE	No	No	No	Yes
Observations	30,193	30,193	30,193	30,193
R-squared (Adj. R^2)	0.0034	0.0037	0.004	0.095 (0.029)
Mean	0.0043	0.0043	0.0043	0.0043

Panel c)	Work accidents without lost working days (1)	Work accidents without lost working days (2)	Work accidents without lost working days (3)	Work accidents without lost working days (4)
BAPP	-0.0024 (0.0019)	-0.0014 (0.0019)	-0.0015 (0.0019)	-0.0007 (0.0019)
BAPP x Time elapsed		-0.0011*** (0.0004)	-0.0011*** (0.0004)	-0.0009*** (0.004)
Observer			0.0011 (0.0019)	-0.0000 (0.0026)
Individual Controls	Yes	Yes	Yes	Yes
Store FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Worker FE	No	No	No	Yes
Observations	30,193	30,193	30,193	30,193
R-squared (Adj. R^2)	0.0023	0.0025	0.0025	0.089 (0.022)
Mean	0.0023	0.0023	0.0023	0.0023
Panel d)	Work accidents with lost working days (1)	Work accidents with lost working days (2)	Work accidents with lost working days (3)	Work accidents with lost working days (4)
BAPP	0.0011 (0.0011)	0.0015 (0.0012)	0.0015 (0.0012)	0.0010 (0.0012)
BAPP x Time elapsed		-0.0004 (0.0003)	-0.0004 (0.0003)	0.0002 (0.0003)
Observer			-0.0014*** (0.0004)	-0.0024** (0.0011)
Individual Controls	Yes	Yes	Yes	Yes
Store FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Worker FE	No	No	No	Yes
Observations	30,193	30,193	30,193	30,193
R-squared (Adj. R^2)	0.0018	0.0018	0.0019	0.097 (0.030)
Mean	0.0020	0.0020	0.0020	0.0020
Panel e)	Commuting accidents (1)	Commuting accidents (2)	Commuting accidents (3)	Commuting accidents (4)

BAPP	0.0003 (0.0018)	0.00013 (0.019)	0.0001 (0.0019)	0.0006 (0.0018)
BAPP x Time elapsed		0.0002 (0.0004)	0.0002 (0.0004)	0.0003 (0.0004)
Observer			0.0008 (0.0019)	0.0014 (0.0021)
Individual Controls	Yes	Yes	Yes	Yes
Store FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Worker FE	No	No	No	Yes
Observations	30,193	30,193	30,193	30,193
R-squared (Adj. R^2)	0.0013	0.0013	0.0013	0.085 (0.017)
Mean	0.0018	0.0018	0.0018	0.0018
Panel f)	Quasi-accidents (1)	Quasi-accidents (2)	Quasi-accidents (3)	Quasi-accidents (4)
BAPP	-0.0022 (0.0019)	-0.0019 (0.0021)	-0.0018 (0.0021)	-0.0018 (0.0022)
BAPP x Time elapsed		-0.0004 (0.0005)	-0.0004 (0.0006)	-0.0001 (0.0005)
Observer			-0.0013 (0.0014)	-0.0018 (0.0014)
Individual Controls	Yes	Yes	Yes	Yes
Store FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Worker FE	No	No	No	Yes
Observations	30,193	30,193	30,193	30,193
R-squared (Adj. R^2)	0.0029	0.0029	0.0029	0.080 (0.011)
Mean	0.0033	0.0033	0.0033	0.0033
Panel g)	Length of leave (1)	Length of leave (2)	Length of leave (3)	Length of leave (4)
BAPP	0.040 (0.045)	0.039 (0.036)	0.040 (0.036)	-0.016 (0.033)
BAPP x Time elapsed		0.001 (0.014)	0.001 (0.015)	-0.008 (0.005)
Observer			-0.030 (0.027)	0.004 (0.034)

Accident with lost time	13.382*** (2.905)	13.382*** (2.905)	13.382*** (2.905)	13.347*** (2.903)
Individual Controls	Yes	Yes	Yes	Yes
Store FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Worker FE	No	No	No	Yes
Observations	30,193	30,193	30,193	30,193
R-squared	0.161	0.161	0.161	0.237 (0.181)
Mean (days per accident)	0.049 (13.4)	0.049 (13.4)	0.049 (13.4)	0.049 (13.4)
OLS in columns (1), (2) and (3), we estimate lineal probability models (results are consistent if we use LOGIT). Errors in parentheses: Robust and clustered at the worker level. * p<0.1,** p<0.05, *** p<0.01.				

Now we turn to the impact of our treatments on workplace accidents. We use the following model:

$$ACCIDENT_{ijt} = b_1 + b_2 \times TREAT1_{ij} + b_3 \times TREAT1_{ij} \times TREAT2_{ij} + b_4 \times TREAT1_{ij} \times TREAT3_{ij} + X_{it} + \tau_{tj} + u_{ijt} \quad (16)$$

As before, $TREAT1_{ij}$ is a dummy taking the value of 1 if that worker is under treatment 1, and zero otherwise. The same applies for TRAT2 and TRAT3. We do not have time indices for the Treatment variables because we estimate this model using the BAPP implementation period, where every worker is assigned to a particular treatment. The vector X is the same as above. Now we introduce a store-month fixed effects (τ_{tj}) for the same reason discussed in equation (12): Treatments 2 and 3 enter latter, potentially biasing the coefficients with uneven ramp-up of observations.

Table 38 presents the results. The panels of the table mirror **Table 37**. The first column uses only Treatment 1 and the second column introduces the interactions with Treatments 2 and 3. Panels b), c) and d) show that again, disaggregation of the treatments is key --- Treatment 1 alone has a positive effect on accident reduction, which is reversed by the anonymity violation of Treatment 2. Importantly, the impact is only on accidents without lost working days, exactly the dependent variable where BAPP has his main impact (see **Table 37**). From panel e) we learn that the effects also translate into commuting accidents: Treatment 1 is associated with a reduction (p-value 0.13)

while Treatment 2 reverts this effect. This result suggests that the group structure, properly implemented, can generate benefits beyond the primary domain of BAPP. Quasi-accidents are not affected by the treatments. Regarding Treatment 3, we find that it generates a significant boost of Treatment 1 in work accidents without lost working days. This result is unexpected as no previous test of Treatment 3 has yielded an impact on observations or worker behavior. In addition to chance, this result could be explained if Treatment 3 generated a higher quality of interaction in observations, something that we have not captured so far in our tests. All of these effects are reflected in panel a) of total accidents.

Regarding the length of leave, we find that, conditional on the presence of lost time accidents, treatment 2 is associated with a reduction in the time (p-value 0.102).

The impact in panel b) of a decrease of 0.3 percentages points for Treatment 1 is equivalent to a one third of the overall BAPP impact, or a 6% reduction in the yearly mean of work accidents. This is a sizable effect, indicating that the treatment can importantly enhance the impact of BAPP. However, our pre-experiment power calculation suggests that we should not have been able to detect this effect if power is set at 80% (the minimum effect to detect was closer to 0.7 percentage points, assuming repeated observations). Our ex-post power calculation indicates that it is around 15%, which is low. The advantage in our case is that we already documented the underlying mechanisms (observations) at a fairly well powered level and with a high PPV (see section 3.4.5). In our case, when calculating the PPV suggested by Ioannidis (2005) we can use this fact and assume a larger R, which we set at 50%. Assuming this, we obtain a PPV of 65%, that is, we have a two in three chance that the statistically significant finding we uncover actually reflects a true effect.

Table 38. Results of the impact of treatments in accidents

Panel a)	Accidents (1)	Accidents (2)
Trat 1	-0.0003 (0.0017)	-0.0031 (0.0026)
Trat 1 x Trat 2		0.0072** (0.0033)

Trat 1 x Trat 3		-0.0035 (0.0034)
Individual Controls	Yes	Yes
Store-month fixed effects	Yes	Yes
Observations	11,277	11,277
R-squared	0.0071	0.0076
Mean	0.0081	0.0081
Panel b)	Work accidents (1)	Work accidents (2)
Trat 1	-0.0007 (0.0012)	-0.0030** (0.0015)
Trat 1 x Trat 2		0.0047** (0.0022)
Trat 1 x Trat 3		-0.0013 (0.0024)
Individual Controls	Yes	Yes
Store-month fixed effects	Yes	Yes
Observations	11,277	11,277
R-squared	0.0071	0.0075
Mean	0.0037	0.0037
Panel c)	Work accidents without lost working days (1)	Work accidents without lost working days (2)
Trat 1	-0.0014* (0.0086)	-0.0022* (0.0012)
Trat 1 x Trat 2		0.0034** (0.0016)
Trat 1 x Trat 3		-0.0030* (0.0018)
Individual Controls	Yes	Yes
Store-month fixed effects	Yes	Yes
Observations	11,277	11,277
R-squared	0.0044	0.0051
Mean	0.0019	0.0019
Panel d)	Work accidents with lost working days (1)	Work accidents with lost working days (2)
Trat 1	0.0069 (0.0080)	-0.0083 (0.0087)
Trat 1 x Trat 2		0.0013 (0.0015)

Trat 1 x Trat 3		0.0016 (0.0017)
Individual Controls	Yes	Yes
Store-month fixed effects	Yes	Yes
Observations	11,277	11,277
R-squared	0.0058	0.0059
Mean	0.0018	0.0018
Panel e)	Commuting accidents (1)	Commuting accidents (2)
Trat 1	-0.0006 (0.0085)	-0.0024 (0.0016)
Trat 1 x Trat 2		0.0031* (0.0017)
Trat 1 x Trat 3		-0.0001 (0.0016)
Individual Controls	Yes	Yes
Store-month fixed effects	Yes	Yes
Observations	11,277	11,277
R-squared	0.0032	0.0035
Mean	0.0019	0.0019
Panel f)	Quasi-accidents (1)	Quasi-accidents (2)
Trat 1	0.001 (0.001)	0.0022 (0.0018)
Trat 1 x Trat 2		-0.0006 (0.0019)
Trat 1 x Trat 3		-0.0028 (0.0018)
Individual Controls	Yes	Yes
Store-month fixed effects	Yes	Yes
Observations	11,277	11,277
R-squared	0.0052	0.0054
Mean	0.0026	0.0026
Panel g)	Length of leave (1)	Length of leave (2)
Trat 1	-0.056 (0.0347)	0.0098 (0.0264)
Trat 1 x Trat 2		-0.103 (0.0678)
Trat 1 x Trat 3		-0.0107 (0.0549)
Accident with lost time	12.978*** (4.438)	12.985*** (4.442)

Individual Controls	Yes	Yes
Store-month fixed effects	Yes	Yes
Observations	285	285
R-squared	0.1819	0.1821
Mean	0.045 (12.97)	0.045 (12.97)
OLS. Errors in parentheses: Robust and clustered at the worker level. * p<0.1,** p<0.05, *** p<0.01. The results are robust to dropping individual level controls as an independent variable, and to separate fixed effects for store and month.		

3.5. Discussion and Conclusion

This paper studies cooperation in large groups, where individuals bear a cost in order to provide a benefit to co-workers and the group at large. Free-riding (or defecting while enjoying the benefits of others' cooperative efforts) makes cooperation in large groups hard to build and sustain. We analyse an empirical setting that is uniquely suited to study cooperation in large groups: the host firms implement a safety methodology where a small group of workers was trained to advice co-workers in terms of workplace safety, and then the initial group expanded by enrolling new workers as additional advice providers. Our setting allowed us to study the evolution of cooperation (i.e., whether the number of cooperators increased over time), the intensity of the cooperative effort (as the cooperator group grew), and the challenges and limitations afflicting cooperation as it expanded, as well as potential solutions to the challenges.

Fine grained archival data and experimental interventions in the field allows us to dissect the anatomy of cooperation in our setting. Using a large-scale dataset of previous implementations of the methodology, we first document that cooperation is beneficial: it is associated with a reduction in accidents and an improvement in workplace culture. However, we document that as the number of cooperators grows, the additional cooperators display lower and less sustained cooperative effort, decreasing the capacity of cooperation to diffuse and to positively impact outcomes in a site. Foreshadowing our experiment, we also document that if cooperators interact within specific areas of a site, that is, interactions are structured and do not occur in a (quasi-)random way, the impact of cooperation is boosted.

We then intervened experimentally in four company sites where the safety methodology was being implemented, applying three treatments. The first treatment structured the growth of cooperation around small groups. This was expected to boost the degree of repeated interactions and the capacity of direct reciprocity to sustain cooperation (Boyd and Richerson, 1988; Axelrod and Hamilton, 1981). Accordingly, we found that this treatment enhanced cooperative effort (i.e., more advice was provided) and the diffusion of cooperation (i.e., more workers enrolled to provide advice), as well as reducing the incidence of risky behaviour and workplace accidents. Paired with the dataset findings, this result provides rare field evidence of cooperation breakdown as groups grow and of a mechanism that is able to mitigate this breakdown.

In our second treatment, we added a name to the groups of treatment 1 as well as providing the group with a list of group members. This treatment was expected to enhance identification with a "minimal group" (Tajfel 1982) and thus support cooperation, on the other hand, the publication of names went against the privacy assurances that were part of the safety methodology, which might cause suspicion and hinder cooperation. The empirical observation clearly indicated that the negative privacy effect outweighed group identification: treatment 2 reverted the impact of treatment 1. Exit interviews, and supplementary tests, indicated the strength of negative feelings about violating a cornerstone of the safety methodology: "no spying, no naming, no blaming". Workers displayed a strong distaste for being "listed" or "under surveillance," generating a cost that weighed against cooperation. This finding suggests two insights. First, the benefits of group structure interact sensitively with the context. In this case, improved group identity was outweighed by valued anonymity. Second, when cooperation includes pointing at erroneous behaviour, and this carries a cost, anonymity might be necessary for cooperation to thrive. This goes against the usual prescription of providing information on identities so that reputation mechanisms can operate.

In our third treatment, we explored how treatment 1 of direct reciprocity interacts with indirect reciprocity by posting public information on cooperative effort (Nowak and Sigmund, 1998 and 2005). Theoretical literature has merely indicated that the interaction of cooperation mechanisms is tricky and highly sensitive to local parameters (Rand and Nowak, 2013; van Veelen et al, 2012;

Roberts, 2008). In our case, we found no statistical impact of treatment 3 on the effectiveness of treatment 1, pointing at independence among these mechanisms. However, we realized during implementation that private peer enforcement of effort naturally occurs among cooperators: among the starting team, observers would frequently be informed about each other's effort, generating peer pressure. We measured this private enforcement and found that it increased effort but that both treatments decreased its impact, probably due to following motives: the group structure of treatment 1 disaggregates responsibility, reducing the legitimate reach of direct peer-to-peer enforcement; and the public nature of treatment 3 may have crowded out costly peer punishment (i.e., "it is better to let the impersonal ranking operate, instead of engaging in costly peer monitoring").

Our study is not without limitations. First, the archival dataset findings only use sites that were selected to implement the methodology we study. Although we showed that causality within the sample is likely, this might not generalize to any site, as the sites might have been chosen based on criteria that themselves influence safety behaviour. Second, power in our experiment is not high. Even if statistical significance and strong a-priori beliefs increase the likelihood of having detected a true effect, replication of our findings is required to be conclusive. Third, although we present a plausible interpretation for the negative impact of treatment 2, we cannot definitively rule out alternative explanations. Instead, we showed, using interviews and several tests, that our interpretation is likely. Finally, the findings around treatment 1 and 3 and their interaction with private enforcement are only suggestive and exploratory. All of these limitations represent good avenues for future research.

Beyond the detailed field evidence of cooperation in large groups, our study contributes to the economics of organization (Gibbons, 2018) and strategic organization (Puranam, 2018). First, we illustrate the general idea of "interaction structure" as a mechanism that supports and sustains cooperation in large groups. When a large number of individuals interact, a structure is required in order to favour cooperation: who is paired with whom and how they interact plays a crucial role in generating cooperation. Appropriate structures, plus a replicator dynamic, ensure that cooperation can spread over time and resist invasion from defecting individuals. Second, large

group cooperation is at the base of key phenomena of interest in these two fields. In organizational economics, there is a strong interest in understanding the root of persistent performance difference among seemingly similar enterprises (PPD among SSE) (Gibbons and Henderson, 2013). We believe that large group cooperation, understood through the lenses of "interaction structures" and replicator dynamics, can complement the advancements generated based on rational actor models of relational contracting. In strategic organization, understanding how organization affects capability formation is central (Argyres, 2011; Argyres et al, 2012). While collaboration is deemed crucial to that process, most attention has focused on the coordination of specialized effort (e.g., Puranam, 2018) and the integration of different knowledge pools (e.g., Grant, 1996) that underlies complementary assets (Argyres and Zenger, 2012). However, the notion of cooperation is almost entirely absent in the mainstream capabilities literature[78]. It may be the case that the theories of capabilities have assumed that cooperation is easily obtained within firms (for example, with incentives or other mechanisms) in order to focus on the details of knowledge integration (e.g., Grant, 1996), search (Rivkin, 2001) or learning (Argote and Miron-Spektor, 2011), much like organizational economics silenced capabilities in order to focus on cooperation incentives (see Gibbons, 2010 for a discussion). Within this discussion, we claim that large group cooperation is a fundamental and basic ingredient for group members to come together, to share ideas, knowledge and goals, and to coordinate in order to produce valuable goods and services that no individual could produce on his/her own. Our proposition is that 'interaction structures facilitate cooperation in big groups, which facilitate capability formation, which in turn generates PPD among SSE'.

We conclude by mentioning two managerial implications that we believe to be important. Calls to teamwork and collaboration are legion, present in almost any company that requires large groups to work together. Yet practitioners rely importantly on leaders with intuitive but inarticulate know-how on how to foster collaboration. Our study points to the crucial importance of the interaction structure of workers in generating cooperation, and thus, to the role that formal organization of

[78] For instance, a word search for "cooperation" in Eisenhardt and Martin (2000) yielded zero hits, while the word "knowledge" yielded 50 hits, a situation that replicates in other canonical papers in the capabilities literature.

companies, departments, and units can play in such a challenge. The second implication concerns change efforts that focus on critical mass approaches. When fostering cooperation, these approaches are limited as they don't address the social dilemma conditions of large-scale cooperation.

3.6. References

Abadie, A., Athey, S., Imbens, G.W. and Wooldridge, J., 2017. When should you adjust standard errors for clustering? (No. w24003). National Bureau of Economic Research.

Akerlof, George A. and Kranton, Rachel E. 2000 "Economics and Identity." Quarterly Journal of Economics, 115(3), pp. 715–53.

Akerlof, George A. and Kranton, Rachel E. 2005 "Identity and the Economics of Organizations." Journal of Economic Perspectives—Volume 19, Number 1—Winter 2005—Pages 9–32.

Akerlof, G., Kranton, R. 2010 Identity Economics: How our identities shape our work, wages and well-being. Princeton University Press

Archeti, M. 2009 The volunteer's dilemma and the optimal size of social group. Journal of theoretical biology, 261: 475-480.

Argote, L. and Miron-Spektor, E., 2011. Organizational learning: From experience to knowledge. Organization Science, 22(5), pp.1123-1137.

Argyres, N.S. and Zenger, T.R., 2012. Capabilities, transaction costs, and firm boundaries. Organization Science, 23(6), pp.1643-1657.

Axelrod, R. and Hamilton, W.D., 1981. The evolution of cooperation. science, 211(4489), pp.1390-1396.

Balliet, D., Van Lange, P. A. (2013). Trust, conflict, and cooperation: a meta-analysis. Psychological Bulletin, 139(5), 1090.

Barnard, C. 1938. *The functions of the executive*. Harvard University Press.

Bernhard, H., Fehr, E. and Fischbacher, U., 2006. Group affiliation and altruistic norm enforcement. American Economic Review, 96(2), pp.217-221.

Blader, Steven and Gartenberg, Claudine Madras and Prat, Andrea, The Contingent Effect of Management Practices (September 2, 2016). Columbia Business School Research Paper No. 15-48. Available at SSRN: https://ssrn.com/abstract=259425

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. Science, 328(5978), 617-620.

Boyd, R. and Richerson, P.J., 1988. The evolution of reciprocity in sizable groups. Journal of theoretical Biology, 132(3), pp.337-356.

Brahm, F., Lafortune, J., Tessada, J. 2018 "Is workplace safety everybody's business? Experimental evidence on prevention information, accidents and compensating wage differentials in SMEs", working paper.

Brahm, F. and Tarzijan, J., 2016. Relational Contracts and Collaboration in the Supply Chain: Impact of Expected Future Business Volume on the Make-or-Buy Decision. Journal of Supply Chain Management, 52(3), pp.48-67.

Brahm, F., Singer, M. 2015. Relación entre productividad y salud y seguridad laboral. Informe preparado para la ACHS.

Buchan, N.R., Johnson, E.J. and Croson, R.T., 2006. Let's get personal: An international examination of the influence of communication, culture and social distance on other regarding preferences. Journal of Economic Behavior & Organization, 60(3), pp.373-398.

Camerer, C. (2003). Behavioral game theory: Experiments in strategic interaction. Princeton University Press.

Cameron, A.C. and Miller, D.L., 2015. A practitioner's guide to cluster-robust inference. Journal of Human Resources, 50(2), pp.317-372.

Charness, G., Rigotti, L. and Rustichini, A., 2007. Individual behavior and group membership. American Economic Review, 97(4), pp.1340-1352.

Chassang, S. (2010). Building routines: Learning, cooperation, and the dynamics of incomplete relational contracts. American Economic Review, 100(1), 448-65

Clement, J. and Puranam, P., 2017. Searching for structure: Formal organization design as a guide to network evolution. Management Science.

Conti, G., 2005. Training, productivity and wages in Italy. Labour Economics, Volume 12, pp. 557-576.

Colombo, E. and Stanca, L., 2014. The impact of training on productivity: evidence from a panel of Italian firms. International Journal of Manpower, 35(8), pp.1140-1158.

Cyert, R.M. and March, J.G., 1963. A behavioral theory of the firm. Englewood Cliffs, NJ, 2, pp.169-187.

Dal Bó, P. and Fréchette, G.R., 2018. On the determinants of cooperation in infinitely repeated games: A survey. Journal of Economic Literature, 56(1), pp.60-114.

Dearden, L., Reed, H. & Van Reenen, J., 2006. The impact of training on productivity and wages: Evidence from British panel data. Oxford Bulletin of Economics and Statistics, 68(4), pp. 397-421

Diekman, A. 1985. The volunteer's game. Journal of Conflict Resolution, 29(4).

Eisenhardt, K.M. and Martin, J.A., 2000. Dynamic capabilities: what are they?. Strategic management journal, 21(10-11), pp.1105-1121.

Fehr, E., & Gachter, S. (2000). Cooperation and punishment in public goods experiments, American Economic Review, 90(4), 980-994.

Fernández-Muñiz, B., Montes, J. M., & Vázquez, C. J. 2009. Relation between occupational safety management and firm performance. Safety Science, Vol 47, p. 980-991.

Gant, J., Ichniowski, C. and Shaw, K., 2002. Social capital and organizational change in high-involvement and traditional work organizations. Journal of Economics & Management Strategy, 11(2), pp.289-328.

Gibbons, R., 2006. What the folk theorem doesn't tell us. Industrial and Corporate Change, 15(2), pp.381-386.

Gibbons, R. 2018. Foundations of Organizational Economics, Princeton University Press, forthcoming 2018.

Gibbons, R., Henderson, 2012. Relational contracts and organizational capabilities. Organization Science, 23(5), pp.1350-1364.

Gibbons, R. and Henderson, R., 2013. What do managers do?: Exploring persistent performance differences among seemingly similar enterprises. Chapter 17 of the Handbook of Organizational Economics, Eds. Robert Gibbons, John Roberts. Princeton University Press.

Gil, R. and Marion, J., 2012. Self-enforcing agreements and relational contracting: evidence from California highway procurement. The Journal of Law, Economics, & Organization, 29(2), pp.239-277.

Gittell, Jody Hoffer, Seidner, Rob, Wimbush, Julian (2010). "A Relational Model of How High-Performance Work Systems Work." Organization Science 21. 2: 490-506.

Goette, Lorenz; Huffman, David and Meier, Stephan. "The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups." American Economic Review 2006, (Papers and Proceedings) 96(2), pp. 212–16.

Graham, John R. and Grennan, Jillian and Harvey, Campbell R. and Rajgopal, Shivaram, Corporate Culture: Evidence from the Field (June 4, 2018). 27th Annual Conference on Financial Economics and Accounting Paper; Duke I&E Research Paper No. 2016-33; Columbia Business School Research Paper No. 16-49. Available at SSRN: https://ssrn.com/abstract=2805602 or http://dx.doi.org/10.2139/ssrn.2805602

Grant, R.M., 1996. Toward a knowledge-based theory of the firm. Strategic management journal, 17(S2), pp.109-122.

Grant, A. 2007. Relational Job Design and the Motivation to Make a Prosocial Difference. The Academy of Management Review, 32(2), pp. 393-417

Grennan, Jillian, 2014 A Corporate Culture Channel: How Increased Shareholder Governance Reduces Firm Value. SSRN working paper.

Guala, F., Mittone, L., Ploner, M. 2013. Group membership, team preferences, and expectations. Journal of Economic Behavior and Organization, 86: 183-190.

Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. Science, 312(5770), 108-111.

Halac, M. (2012). Relational contracts and the value of relationships. American Economic Review, 102(2), 750-79.

Hämäläinen, P., Takala, J. and Saarela, K.L., 2006. Global estimates of occupational accidents. Safety science, 44(2), pp.137-156.

Hauert, C., Michor, F., Nowak, M.A. and Doebeli, M., 2006. Synergy and discounting of cooperation in social dilemmas. Journal of theoretical biology, 239(2), pp.195-202.

Hermalin, B, 2013. Leadership and corporate culture. In the Handbook of Organizational Economics, Eds. R. Gibbons, J. Roberts. Princeton University Press.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies, Science societies. Science, 319(5868), 1362-1367

Ioannidis, J.P., 2005. Why most published research findings are false. PLoS medicine, 2(8), p.e124.

Knez, M., & Simester, D. (2001). Firm-wide incentives and mutual monitoring at Continental Airlines. Journal of Labor Economics, 19(4), 743-772

Kosfeld and Rustagi (2015), Leader punishment and cooperation in groups: experimental field evidence from commons management in Ethiopia. American Economic Review, 105(2): 747-783.

Kraft-Todd, G., Yoeli, E., Bhanot, S., & Rand, D. (2015). Promoting cooperation in the field. Current Opinion in Behavioral Sciences, 3, 96-101.

Kramer, R.M., 1991. Intergroup Relations and Organizational Dilemmas-The role of categorization processes. Research in organizational behavior, 13, pp.191-228.

Latane, J., Darley, J. 1968 Bystander intervention in emergencies: diffusion of responsibility. Journal of Personality and Social Psychology, 8(4): 377-383.

Lindenberg, S., & Foss, N. J. (2011). Managing joint production motivation: The role of goal framing and governance mechanisms. Academy of Management Review, 36(3), 500-525.

Lee, S. and Puranam, P., 2017. Incentive redesign and collaboration in organizations: Evidence from a natural experiment. Strategic Management Journal, 38(12), pp.2333-2352.

Loch, C.H. and Wu, Y., 2008. Social preferences and supply chain performance: An experimental study. Management Science, 54(11), pp.1835-1849.

Mathieu, J., Maynard, M.T., Rapp, T. and Gilson, L., 2008. Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future. Journal of management, 34(3), pp.410-476.

McEvily, B., Soda, G. and Tortoriello, M., 2014. More formally: Rediscovering the missing link between formal organization and informal social structure. The Academy of Management Annals, 8(1), pp.299-345.

McKenzie, D., 2012. Beyond baseline and follow-up: The case for more T in experiments. Journal of development Economics, 99(2), pp.210-221.

Milgrom, P. and Roberts, J., 1995. Complementarities and fit strategy, structure, and organizational change in manufacturing. Journal of accounting and economics, 19(2-3), pp.179-208.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. Science,314(5805), 1560-1563.

Nowak and Rand (2013). Human Cooperation. Trends in cognitive sciences, 17(8), pp.413-425.

Nowak, M.A. and May, R.M., 1992. Evolutionary games and spatial chaos. Nature, 359(6398), p.826.

Nowak, M.A. and Sigmund, K., 1998. Evolution of indirect reciprocity by image scoring. Nature, 393(6685), p.573.

Nowak, M.A. and Sigmund, K., 2005. Evolution of indirect reciprocity. Nature, 437(7063), p.1291.

Ohtsuki, H., Hauert, C., Lieberman, E. and Nowak, M.A., 2006. A simple rule for the evolution of cooperation on graphs and social networks. Nature, 441(7092), p.502.

Organ, D.W., Podsakoff, P.M. and MacKenzie, S.B., 2005. Organizational citizenship behavior: Its nature, antecedents, and consequences. Sage Publications.

Oster, E., 2017. Unobservable selection and coefficient stability: Theory and evidence. Journal of Business & Economic Statistics, pp.1-18.

Podsakoff, N.P., Whiting, S.W., Podsakoff, P.M. and Blume, B.D., 2009. Individual-and organizational-level consequences of organizational citizenship behaviors: A meta-analysis. Journal of applied Psychology, 94(1), p.122

Puranam, P, 2018. The Microstructure of Organizations. Oxford University Press, Oxford.

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. A. (2009). Positive interactions promote public cooperation. Science, 325(5945), 1272-1275.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. Trends in cognitive sciences, 17(8), 413-425.

Rivkin, J.W., 2000. Imitation of complex strategies. Management science, 46(6), pp.824-844.

Roberts, G., 2008. Evolution of direct and indirect reciprocity. Proceedings of the Royal Society of London B: Biological Sciences, 275(1631), pp.173-179.

Schein, E. 2010. Organizational culture and leadership. John Wiley & Sons; 4th edition

Tajfel, Henri 1970 Experiment in intergroup discrimination. Scientific American 223, 96-102.

Tajfel, Henri. 1982 "Social Psychology of Intergroup Relations." Annual Review of Psychology, 33, pp. 1–39.

Takezawa, M., & Price, M. E. (2010). Revisiting "The Evolution of Reciprocity in Sizable Groups": Continuous reciprocity in the repeated n-person prisoner's dilemma. Journal of theoretical biology, 264(2), 188-196

Traulsen, A., Nowak, M. 2006. Evolution of cooperation by multilevel selection. Proceeding of the National Academy of Sciences, 103, 10952-10955.

van Veelen, M., Garcia, J., Rand, D., Nowak, M. 2012. Direct reciprocity in structured populations PNAS, 109 (25) 9929-9934

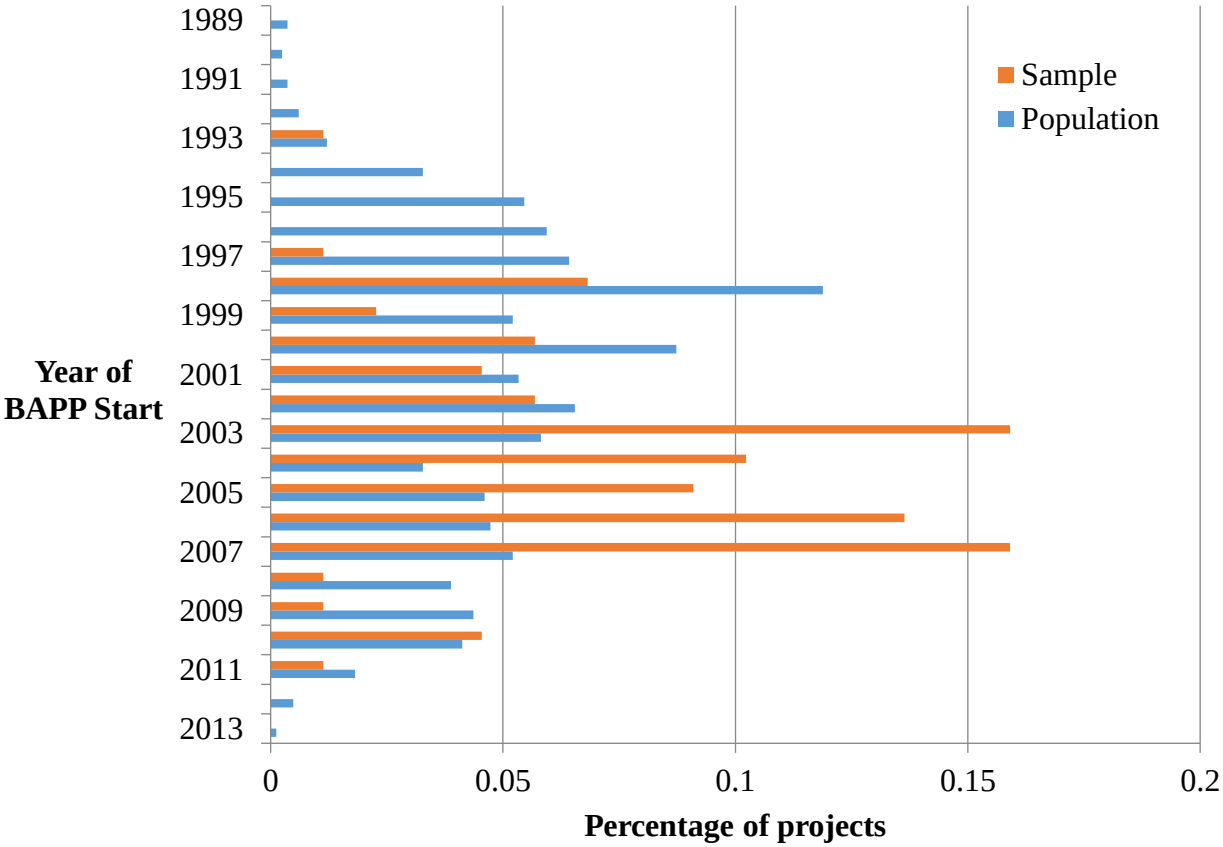
Vanneste, B.S., Puranam, P. and Kretschmer, T., 2014. Trust over time in exchange relationships: Meta-analysis and theory. Strategic Management Journal, 35(12), pp.1891-1902.

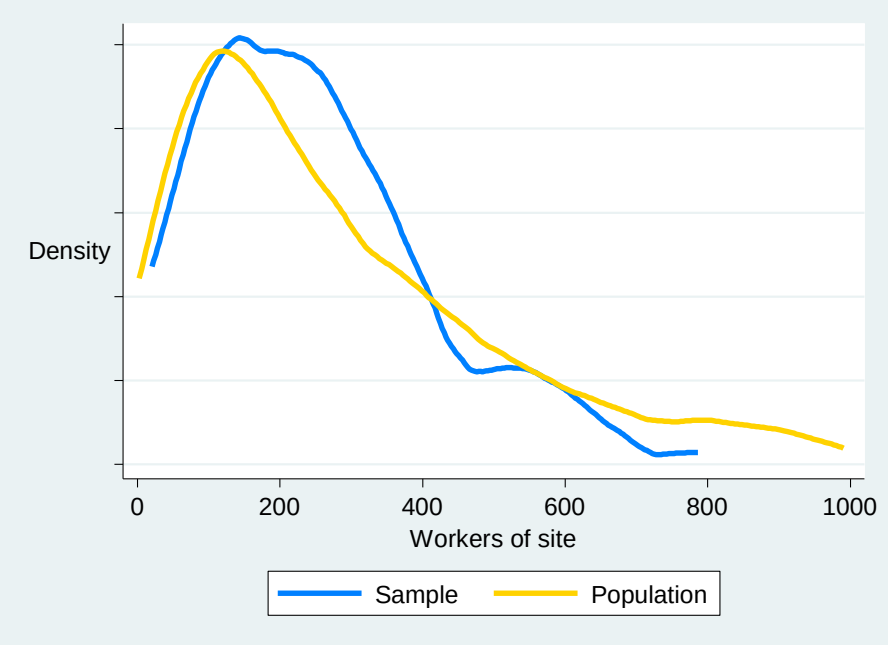
Viscusi, W.K. and Aldy, J.E., 2003. The value of a statistical life: a critical review of market estimates throughout the world. Journal of risk and uncertainty, 27(1), pp.5-76.

Yamagishi, T., Mifune, N. 2008 Does shared group membership promote altruism? Rationality and Society, 20 (2008), pp. 5-30

3.7. Appendices

3.7.1. Additional descriptive statistics of the DEKRA administrative data


Year of BAPP Start
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
Sample
Population
0
0.05
0.1
0.15
0.2
Percentage of projects



3.7.2. Heterogeneity of impact of BAPP

In order to explore the conditions that would maximize the impact of BAPP, we computed several variables at the site-month level that capture different choices in terms of how to structure/implement the BAPP methodology. Then, we studied the heterogeneity of impact using the following model:

$$ACC_{it} = b_1 + b_2 \times BAPP_{it} + b_3 \times TREND_{it} + b_4 \times (BAPP_{it} \times TREND_{it}) + b_5 \times BAPP_{it} \times Z_{it}$$

$$+ b_6 \times \ln(WORKERS_{it}) + U_i + ERROR_{it} \quad (10)$$

This model is the same as equation (1) with the addition of the interaction between BAPP and the variable Z, the moderator of the impact of BAPP. Given that Z captures a BAPP implementation variable it enter multiplies by the BAPP dummy. We also added to Z important variables of the implementation to act as control, such as participation and intensity. Below we describe the variables we used and the results we obtained (all the tables and figures for these results are available upon request).

i. *Tenure*. We computed the tenure of observers in their role of observer. This is computed as the time elapsed between the first observation and the focal observation. Then we

averaged these at the site-month level. We found evidence of a negative correlation of average observer tenure in the site with the reduction in accidents coming from BAPP. That is, long tenure in observers is detrimental to the impact of BAPP. However, just like participation, the average effect covered interesting underlying dynamics. We split tenure at the site on tenure for observers of different cohorts, and we found that tenure is detrimental for the first cohorts only, while it is beneficial on new additional observers. From our conversations with DEKRA, we learned that the observers that start early in the process, particularly those in the committee, tend to get demotivated of being the ones that contribute to large part of the cooperative effort in the site. They feel that other should contribute as well. On the contrary, for the newer cohorts, low tenure is simply a matter of low compromise towards cooperation.

ii. *Coaching*. On average, 20% of the observations are executed with a fellow observer that acts as a coach, who provides feedback to the observer at the end. A larger share of the coaching is executed by the enabler of the site. We find that coaching generates an increase in the reduction of accidents generated by BAPP. Again this is not surprising: quality of observations should improve with coaching, and effort/intensity might be boosted in response to a helping/caring leader.

iii. *Training method*. To train new observers, BAPP considers two methods. The typical method is that the enabler and the members of the starting team of observers are "trained to train" the new observers. This method is used 80% of the time. In the remaining cases, DEKRA is training all the observers that enter a specific implementation. We find that "training the trainers" is related to an improvement in the impact of BAPP. This suggest that it is useful to empower the starting team and to make them "own" their process by making them responsible for the successful training enrollment of new observers.

iv. *Use of the observation sheets*. There are two areas of the sheets that the observer can fill partially or fully depending on the care and effort they display. The first is general information at the header of the information sheet regarding the time, the place, the presence of coaching, the number of workers being observed (only occasionally is more than one), shift and hours at work of the observed worker, among a few others. The second area of the sheet he has to fill is the area known as "flags". When an observer has indicated that a particular behavior of the CBI was performed in a risky way, the observer, after providing feedback, has to fill out several fields describing the risky behavior as well as the quality/depth of the interaction with the observed worker. We computed the extent of use of these two areas of the sheet for each observation and then added this up at the site-month level. We found that these two area correlated with an increase in the impact of BAPP: The quality and care in filling out the information in sheets is generating a boost on BAPP. This is not surprising: this information is aggregated and analyzed by the enabler and the site committee in order to evaluate progress, plan the next steps and set priorities (e.g., are we performing enough coaching? are the observations balanced across the areas of the site? is the reported quality of the interaction with observed workers good enough?).

v. *Focused observations*. BAPP also includes the notion of focused observation, where the observers are instructed by the enabler (in conjunction with the committee) to observe specific tasks, specific times of the day, or specific areas of the site. The share of observations that have this condition is approximately 5%. The need to focus the

observations comes from observing patterns in the gathered data that show that risky behavior is happening mostly in specific areas, times and/or tasks[79]. We find that focused observations is related to an increase in the impact of BAPP. This suggest that is beneficial to analyze and put to use the information that is gathered by the observations sheets.

vi. *Number of behaviors tracked*. The enabler and committee at the start of BAPP define a set of critical behavior for the site –the CBI. However, as the implementation elapses the data might point at behaviors that are missing that became obvious after observation. In general, this feedback is placed into the comments section of the observations sheet but it can also be given informally as well. When this happens, the enabler and the committee might decide to take action and update the CBI. We find that increases in the CBI is related to an improvement in the impact of BAPP. This suggest that the capacity of the enabler and committee to gather and consider the feedback coming informally or through the sheets is beneficial.

From this exercise, plus a concise and interesting picture emerges regarding the type of BAPP implementation that is related to a higher reduction in accidents. First, on the organizational aspect of BAPP, we find that is useful to: 1) keep the group of observers small –at 10% of the site or 25 workers– so to focus on intensity, 2) have high tenure and specialize observers, and 3) delegate training to workers and promote coaching so to boost the collaborative spirit in the group. Second, on the information aspect, we find that is useful to: 4) be thorough in gathering and coding the operational data, and 5) have a leader –the enabler and committee– that acts upon the patterns that can be extracted from the data.

3.7.3. Letter handed out to workers

Letter handed out under treatment 1

Estimado Colaborador,

En nuestra tienda estamos implementando la metodología BAPP cuyo propósito es ayudarnos a trabajar de forma segura, sin accidentes y enfermedades laborales.

[79] It may also come from the occurrence of accidents. However, accidents are not frequent, and thus are a noisier source than leading/behavioral indicators such as risky behavior.

En esta metodología mi rol es ser tu "observador". Esto significa que de forma frecuente, por ejemplo una vez al mes, observaré cómo ejecutas tu trabajo, tomaré nota de lo observado y te entregaré retroalimentación. Si estás haciendo alguna tarea o actividad de forma insegura, intentaré hacértelo ver y podremos discutir cómo mejorar; si estás haciendo las tareas de forma segura, reforzaremos en conjunto la importancia mantener ese comportamiento en el futuro.

Todas las "observaciones" serán anónimas, tú nombre no quedará registrado en ninguna parte del proceso. Asimismo, yo seré tu único observador. Si algún otro observador se acerca por error a observarte, por favor indícale gentilmente que ya tienes un observador asignado.

Yo estaré haciendo observaciones a ti y a [NUMERO] otros trabajadores de la tienda.

Finalmente, es importante que sepas que TÚ también puedes ser un observador como yo. Si en el futuro decides serlo, yo te podré entrenar y podrás realizar observaciones a los mismos [NUMERO] trabajadores que yo observo. Podremos trabajar codo a codo, ayudando a nuestro compañeros a trabajar de forma segura!

Si tienes cualquier duda o comentario, no dudes en contactarme.

Cordialmente,

[FIRMA DEL MIEMBRO DEL EQUIPO IMPLEMENTADOR]

[NOMBRE DEL MIEMBRO DEL EQUIPO IMPLEMENTADOR]

Letter handed out under treatment 2 (the areas highlighted in grey are added to

Estimado Colaborador,

En nuestra tienda estamos implementando la metodología BAPP cuyo propósito es ayudarnos a trabajar de forma segura, sin accidentes y enfermedades laborales.

En esta metodología mi rol es ser tu "observador". Esto significa que de forma frecuente, por ejemplo una vez al mes, observaré cómo ejecutas tu trabajo, tomaré nota de lo observado y te entregaré retroalimentación. Si estás haciendo alguna tarea o actividad de forma insegura, intentaré hacértelo ver y podremos discutir cómo mejorar; si estás haciendo las tareas de forma segura, reforzaremos en conjunto la importancia mantener ese comportamiento en el futuro.

Todas las "observaciones" serán anónimas, tú nombre no quedará registrado en ninguna parte del proceso. Asimismo, yo seré tu único observador. Si algún otro observador se acerca por error a observarte, por favor indícale gentilmente que ya tienes un observador asignado.

Yo estaré haciendo observaciones a ti y a [NUMERO] otros trabajadores de la tienda. Más abajo encontrarás un listado con los trabajadores que forman parte este grupo. Hemos bautizado a este grupo con el nombre "[GRUPO NUMERO XX]".

Finalmente, es importante que sepas que TÚ también puedes ser un observador como yo. Si en el futuro decides serlo, yo te podré entrenar y podrás realizar observaciones a los mismos [NUMERO] trabajadores que yo observo (es decir, a los trabajadores del listado de abajo). Podremos trabajar codo a codo, ayudando a nuestro compañeros a trabajar de forma segura!

Si tienes cualquier duda o comentario, no dudes en contactarme.

Cordialmente,

[FIRMA DEL MIEMBRO DEL EQUIPO IMPLEMENTADOR]

[NOMBRE DEL MIEMBRO DEL EQUIPO IMPLEMENTADOR]

Observador asignado al "[GRUPO NUMERO XX]"

Integrantes del "[NOMBRE DEL GRUPO]"

	NOMBRE COMPLETO	CARGO

1	xxx	xxx
2	xxx	xxx
3	xxx	xxx
4	xxx	xxx
5	xxx	xxx
6	xxx	xxx
7	xxx	xxx
8	xxx	xxx
9	xxx	xxx
10	xxx	xxx
…	…	…
…	…	…
…	…	…

3.7.4. Report used in treatment 3

Report

Listado observadores y observaciones BAPP

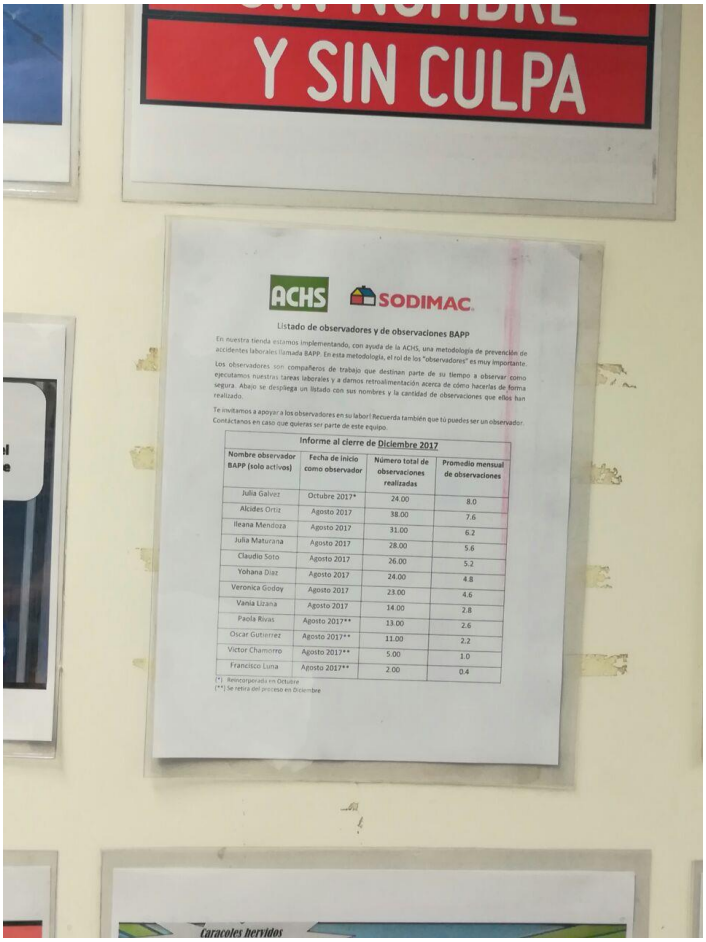
En nuestra tienda estamos implementando, con ayuda de la ACHS, una metodología de prevención de accidentes laborales llamada BAPP. En esta metodología, el rol de los "observadores" es muy importante.

Los observadores son compañeros de trabajo que destinan parte de su tiempo a observar como ejecutamos nuestras tareas laborales y a darnos retroalimentación acerca de cómo hacerlas de forma segura. Abajo se despliega un listado con sus nombres, y la cantidad y la calidad de las observaciones que ellos han realizado.

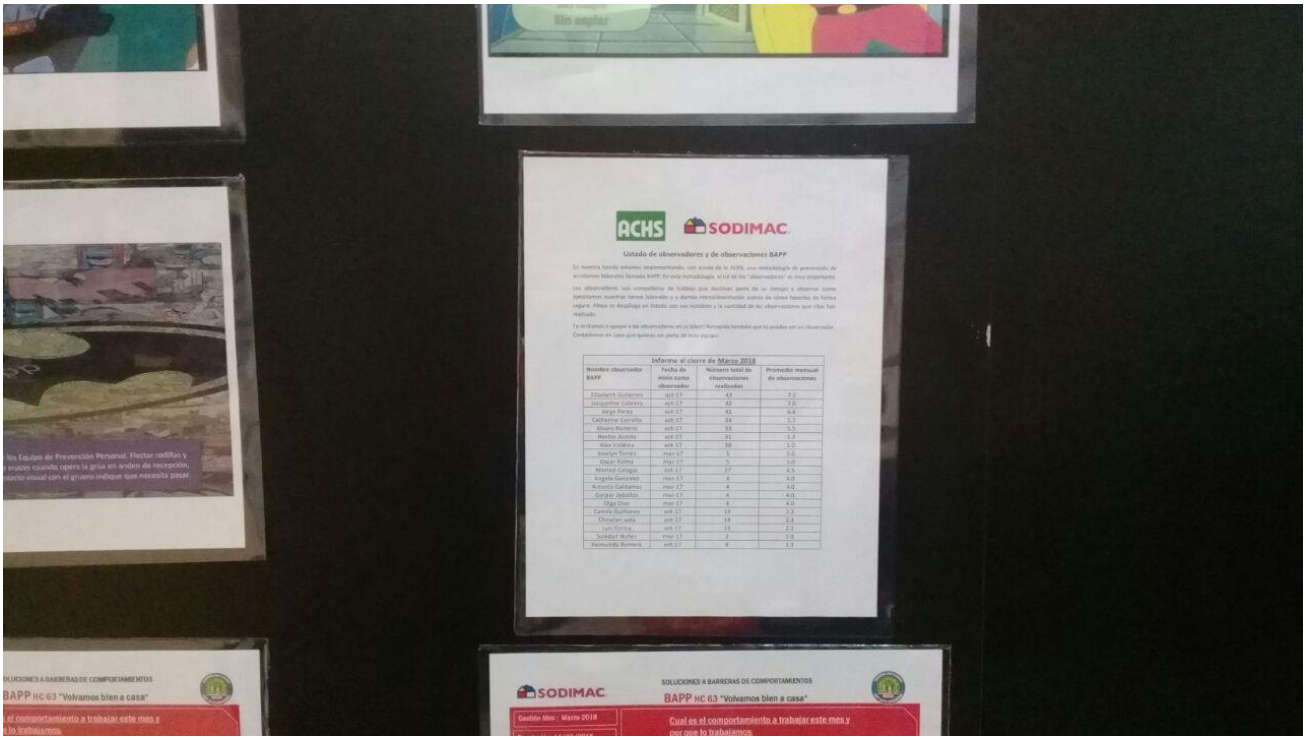
Te invitamos a apoyar a los observadores en su labor! Recuerda también que tú puedes ser un observador. Contáctanos en caso que quieras ser parte de este equipo.

Nombre observador BAPP	Fecha de inicio como observador	Número total de trabajadores observados	Promedio mensual de trabajadores observados
Prueba probando			
Prueba probó			
….			
….			
Etc.			

Picture of published report in “La Reina” store



Picture of published report in “Huechuraba” store



3.7.5. Identifying the treatment effects

Consider T1, T2, and T3 as treatments, and A, T, H, L as the differential impact that the treatments have above the control group within a site (where T2 and T3 really capture the incremental generated on T1). The treatments effect can be identified using the following system of equations:

$T1 = A$ (1) "Antofagasta"

$T1 + T2 = T$ (2) "Temuco"

$T1 + T3 = H$ (3) "Huechuraba"

$T1 + T2 + T3 = L$ (4) "La Reina"

Adding (1) and (2), we get the following system,

$2T1 + T2 = A + T$ (2')

$T1 + T3 = H$ (3)

$T1 + T2 + T3 = L$ (4)

The solution is:

$T1 = (A + H – L + T) / 2$; $T2 = L – H$; $T3 = (H – A + L – T) / 2$

If we add (1) to (3), we get,

$T1 + T2 = T$ (2)

$2T1 + T3 = H + A$ (3')

$T1 + T2 + T3 = L$ (4)

The solution is:

$T1 = (A + H – L + T) / 2$; $T2 = (L + T – A – H) / 2$; $T3 = L – T$

The average of the previous solutions is:

$T1 = (A – [L – (T + H)]) / 2$; $T2 = (3L – 3H + T - A) / 4$; $T3 = (3L – 3T + H – A) / 4$

From these solutions it is easy to see that the information from all sites is used to identify the treatment effects.

Other solutions are restrictive. For example, if we add (1) to (4), we get,

$T1 + T2 = T$ (2)

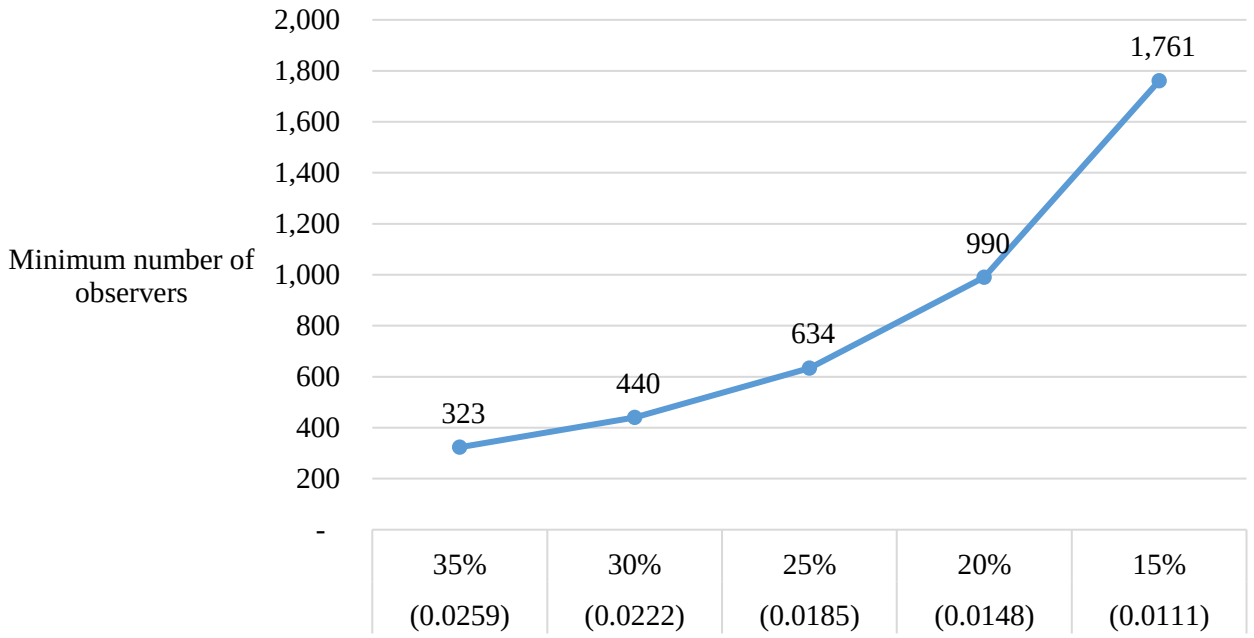
T1 + T3 = H (3)

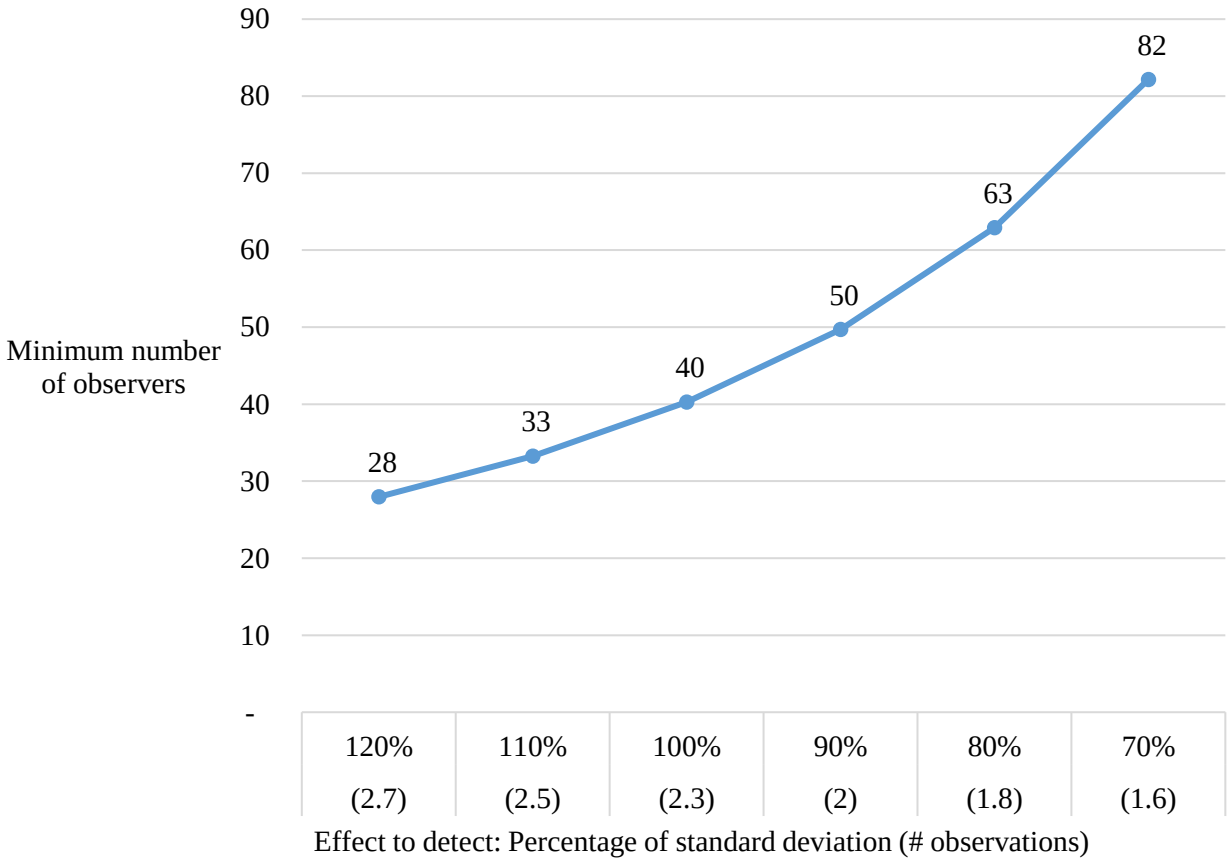
2T1 + T2 + T3 = L + A (4')

The solution is:

T1 = L + A - H ; T2 = H – L + T - A ; T3 = 2H – A – L , ***provided that H + L = A + L***

3.7.6. Power calculations





3.7.7. Impact of treatment on number of CBI items recorded

	CBI Items (1)	CBI Items (2)	CBI Items (3)
Treat 1	-1.51 (1.80)		
Treat 1 x Committee observer		-1.67 (2.69)	
Treat 1 x Committee observer alone			-1.28 (2.20)
Treat 1 x Committee observer with group			-2.01 (3.79)
Treat 1 x New observer		-1.34 (2.08)	-1.28 (2.12)
Treat 1 x Treat 2	-2.94 (2.67)	-2.96 (2.70)	-3.02 (2.75)
Treat 1 x Treat 3	-0.50 (3.02)	-0.41 (3.29)	-0.58 (3.30)
Enabler	1.23 (3.29)	1.19 (3.25)	1.20 (3.27)
Tenure	1.07** (0.48)	1.07** (0.48)	1.09** (0.46)

Tenure x New observer	-0.19 (0.54)	-0.19 (0.55)	-0.21 (0.54)
New	4.23 (2.96)	4.28 (2.96)	4.16 (3.49)
Sheets	5.86*** (0.31)	5.86*** (0.31)	5.86*** (0.31)
Store-month fixed effects	Yes	Yes	Yes
Observations	585	585	585
R-square	84.0%	84.0%	84.0%
Mean (mean per observation)	30.66 (6.11)	30.66 (6.11)	30.66 (6.11)
OLS. Errors in parentheses: Robust and clustered at the observer level. * p<0.1, ** p<0.05, *** p<0.01			

4. Concluding remarks

In this dissertation, I have applied ideas developed in evolutionary sciences to address two questions: what is the origin of firms? and, how can firms accomplish large scale cooperation? The answers that I propose are obtained using a mixture of formal models, regression analyses and field experiments: Firms evolved because they facilitate the conditions that lead to cumulative culture; firms can favor cooperation by exploring which interaction structures favor cooperators over defectors in their specific setting.

In the remaining of these remarks, I will briefly describe three (early stage) projects that, in a joint effort with collaborators, plan to take the cultural evolution agenda further.

The first project attempts to introduce cultural evolution to management. Several evolutionary approaches exist management (Nelson and Winter, 1982; Aldrich, 1999; Hannan and Freeman, 1997; Levinthal, 1997). In this project we detail how cultural evolution can provide novel and valuable contributions to management:

i) it provides an overarching framework that can integrate and organize the cacophony of (highly insular and disconnected) research branches in our field --in a similar way how evolution integrated the different biological sciences (e.g., paleontogy, genetics, ecology, developmental physiology, cytology) under a common framework,

ii) it allows to rigorously study the evolutionary origin of features widely present in economies, industries and companies (section 2 of this dissertation is an example),

iii) it introduces micro-evolutionary models, mainly based in evolutionary game theory and replicator dynamics, which provide a powerful tool to explain macro-phenomena using micro-founded mechanics --something that has eluded extant approaches; moreover, these models allows to engineer change in culture --something that is currently left to leaders or simplistic critical mass approaches (section 3 of this dissertation is an example).

The second project is the continuation of section 2. It uses verbal theory, a NK model and evidence to explain the evolution of the modern firms such as corporations and limited liabilities companies. The key idea is that modern firms are an excellent mechanism to accelerate the evolution of group-level traits, defined as complex traits generated by synergistic collaboration between specialized individuals (e.g., complex technologies, processes, practices, strategies). Since the imitation of complex group traits easily breaks down across firms (Rivkin, 2000), group selection, *and therefore group properties*, becomes crucial. We discuss how some key and historically non-obvious properties of modern firms affected the selection, variation and inheritance forces and estimate a twenty-fold increase the speed of the evolutionary process.

The third project is about distilling insights from the large literature on cultural micro-evolutionary models and then generate practical guidance for cultural change. These are insights on the likelihood that an inherited culture is adaptive, and in case it is not, insights on how to effective drive change using the mechanisms of social learning biases, conformity, interaction structures (for cooperation), norm psychology (for coordination), and complementarity (to avoid clashes with managerial/operational practices). The target audience is both academics and practitioners.

References

Aldrich, H. 1999. Organizations Evolving. Sage Publications, London.

Hannan, M.T., Freeman, J. 1977. "The Population Ecology of Organizations." American Journal of Sociology 82(5): 929-964.

Levinthal, D. 1997. "Adaptation on Rugged Landscapes." Management Science 43 (7): 934-950.

Nelson, R., Winter, S. 1982. An Evolutionary Theory of Economic Change. Belknap Press of Harvard University Press, Cambridge, Massachusetts

Rivkin, J. 2000. Imitation of complex strategies. Management Science, 46(6): 824-844