The Breakdown and Recovery of Cooperation in Large Groups: Exploring the Role of Formal Structure Using a Field Experiment

Francisco Brahm, Christoph Loch and Cristina Riquelme[1]

Abstract

Cooperation among employees is crucial for the success of organizations but can unravel with size. We study cooperation in the context of a workplace safety methodology: workers are voluntarily enrolled and are trained to provide advice to co-workers on safe behavior. Using archival data, we show that while cooperation is useful in reducing accidents, it breaks down as the number of enrolled workers increase. We show that this is due to decreasing marginal rewards for the additional enrolled worker. This supports recent research that highlights the crucial role of marginal benefits in shaping the relationship between cooperation and size. Then, we manipulate the safety methodology in a field experiment by randomly structuring workers in groups. This produces a recovery of cooperative effort and a reduction in risky behavior and accidents. We show that the likely mechanism behind this recovery are repeated interactions among advisors and workers, rather than group identity or social control such as peer pressure or reputation. This result suggests that the core function of structure is not only grouping workers to favor the division of specialized labor –as emphasized in prior research–, but also to promote cooperation at scale.

Keywords: Cooperation, Field experiment, Formal structure, Repeated interactions, Identity, Reputation, Workplace safety

1. Introduction

Achieving and sustaining cooperation – exerting effort for the benefit of the group and co-workers – is a crucial enabler of success in large organizations (Gibbons and Henderson, 2013; Organ et al., 2005). Cooperation is necessary to unlock the potential of the specialized and complementary assets and activities that comprise the firm (Milgrom and Roberts, 1995; Argyres and Zenger, 2012) and an essential condition for collective investments in valuable assets, such as the firm's reputation (Fehr, 2018). Research has documented a strong positive association between the cooperative behavior of employees and the performance of their organizations (Podsakoff et al., 2009), with recent causal evidence provided by Grennan (2014).

Several authors argue that a central role of the CEO is to foster cooperation in the organization (Barnard, 1938; Schein, 2010; Hermalin, 2013). However, large organizations often struggle to achieve cooperation. A survey of 1,348 CEOs of large US firms ranked cooperation among employees as the main driver of an effective culture, but

[1] Brahm: London Business School, 26 Sussex Pl, London NW1 4SA, UK, fbrahm@london.edu (corresponding author); Loch: Cambridge Judge Business School, Trumpington St, Cambridge CB2 1AG, UK, c.loch@jbs.cam.ac.uk; Riquelme: University of Maryland, Economics department, Tydings Hall, 3114 Preinkert Dr, College Park, MD 20742, USA, riquelme@econ.umd.edu. We are grateful for comments received by Bart Vanneste, Vincent Mak, Dmitry Sharapov and Jerker Denrell, Robert Gibbons and participants at seminars in the London Business School and Business Economics Department at Pompeu Fabra University and at the London50 Conference and the Berkeley Haas Culture Conference. The experiment was pre-registered on the AEA registry (AEARCTR-0002350). Usual disclaimers apply.

only 16% believe their culture is where it should be (Graham et al., 2018). One often cited culprit behind this problem is the organization's size: many authors have argued that it makes cooperation harder to sustain due to an increased free-riding temptation (Holmstrom, 1982; Alchian and Demsetz, 1972; Olson, 1965; Ostrom, 1990)[2].

At the same time, this decay of cooperation with size is a contested claim (Barcelo and Capraro, 2015; Pereda et al., 2019). A stream of lab research has documented that contributions in public good games do not decrease with the number of players, if anything, they tend to slightly increase (Zelmer, 2003; Isaac et al., 1994; Carpenter, 2007), a pattern that holds outside the laboratory for contributions to the Chinese Wikipedia (Zhang and Zhu, 2011) and free-riding in office candy bars (Haan and Kooreman, 2002). Other lab studies have found an inverted U-relationship between cooperation and group size (Capraro and Barcelo, 2015), which is reflected in the common pool resources literature where, in general, medium-size groups tend to cooperate more (Ostrom, 1990; Yang et al, 2013; Pereda et al., 2019).

Firms formalize their organization as they grow (Davila et al, 2010), usually adding a formal organizational structure (Colombo and Grilli, 2013). At its core, a formal structure entails separating workers into units or areas to favor the division of specialized labor (then these units get middle managers, reporting lines and other formal organization elements such as monitoring and incentive systems) (Puranam, 2018; Garicano and Wu, 2012). What is the relationship between structure cooperation at scale? Can structure increase cooperation even if it is imposed randomly, that is, without any attention to specialization? While some research suggests that infusing structure into groups can influence some aspects of the informal organization, such as the emergence of networks and coordination (McEvily et al., 2014; Clement and Puranam, 2018) or the presence of "real" authority (Aghion and Tirole, 1997), a role for structure in solving social dilemmas such as cooperation has not been considered in the organizational theory and organizational economics literatures[3]. If any, prior research seems to suggest that the separation into units creates challenges to inter-unit cooperation, requiring additional organizational elements, such as incentive systems, to reinforce it (Puranam, 2018; Gibbons and Roberts, 2013). This is surprising, as a sizeable literature in

[2] There are three main reason for this. Cooperation poses a social dilemma: while cooperation benefits the group, the temptation to free-ride by individuals usually increases with the size of the group (Holmstrom, 1982; Alchian and Demsetz, 1972; Olson, 1965; Ostrom, 1990). Second, given that the cooperation level in a group is a self-enforcing equilibrium, and thus stable and hard to change (Gibbons, 2006), a larger size hinders the coordinated change that is required to move out of a bad equilibrium. And third, cooperation becomes increasingly voluntary in larger groups, as it becomes harder to enforce using managerial levers such as monitoring or formal contracting (Gibbons and Henderson, 2012 and 2013; Organ, Podsakoff and Mackenzie, 2005).

[3] Of course many drivers of large scale cooperation have been studied in organizations. A partial list of these drivers is: the role of leaders as guides and enforcers (Barnard, 1938; Schein, 2010; Kosfeld and Rustagi, 2015; Hermalin, 2013); the identification of workers with the organization (Akerlof and Kranton, 2005); firm-wide financial incentives coupled with small groups (Knez and Simester, 2001); punishment either by individuals (Fehr and Gachter, 2000) or centralized institutions (Gurerk et al., 2006; Boyd et al., 2010); a set of organizational principles (Ostrom, 2000); and governance that focuses on the long term (Grennan, 2014).

evolutionary biology/anthropology strongly suggests that adding structure to populations can generate and sustain cooperation (Nowak, 2006 and 2010; Rand and Nowak, 2013; van Veelen et al., 2012; Allen et al., 2017).

In this article, we tackle these two related issues in a field study: first, we document that cooperation does decline with size and then, we experimentally show that adding structure (creating groups) is a good remedy. We use theory to develop hypotheses that guide us in probing mechanisms: we show that cooperation decays with scale due to decreasing marginal benefits, and we show that structure safeguards cooperation by increasing repeated interactions, rather than by group identity or social control (e.g., peer pressure, reputation).

We collaborate with a consulting company that implements a workplace safety methodology, a setting that is well suited to document a breakdown of cooperation with group size. This methodology works with employees of a site (e.g., a plant or a store) who volunteer to be trained and then provide safety feedback to their colleagues. This entails cooperation: training and feedback provision is costly to the volunteers and benefits mostly their colleagues (in the form of fewer accidents). During implementation of the methodology, an initial group of about ten volunteers typically expands to several dozen – a unique "field laboratory" to study cooperation as it scales.

In the first part of the paper, we hypothesize and document the breakdown of cooperation using archival data on 88 implementations with roughly 1.3 million feedback data points. On the one hand, the method promotes cooperation: volunteers expand within the site, exert effort and reduce accidents. On the other hand, its impact suffers significantly as the number of volunteers expands, especially beyond fifteen to twenty volunteers. We hypothesize and document that this is because as volunteers join over time, they provide increasingly lower levels of feedback and are quicker to drop out, so cooperation weakens. We show that this behavior is likely due to decreasing career and status benefits flowing from cooperation, such that the first cooperators are likely to enjoy better promotion prospects or reputational gains from colleagues. This dynamic may well be prevalent in many other social dilemmas in firms where organizational size leads to a dilution of cooperation benefits; this contrasts to, for example, an open source context (such as Wikipedia), where cooperation does not suffer with the size of the community because the contributors' warm glow increases correspondingly (Zhang and Zhu, 2011) as well as lead the programmers' status benefits (Raymond, 1998). Our result supports research suggesting that changes in the marginal benefits or costs of cooperation are crucial for understanding how scale impacts cooperation (Pereda et al., 2019; Hauert et al, 2006).

In the second part of the paper, we hypothesize how formal structure can avoid the decay of cooperation with scale and then introduce a pre-registered field experiment where we add structure to the safety methodology. While in a regular implementation feedback is provided quasi-randomly (i.e., any volunteer can provide feedback to any worker), we experimentally create groups that structure who is providing feedback to whom. In the treatment, half of the site's volunteers were assigned to groups of workers to whom they exclusively provided feedback; in the control, implementation happened as usual with the remaining half of volunteers and workers. By creating this structure of smaller units within the methodology, we reduced the number of persons that interacted with one another. As a consequence, a crucial mechanism that favored cooperation was turned on: the likelihood of *repeated interactions* between volunteers and employees receiving feedback increased by a factor of five, enabling the use of conditional strategies that produce self-enforcing cooperation[4] (Dal Bo and Frechette, 2018; Gibbons and Henderson, 2012; Axelrod and Hamilton, 1981; Rand and Nowak, 2013; Nowak, 2006).

Interacting with the baseline treatment of group structure, we add two treatments to probe two additional mechanisms that might be triggered by structure. First, the creation of small groups can facilitate *group identity* (Akerlof and Kranton, 2010). Extensive research shows that minimal group identity cues, together with a brief joint history, can foster cooperation among group members (Tajfel, 1982; Bernhard et al., 2006; Goette et al., 2006; Loch and Wu, 2008). Thus, in some sites we named the groups and revealed, within the groups, the identity of their members. If cooperation increases in the groups of these sites, then it is likely that identity more than repeated interactions is the driving mechanism of the impact of structure. Second, small groups can tap more easily into *social control* such as the withdrawal of cooperation if too many players defect (Boyd and Richerson, 1988; Rayo, 2007)[5], the application of peer pressure or punishment on defecting members (Kandel and Lazear, 1992; Bandiera et al., 2005; Carpenter, 2007; Fehr and Gächter, 2000) or cooperation conditional on reputational standing (Nowak and Sigmund, 1998 and 2005; Suzuki and Akiyama, 2005 and 2007). Given that observability is what enables these various ways of social control, in some sites we posted public lists displaying the amount of feedback provision performed by volunteers, ranked in decreasing order. Again, if the baseline treatment is impactful only in these sites, then social control, more than dyadic interaction, is the key driver of the impact of structure.

We find that the baseline group treatment was highly effective: it increased volunteering and the amount of feedback provision, and it reduced the incidence of risky behavior and accidents. Regarding the remaining treatments, we found that the identity treatment reversed the impact of baseline treatment while the observability

[4] When interactions are repeated, the player in a social dilemma can condition its behavior on the past behavior of the other player(s). There are many strategies that condition behavior (e.g., tit-for-tat, grim, generous tit-for-tat, win-stay-lose-shift), and all share the notion of reciprocating the other player's move: cooperate but punish defection by withdrawing cooperation. In organizational economics, this is associated with the idea of "relational contracting" (see Gibbons and Henderson, 2013). In evolutionary studies, it is associated with the idea of "direct reciprocity" (Nowak, 2006; Rand and Nowak, 2013) or "reciprocal altruism" (Boyd and Richerson, 1988).

[5] This is different than dyadic repeated interaction between worker and volunteer of the baseline. Here, the social control is between the observers regarding how much effort they exert.

treatment was neutral[6]. These two results plus the execution of several additional tests (which we explain in the body of the paper), strongly suggest that repeated interaction is the main mechanism through which structure safeguards cooperation in our setting, not the facilitation of identity or social control. Overall, our results suggest that a fundamental role of formal organizational structure is promoting cooperation within smaller units where repeated interactions increase. This result provides a novel explanation for the nature and function of organizational structure, complementing an existing focus on how structure influences non-social dilemmas, such as social network formation and coordination (Gibbons and Roberts, 2013; Puranam, 2018; McEvily et al., 2014; Clement and Puranam, 2018).

Our result on the role of structure is intriguingly consistent with recent theoretical results using evolutionary game theory. First, van Veelen et al. (2012) show that, in large populations, repeated interactions can favor cooperation, but that it is very unstable and infrequent as compared to defection. However, by adding a bit of population structure, repeated interactions can successfully stabilize higher levels of cooperation. They suggest that structure is crucial to deliver the type of stable cooperation seen in humans. Second, Allen et al. (2017) solve cooperation games in any type of population structure in order to find which type of structure favors cooperation the most. They find that, for any given population structure, cooperation gets maximally boosted if strong pairwise interactions are infused into the structure.

The rest of the study is organized as follows. Section 2 provides a detailed description of our safety methodology setting. Section 3 discusses theory and lays out two hypotheses that we study empirically in the two subsequent sections. Section 4 provides evidence of cooperation breakdown with size and its causes using archival data of a large sample of previous implementations. Section 5 introduces and analyzes our field experiment where we show how and why formal structure recovers cooperation levels. Section 6 concludes.

2. Setting: BAPP Methodology

We collaborated with DEKRA Insight, a global company specialized in workplace safety prevention services. One of its services is BAPP (Behavioral Accident Prevention Process), a methodology that improves workplace safety among the employees of a treated site (such as a plant, a store or a warehouse typically employing 250 employees) based on co-worker feedback. The BAPP methodology works as follows. After two months of assessment and

[6] Several tests confirmed a plausible explanation for the negative impact of treatment 2: this treatment lifted anonymity, triggering suspicion and distaste for surveillance and blame. The treatment clashed with the motto of the methodology (“no spying, no naming, no blaming”) and its voluntary character, which overwhelmed any group identity that might have been created. This result raises an interesting novel angle for cooperation research: when the benefit entails pointing at erroneous behaviors, anonymity might be necessary.

planning, a team of 8 to 12 employees (depending on the site's size) is constituted[7]. The selection of employees does not follow pre-defined criteria, other than focusing on front-line workers (supervisors or managers are not eligible) and being voluntary. One team member is consensually selected to the role of BAPP enabler, who is 100% devoted to the project. The enabler reports directly to the site manager, who serves as the project sponsor. Over the course of BAPP, the enabler and the team meet once a month in order to monitor and manage progress. In the fourth month, in order to become 'observers', the workers receive training on how to execute 'observations'. An observation consists of approaching a worker and, with his/her consent, observing his/her behavior for 10 to 20 minutes and filling out a detailed observation sheet. This sheet contains general information (e.g., date, place of the site, time of day) and a list of site-specific critical behaviors (e.g., driving a forklift, working at height), which are marked as performed either in a safe or a risky manner. If a risky behavior is identified, verbal feedback is provided to the worker. The sheet has space to provide written details about the behavior and the interaction with the worker. Only front-line workers are observed. BAPP is a method "by the workers for the workers". BAPP doesn't establish any pre-defined criteria regarding who observes whom, and the observed workers remain anonymous: their identities are never recorded. This is made clear to workers in advance as a critical feature of BAPP. Observers do not "spy", they ask for permission. BAPP has a frequently repeated mantra: "no spying, no naming, no blaming". In the fifth month, the initial observers are trained to enroll and train workers that are willing to become observers themselves. From the sixth month onwards, the enabler and observers have the goal of expanding the number of new observers; again, selection is voluntary and limited to front-line workers. The new observers do not participate in the monthly progress meetings. In addition to observations, observers also perform coaching, or observing a fellow observer execute an observation and then providing suggestions for improvement. Between the sixth and the twelfth month, the main challenge is ramping up observations and enrolling new observers. In the twelfth month the consultant performs a sustainability review and report, after which the site is left to its own devices.

This setting is well suited to studying large scale cooperation for two reasons. First, BAPP requires observers to devote time and effort in order to provide feedback to workers (and to provide coaching to fellow observers). This is textbook cooperation: private cost, and benefit to a third party. The cost is not small as BAPP observations are performed on top of regular work at the site. Observers who are part of the initial team, DEKRA estimates, devote approximately 7% of their time to BAPP during the first year, and later about 5%. Later observers spend a bit less, 3% to 5%. Furthermore, there is no pre-defined monetary compensation provided to observers. Sites attempt to provide flexibility to workers, but this is hard to achieve and regularly leads to role tensions. Informal rewards

[7] In the third month of implementation, the consultant and the team develop an inventory of critical behaviours in terms of safety (known as CBI, "critical behaviour inventory"). The behaviours are adapted to their site and the inventory typically includes between 15 and 30 behaviours (e.g., placing your body in front of the line of fire, not using the safety equipment, cluttered workspace).

(reputation or future promotions) typically exist but are by no means guaranteed, nor formally promoted by BAPP. The second reason for this being a good context to study cooperation is that as the number of observers grows, BAPP allows us to study in detail how cooperative effort is affected by scale – the dynamics of cooperation "in the wild". To understand cooperation in our setting, it is useful to define three terms using the following equation:

$$\textit{Contact rate} = \text{observations/worker} = \text{observations/observer x observers/worker} = \text{“effort” x “diffusion”} \quad (1)$$

"Contact rate" is the number of observations per worker at a site in a given month. The contact rate can be broken down into two components: "effort", which captures the number of observations per active observer (active indicates that the observer has done at least one observation in the month); and "diffusion", which captures the share of workers that are active observers. Effort captures the cooperative effort by an observer, and diffusion captures the expansion of cooperation in the site.

3. Theory and Hypotheses

We have mentioned earlier that the impact of scale on cooperation is inconclusive. In lab studies of the public good game, increasing the number of players can reduce or increase cooperation (Pereda et al, 2019), and this holds also in the field (Zhang and Zhu, 2011; Yang et al 2013). Recent research explores the conditions that might explain this (Pereda et al, 2019; Hauert et al., 2006). We illustrate the gist of the explanation using the classic public good game. In this game of n players, each player may cooperate by bearing a cost c to generate a benefit of b, of which everyone receives the share b/n. If everyone cooperates, each player receives $b/n \times n - c = b - c$. However, there is a temptation to free ride because payoffs are assumed to be such that the inequality $b/n \times (n - m - 1) > b/n \times (n - m) - c$ holds for any n or m, where m is the number of players that free-ride ($m \leq n$). This condition simplifies to $b/n < c$, which means that an individual free-rider loses less (shared) benefit than it saves in costs; therefore, the free riding temptation rises as n grows and the benefit is further diluted and cooperation suffers with scale (Holmstrom, 1982; Alchian and Demsetz, 1972; Olson, 1965). However, if b is a function of n with $b'(n)>0$, then free-riding can be overcome (Pereda et al, 2019; Hauert et al, 2006). This might be possible, for example, if players receive an additional subjective benefit, such as a warm glow, moral satisfaction, status, group identity or simply a "joy to give", that increases with others' cooperation (Andreoni, 2007; Zhang and Zhu, 2011). Also, the functional relationship $b(n)$ might be more complex; for example the marginal benefit might increase sharply with n at first, and then decrease after a certain threshold of cooperators. This would yield an inverted U relationship between cooperation and size (Capraro and Barcelo, 2015). The non-linearity of $b(n)$ may reflect the nature of activity. For example, in volunteer firefighting the marginal benefit is larger for the first group of volunteers, shrinking as firefighting capacity becomes sufficient. Also, reputational benefits may shrink, as the initial firefighting volunteers become heroes to fellow town members, while latecomers join something “routine”.

In BAPP, the free riding temptation of workers is the inequality $b/w \times (w\text{-}m\text{-}1) > b/w \times (w\text{-}m) - c$, where b is the average number of observations performed by each observer (and the benefits they generate), w is the number of workers in the site, and $(w\text{-}m)$ is the number of observers. The term $b/w \times (w\text{-}m)$ equals the contact rate, the number of observations each worker receives (this includes observers, as they observe each other). Free riding here means two things: either not becoming an observer (in which case $b=0$) or conditional on being an observer ($b>0$), how many observations are performed. The inequality simplifies to $b/w < c$, and thus, differently to the public good game, the free-riding temptation in BAPP increases with the size of the site. While we do not formalize a hypothesis, in our empirical analysis we will test this intermediate free-riding condition.

In order for the number of observers to affect the free riding temptation, it is required that b is a function of $(w\text{-}m)$, not a fixed parameter. We suggest that there are two reasons to expect this in our setting. First, the marginal impact of observations might be a decreasing function of $(w\text{-}m)$ such that $b'(w\text{-}m) < 0$. If a worker has been observed already a few times, it is likely that the marginal observation might generate less impact. And the higher the number of observers, the more likely that the worker have already been observed. Second, b can also be a decreasing function of $(w\text{-}m)$ if there exist some indirect career benefits of being an observer, such as higher likelihood of promotions or enhanced status/reputation among managers and colleagues, which decrease with the number of observers. Although these rewards are not formally included in any BAPP implementation, they might exist at the host organization and they are particularly salient for the first observers and then decay as the number of observers expands. Why? It is quite typical that signals of good citizenship tend to be valued and rewarded in firms (Podsakoff et al., 2009), and cooperating at the start represents a more credible signal of effort and engagement for managers and co-workers because BAPP is risky at the outset – not all implementations succeed. Given this context, we introduce our first hypothesis:

Hypothesis 1: Cooperation in BAPP, and therefore the marginal impact of BAPP on accidents, will be reduced as the number of observers increase.

Notice that while both mechanisms underlying this hypothesis – decreasing impact of additional observations and decreasing reputational/career benefits – have the same aggregate impact on cooperation, they differ in their details. The first mechanism decreases the marginal impact of observations for all observers, irrespective of their entry order. Therefore, when diffusion increases and more observers enter, all observers should display a similarly lower effort, with no heterogeneity by order of entry. In contrast, in the second mechanism, the differential career/reputational rewards that early observers may enjoy do not depend on the number of observers. Therefore, we should observe that there is heterogeneity across order of entry: later entrants should display increasingly lower levels of effort. The empirical analysis will distinguish between these two mechanisms.

We now discuss how adding structure to BAPP can remedy the hypothesized breakdown of cooperation. We start with the premise that adding formal structure entails, essentially, grouping workers into units or areas to favor the division of specialized labor (Puranam, 2018; Garicano and Wu, 2012). We claim that by breaking a large group into smaller sub-groups, even if they are set at random (i.e., specialization is not taken into account), structure favors three types of mechanisms that incentivize cooperation.

First, small groups increase the number of dyadic repeated interactions, which boost cooperation (Dal Bo and Frechette, 2018; Gibbons and Henderson, 2012; Axelrod and Hamilton, 1981; Rand and Nowak, 2013; Nowak, 2006). BAPP is well suited to test this idea. BAPP by itself introduces no division of labor among observers, all of them observing workers in a quasi-random way, so that the likelihood of repeating interactions is rather low. For example, suppose there are w workers, f observers execute j observations a month, and an observer selects a worker randomly each time. Then the likelihood that a worker repeats observations with a specific observer the next month is P(Repeat Interaction) = P(RI) = P(Being observed) x P(Same observer) = $j \times f/w \times 1/f = j/w$. In a typical implementation $j=5$ and $w=200$, so P(RI) = 2.5%. Compare this to a structure that restricts who observes whom: suppose the w workers are divided into g groups of w/g workers and f/g observers each, and observations remain random but with likelihood p the f/g observers observe outside of their group. Then, within a group, P(RI) = $j \times f/w \times [\,(1-p)/(f/g) + (p/g)/f] = j/w \times [(1-p) \times g + p/g]$. If $p = 0$, so that observers are fully bound to their group, then P(RI)=$j/w \times g$; that is, creating groups dramatically boost repeated interactions. If in the example above, there are $g=10$ groups, then P(RI) = 25%, a tenfold increase. (If some observations are carried out in other groups, namely $p>0$, this boost in repeated interaction is merely somewhat diluted.)

Second, group identity can foster cooperation (Akerlof and Kranton, 2005), particularly if groups are smaller (Wichardt, 2008). A long tradition in social psychology has used the minimal group paradigm of social to study identity (Tajfel, 1970). In this method, even a random assignment of experimental subjects to groups that are tagged using a trivial or arbitrary label leads to higher help for in-group members (Tajfel, 1982). However, in this tradition subjects do not face a social dilemma (Bernhard et al., 2006). Recent research suggests that the positive effect of the minimal group paradigm on help might not hold when individual and group welfare conflict (e.g., Buchan et al., 2006; Charness et al., 2007). Recent studies show evidence that groups require a joint history (Bernhard et al., 2006; Goette et al., 2006), even if this is minimal, like a short introduction (Loch and Wu, 2008), and common knowledge of group affiliation (Guala et al., 2013; Yamagishi and Mifune, 2008). We follow these ideas to design the "identity" treatment which we detail below.

Third, different types of group dynamics we label as "social control" facilitate cooperation, especially in small groups. Peer pressure and punishment targeted at free-riding individuals allow groups to enforce norms of effort and cooperation (Kandel and Lazear, 1992; Mas and Moretti, 2009; Bandiera et al., 2005; Fehr and Gächter, 2000). Research has convincingly shown that these means of enforcing behavior are more effective in smaller groups, because groups members have an easier time to coordinate around a norm (Bandiera et al., 2005) or because monitoring is facilitated (Carpenter, 2007). Also, untargeted punishment suffers with scale: the effectiveness of enforcing group cooperation by punishing the whole group by withdrawing one's cooperation, if a given percentage has defected, becomes ineffective in large groups (Boyd and Richerson, 1988). Finally, reputation may also be at play: when a person A (does not) help B, then C observes this and is therefore (not) willing to help A back (Nowak and Sigmund, 1998 and 2005; Rand and Nowak, 2013; Kraft-Todd et al., 2015; Khadjavi, 2016). This mechanism also works better in small groups, where reputation standings are easier to track (Suzuki and Akiyama, 2005 and 2007).

Notice that all of these distinct mechanics – peer pressure, withdrawal of one's cooperation, and reputation, which we lump into the label of "social control" – share the common *requirement of observability of effort to operate*. Without observability, it is difficult to: know whom to pressure (Mas and Moretti, 2009); to coordinate and enforce a norm (Bandiera et al, 2005); to monitor effort for potential targeted punishment (Carpenter, 2007; Fehr and Gachter, 2000) or effort withdrawal (Boyd and Richerson, 1988); and to track reputations (Nowak and Sigmund, 2005; Kraft-Todd et al., 2015).

Given this discussion, we introduce our second hypothesis:

> *Hypothesis 2*: Adding structure to BAPP in the form of small groups will counteract the reduction in cooperative effort and therefore restore BAPP's marginal impact on accidents.

We propose that the mechanism of repeated interactions is "primordial" in the sense that it is at the core of what structure does. Even if a structure is implemented at random, and nothing else is added so that structure is completely "blind" to agents, then a change in repeated interactions will still occur. This is not the case with the mechanisms of identity or social control. These mechanism are not "primordial", in the sense that they require the addition of something extra: in the case of identity, it requires the addition of a label, common knowledge of affiliation and joint history; in the case of social control, it requires the addition of observable information. Of course, it may well be that structure does not boost cooperation without an identity or social control effect, even if it fosters repeated interactions at the most basic level. In the end, parsing this out is an empirical question. As we discussed under the first hypothesis, the empirical analysis will distinguish the underlying mechanisms at play.

4. Breakdown of Cooperation: Evidence from Archival Data

4.1. Data and descriptive statistics

DEKRA provided an administrative data set of 1,352 sites with BAPP implementations, executed between 1989 and 2013. These projects cover a substantial percentage of their BAPP activity over the years. For each site and month, we have detailed information on implementation.[8] We have accidents information, which DEKRA took great care to harmonize it across countries, as there might be different rules in reporting accident data.

We restricted the sample to those projects that had information on workplace accidents at least two years before and three 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 online appendix A.1 we show that the sample is not significantly different from the population. In addition, we collected observation-level data for the 88 projects in our sample. The data set contains 1,265,176 observation sheets in total, each indicating site, date, name of observer, area of the site, among many other information of the observation sheet (e.g., number of items in the CBI that were observed/recorded, time of the day, etc.).

In **Figure 4-1**, we display some important descriptive statistics --the average and percentiles 25 and 75 for contact rate, effort and diffusion at the site level-- for the first 36 months of BAPP implementation (considering the 88 sites of our sample). Contact rate (the green line) approaches 1 by the end of year 3, but there is considerable variation across sites (dotted green lines). Effort (the red line) is stable over time, displaying a 10% decrease from ~5.3 in the first year to ~4.8 in the third year. Variation is also high (red dotted lines): the twenty-fifth percentile displays around 3 observations, while at the seventy-fifth percentile this increases to 6.5. Diffusion has a steady and uniform increase from 4% in the first couple of months to 21% in 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. These indicators suggest that as cooperation diffuses and the number of observers increase, the average cooperative effort goes slightly down. However, this decrease is not significant and therefore, is not

[8] Variables of the data: date, name of site, company of site, industry of company, country of site's location, name of consultant, presence of a culture survey, 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 amount of cases, training of new observers is done by DEKRA and not the observers of the starting team), number of critical behaviors that are tracked, the number of critical behaviors that were observed, the number of observed critical behaviors that were safely and riskily executed, number of workers on the site, and number of accidents.

consistent with hypothesis 1. Below we will show that this masks strong variance across observers with different order of entry.

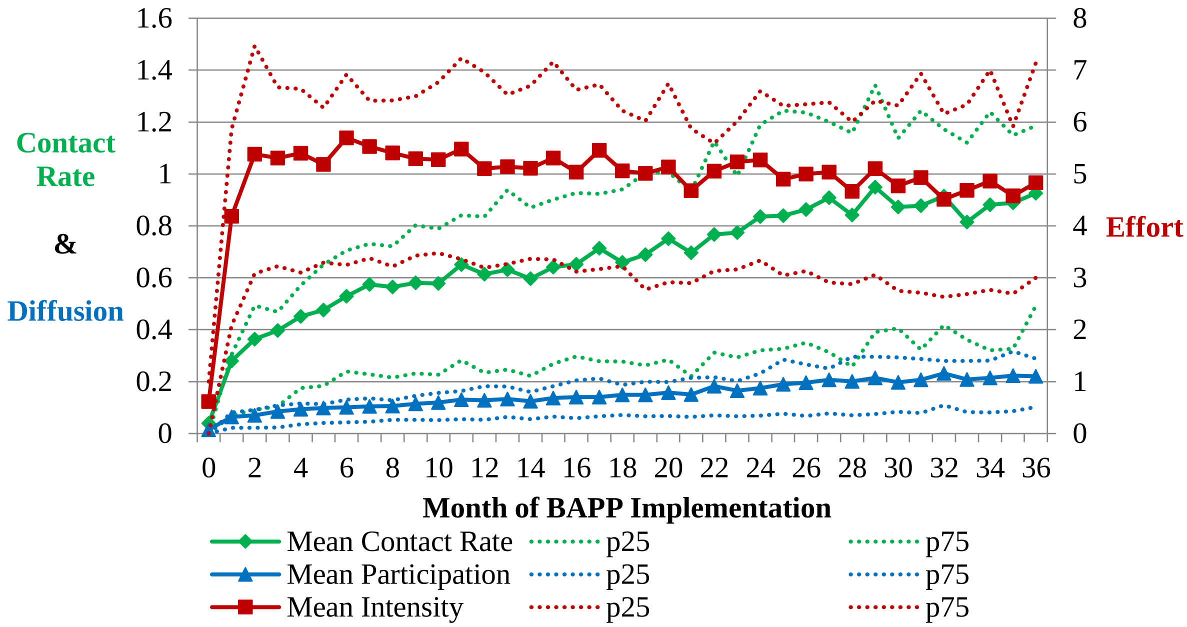


Figure 4-1. Evolution of contact rate, effort and diffusion over BAPP implementation

4.2. Checking the benefit of BAPP and the free-riding condition

Our theory requires that observations from BAPP actually generate a benefit to workers. We examine this assumption by studying the impact of BAPP on accidents with the following model:

$$ACCIDENTS_{it} = b_1 + b_2 \text{ x } BAPP_{it} + b_3 \text{ x } TREND_{it} + b_4 \text{ x } (BAPP_{it} \text{ x } TREND_{it}) + b_5 \text{ x } \ln(WORKERS_{it}) + U_i + ERROR_{it} \quad (2)$$

Equation (2) models the accidents at 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 at the site. TREND equals $(t - \theta_i)$, where t is the month and θ_i is the month when the BAPP started at the site. Given our sampling, this variable goes from -24 to +36. We add a site fixed effect U_i to the estimation in order to control for time-invariant store unobservables. As a control, we add the natural logarithm of workers, as more workers translate into more accidents.[9] 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.

[9] 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 slightly stronger.

Table 4-1. Impact of BAPP on accidents

	Accidents – OLS (1)	Accidents – OLS (2)	Accidents – OLS (3)	Accidents – POIS (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	4,762
Mean of dependent variable before BAPP	1.338	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<0.001 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 statistically significant at p<0.05.

In **Table 4-1** we display the results. Column (1) indicates that BAPP is significantly associated with a decrease in accidents. Column (2) shows that the TREND is negative and statistically significant. BAPP loses its statistical significance; this is due to collinearity but could also reflect that it is the trend that matters, not BAPP. Column (3) dispels this concern: the trend turns negative only after BAPP. The trend without BAPP is flat and non-significant. 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% (the variance inflation factor is above 6 for these variables). In model (4) we display POISSON fixed effect estimates as robustness (accidents tend to follow a count distribution). The results do not change. Using column (3), we find that BAPP is related to a decrease in the level of accidents of 0.2 accidents and, regarding the slope, with a decrease of 0.132 accidents after 12 months. At the end of the first year, BAPP is associated with an overall decrease of 30% in accidents.

These estimates are subject to endogeneity bias. The main threat to identification is posed by time-variant unobservables at the site level (e.g., a change in site manager). To tackle this issue, we execute three analyses: a placebo test, and we add a site-specific trend and probe the mechanisms (see the online appendix A.2). These analyses provide evidence for the documented impact of BAPP being causal. Overall, our analysis shows that BAPP does generate a large benefit on workers, as required in our theory.

We now test the first free-riding condition that predicts that cooperation is affected by the size of the site. In **Table A-8** of the online appendix we exploit within site changes in the number of workers to show if the site doubles in the number of workers, then diffusion is reduced by 17 percentage points, a substantial amount (however, we find no relationship with effort). This provides additional validity to our theoretical framework.

4.3. Cooperation is lower for newer observers

We use observation level data to show that cooperation effort decreases with the observers' order of entry. We broke down the number of active observers into five quintiles of entry order, that is, into five cohorts of observers. For all observers that have participated in BAPP, we record the "date of entry" as the date of their first observation and then compute an "order of entry" for each observer within their site. See Appendix A.5 for details of the cohorts. Then, for each quintile and each month, we compute the mean effort for each quintile. The results, which we summarize at year level, are displayed in **Figure 4-2**, where the dotted lines display a 95% confidence interval. We can see that effort experiences an important drop as we move up in the quintiles, and that the differences are statistically significant. In the first year, the first quintile executes 7 observations per month, while the fifth quintile only executes 3.5. Effort levels converge slightly over time. The first quintile executes 5.7 monthly observations in year 3, the fifth quintile executes 4.5. In addition, as detailed in the appendix, the descriptive analysis of the data also shows that the rotation of observers increases with the cohorts; that is, diffusion becomes more fragile if the number of observers increases.

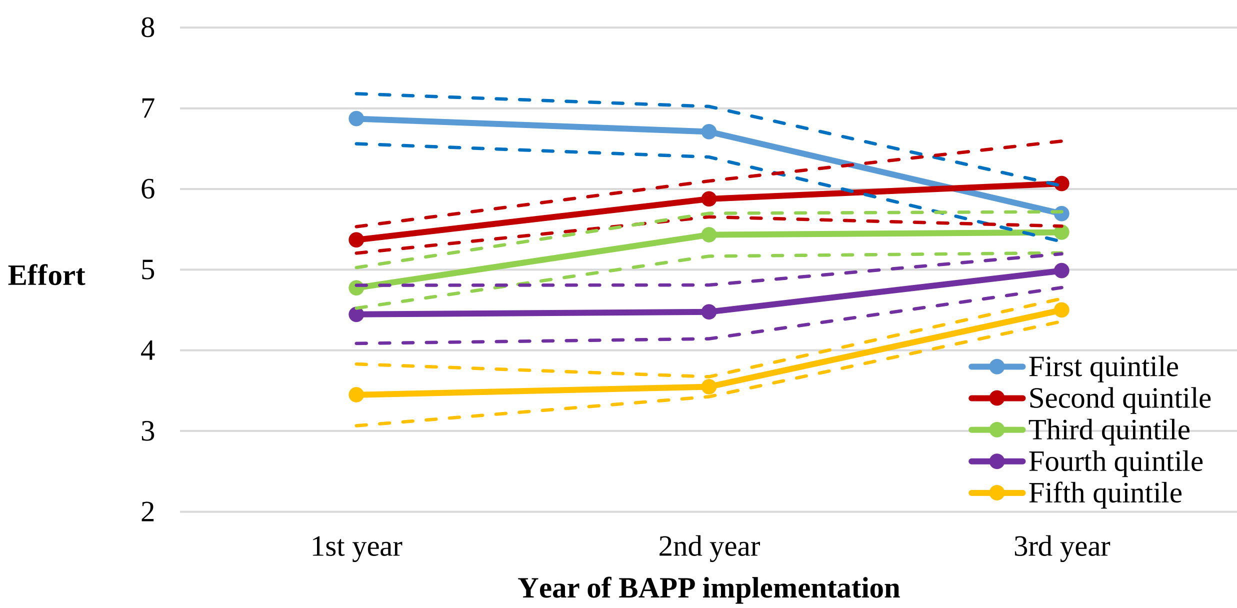


Figure 4-2. Newer observers execute fewer observations

However, this descriptive analysis is subject to site-specific confounding factors. For example, it could be that the lower effort of higher quintiles is due to a higher diffusion rate: in order to achieve a pre-defined contact rate, low effort might be needed if diffusion is high. To check this, in the online Appendix A.6 we regress the number of observations per observer per month on the entry cohorts (measured within the site), adding several controls, such as diffusion at the site level, and observer and month fixed effects. We repeat this regression using tenure as observer as the dependent variable. These analyses confirm that higher cohorts display lower effort and higher rotation. The values are similar in magnitude to those of **Figure 4-2**. Overall, the analysis of observation-level data provides strong support for Hypothesis 1: cooperation in BAPP, both in terms of effort and diffusion, suffers as the number of observers increases.

Figure 4-2 also allows to distinguish between the two mechanisms driving the breakdown in cooperation with size. The first mechanism is that observations becomes less impactful, reducing b in the free-riding temptation condition, as the cumulated number of observations increases (which is necessary consequence of adding more observers). This would predict that *all observers* would reduce their effort to a homogeneous but lower level after more observers enter. In short, there would be no heterogeneity by cohort, simply a reduction in the average. In **Figure 4-2** we observe quite the contrary: heterogeneity is large and convergence (statistically) inexistent in the second year and partial in the third year. Additional tests also show further evidence against this mechanism: i) Using an interaction model, in **Table A-9** of the online appendix we show that the impact of effort is independent of the degree of diffusion, ii) In **Table A-10** of the online appendix we show that the cumulative number of observations by observers doesn't affect the impact of BAPP, iii) Consistent with **Figure 4-1**, in the online appendix A.7 we confirm, using regression, that the *average* effort does not decrease with diffusion[10].

4.4. The impact of BAPP decreases as the number of observers expand

Hypothesis 1 indicated that this cooperation breakdown will be reflected in the impact of BAPP on accidents. To do so we examine the two cooperation elements of: effort and diffusion. If the increase in the number of observers affected cooperation we should observe that diffusion produces a reduction in the impact of BAPP.

This exercise exploit the facts that a high contact rate can be achieved using two generic strategies: high effort and low diffusion, or low effort and high diffusion. BAPP does not impose an execution strategy in this regard: sites decide, leading to naturally occurring variance across implementations. In **Figure 4-3** we display all the month-site combinations of diffusion and effort for the three years of BAPP implementation. In red we display a site that

[10] The fact that effort decreases with cohorts but that *average* cooperation effort doesn't decrease with diffusion seems puzzling. However, the reason is simple, and is a combination of convergence in effort (depicted in **Figure 4-2**) and change in composition (as depicted in **Figure A-4** of the online appendix A.5). Given that first (last) cohorts decreased (increased) their effort over time, and their relative weight decreased (increased), the average effort only decreases slightly but not sufficient to be statistically significant. Further, as we will show in the next section, this distinction between average effort and cohorts is consequential for the impact of BAPP: The cohort effect implies that there is a combination of diffusion and effort that maximizes the impact of BAPP.

achieved a high contact rate by growing effort while keeping diffusion low. In green we display a site that achieved a high contact rate by growing diffusion while keeping its effort low.

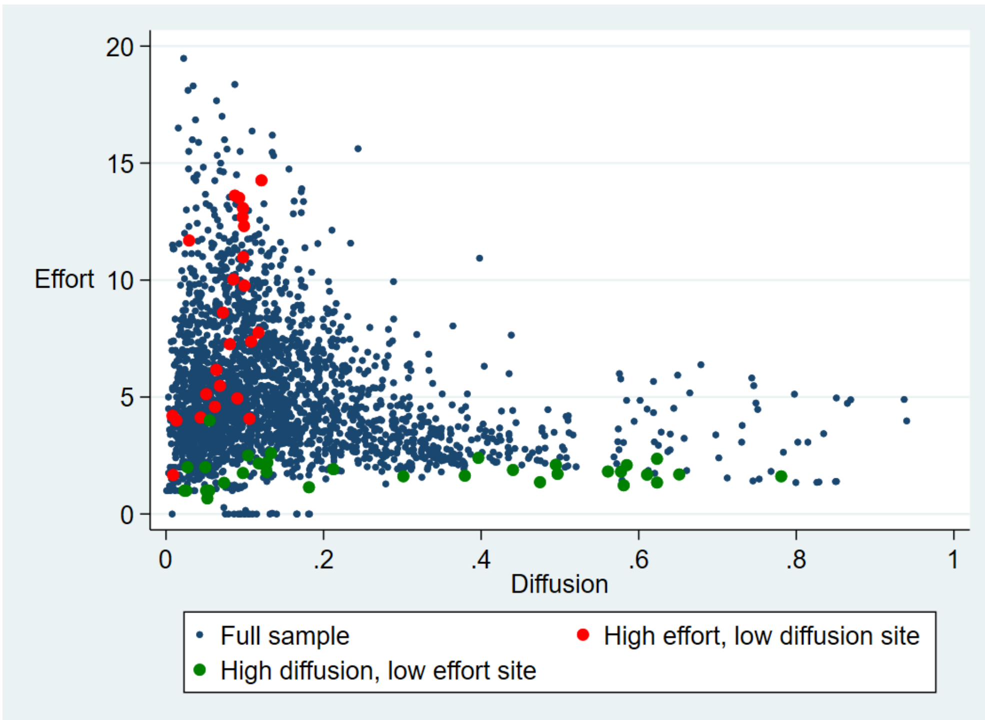


Figure 4-3. Two strategies to increase contact rate.

We exploit this variation in strategies to isolate the impact of effort and diffusion. We use the following model:

$$ACC_{it} = b_1 + b_2 \text{ x } BAPP_{it} + b_3 \text{ x } TREND_{it} + \sum_j b_{4j} \text{ x } BAPP_{it} \text{ x } QUINT\_EFFORT_{jt} + \sum_j b_{5j} \text{ x } BAPP_{it} \text{ x } QUINT\_DIFF_{jt} + b_6 \text{ x } \ln(WORKERS_{it}) + U_i + ERROR_{it} \quad (3)$$

This model introduces two sets of five quintiles of effort and diffusion. In **Table 4-2** we present the results. The results indicate that increases in effort unambiguously decrease accidents. In contrast, diffusion decreases accidents at first but then increases them. We use a joint t-test because of collinearity (if we use dummies of high/low diffusion and high/low effort, the results are statistically significant without joint t-test; see **Table A-10** in the online appendix). Adding the control of BAPP times TREND in column (2) does not change the results.

Table 4-2. The role of effort and diffusion in the impact of BAPP.

	Accidents (1)	Accidents (2)
BAPP	0.016 (0.149)	-0.039 (0.152)
BAPP X 1ST QUINTILE OF EFFORT	(omitted)	(omitted)

BAPP X 2ND QUINTILE OF EFFORT	-0.113 (0.089)	-0.118 (0.091)
BAPP X 3RD QUINTILE OF EFFORT	-0.144 (0.101)	-0.147 (0.103)
BAPP X 4TH QUINTILE OF EFFORT	-0.218* (0.126)	-0.226* (0.130)
BAPP X 5TH QUINTILE OF EFFORT	-0.267** (0.117)	-0.266** (0.119)
BAPP X 1ST QUINTILE OF DIFFUSION	(omitted)	(omitted)
BAPP X 2ND QUINTILE OF DIFFUSION	-0.169† (0.119)	-0.144† (0.113)
BAPP X 3RD QUINTILE OF DIFFUSION	-0.016 (0.110)	0.015 (0.116)
BAPP X 4TH QUINTILE OF DIFFUSION	0.037 (0.096)	0.084 (0.094)
BAPP X 5TH QUINTILE OF DIFFUSION	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 DIFFUSION and BAPP X 2ND QUINTILE OF DIFFUSION is rejected at 20% and 10% significance in column (1) and (2), respectively.		

In **Figure 4-4** and **Figure 4-5** we separate the impacts of effort and diffusion. Each figure keeps one dimension constant at its second quintile, and then displays the impact of changing quintiles in the second dimension. **Figure 4-4** shows the monotonically increasing impact of effort. That is, a higher cooperative effort by observers always pays off. **Figure 4-5** displays a clear inverted-U relationship between diffusion and accidents: conditional on effort, increasing diffusion is beneficial up to a diffusion of 8%, which for the average site of 245 employees, is equivalent to 20 observers. After this point, adding more observers is detrimental[11]. This result provides further evidence for Hypothesis 1: cooperation breaks down with size with an important effect on the impact of BAPP on accidents.

Note that we can trace the negative impact of diffusion back to the lower level of effort from higher cohorts of observers. Given that the average effort is not affected by diffusion (see footnote 10), the impact of heterogeneous effort due to order of entry would not be captured by the coefficients b_{4j} of Equation 3. If one uses the impact of effort plotted in **Figure 4-4** and the lower effort associated with the marginal observer plotted in **Figure 4-2**, then

[11] Consistent with the fact that contact rate = effort x diffusion, in the online appendix A.4 we show that the contact rate has an inverted-U relationship with the reduction in accidents, with a maximum reduction at a contact rate of 30%. The non-linearity of the impact of diffusion means that, if a constraint on contact rate and individual effort is assumed, there is a combination of effort and diffusion that maximizes the impact of BAPP.

one can explain a big chunk of the drop in the impact of higher diffusion depicted in **Figure 4-5**. An alternative explanation for the detrimental impact of diffusion could be a reduction in the quality of observations. In appendix A.10 we examine whether the order of entry of the observer affected the quality of the information on their observation sheets, as measured by the number of fields they should complete. We find no evidence that order of entry affected the information in observations.

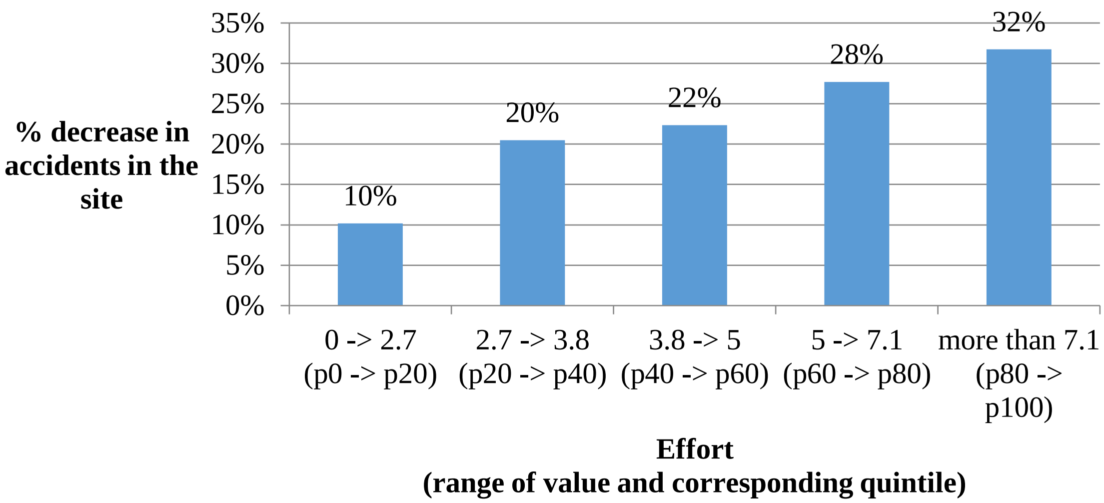


Figure 4-4. The impact of BAPP varies according to Effort. <u>Note for figure:</u> To build this graph we plot the derivative of accident on BAPP, and assume that the sites keep a fixed diffusion in the second quintile (0.04 to 0.08) and then activate the different effort dummies.

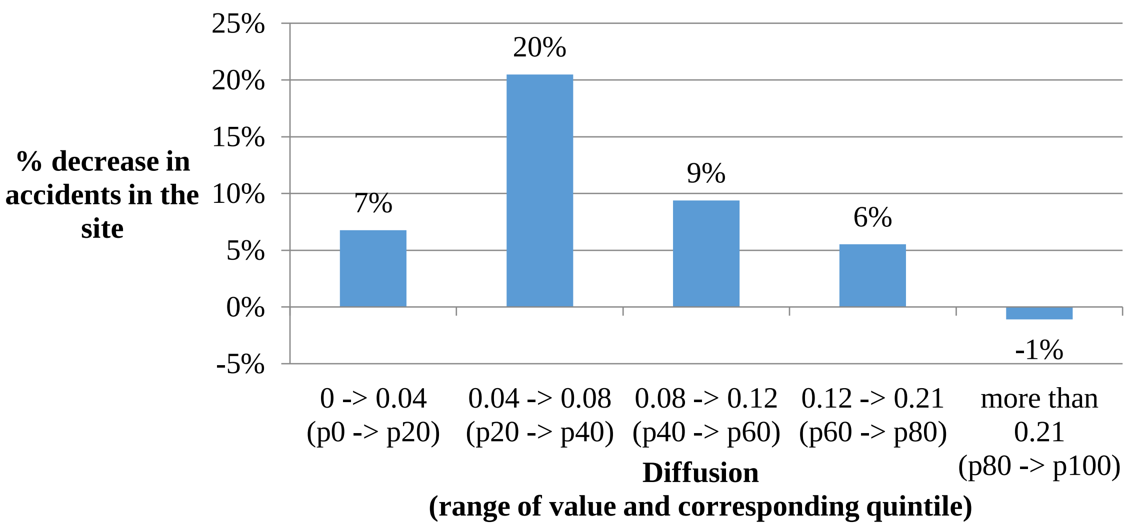


Figure 4-5. The impact of BAPP varies according to diffusion. <u>Note for figure:</u> To build this graph we plot the derivative of accident on BAPP, and assume that the sites keep a fixed effort in the second quintile (2.7 to 3.8) and then activate the different diffusion dummies.

5. Recovery of Cooperation: Evidence from a Field Experiment

In the previous section, we found support for Hypothesis 1. In this section we test Hypothesis 2 using a field experiment.

5.1. Setting

We executed the experiment in the years 2017 and 2018 in Chile. We collaborated with the Chilean Safety Association (ACHS) and one of its clients, SODIMAC. ACHS is one of the three non-profit organizations that provide services in occupational safety and health (OSH) (prevention, medical treatments, disability pensions and subsidies). ACHS partnered with DEKRA in 2012 in order to implement BAPP in its affiliated firms. DEKRA provided deep training to ACHS personnel for several years, generating the capability to deliver BAPP. This included the training and mentoring of a cadre of BAPP consultants within ACHS, sharing handbooks, guidelines, IP and software. DEKRA also allocated permanent DEKRA staff to ACHS.[12] SODIMAC is a home-improvement-store company that has operations across South America. In Chile they employ 20,000 employees and own approximately 75 stores scattered across the country. A SODIMAC store typically employs between 200 and 350 workers. SODIMAC had already implemented BAPP in five stores and a distribution center, all of which started in 2014. In 2017 they announced the implementation of BAPP in four additional stores, which started their implementation in a staggered fashion between June 2017 and October 2017 (see Appendix **Table A-12** for exact dates). We were allowed to experimentally modify these implementations from their start to June 2018[13].

5.2. Experiment Design

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 of Chile, the "Antofagasta" store. These stores had BAPP eligible workforces of 258, 268, 334 and 234 workers, respectively (excluding managers such as supervisors and area/line managers) (see **Table A-12**). Three BAPP consultants executed the BAPP implementation (the two Santiago stores shared the same consultant). We discussed the experimental treatment guidelines with

---

[12] One big difference between BAPP implementation in ACHS and implementation normally executed by DEKRA is that firms affiliated to ACHS do not pay the (high) cost of BAPP implementation. Just like other prevention services, ACHS finances BAPP with the insurance premium paid by firms. We believe that this, if anything, can play against the success of BAPP, as payment typically provides 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, in the context of the theory discussed, the condition of voluntary cooperation).

[13] At the start of the experiment, the end date was defined as "mid-2018". The participants of the experiment were not informed about this approximate date. Consultants were informed but requested not to tell any person in the intervened stores about it. Around January 2018, it was agreed with the senior SODIMAC manager sponsoring the experiment to run the experiment until June 2018. Thus, given non-negligible possibility of leakage, and in order to avoid a "last-period" drop in the collaboration of the sites, we decided to communicate to the consultants in early May that the experiment would end in June 2018, but we internally committed to executing the analysis of the experiment with the data until the end of May 2018 only.

them. These guidelines included the context of the research, the design of each treatment, a detailed implementation protocol, a communication protocol and materials. The communication of the research project was scripted in order to avoid elements that might affect or bias the reaction to our experiment (see Appendix A.12 for details). The three treatments were designed during the last quarter of 2016, after which they were revised and approved by the IRB of the Cambridge Judge Business School. The experiment was pre-registered in July 2017 on the AEA registry for randomized controlled trials (ID: AEARCTR-0002350).

Treatment 1 ("structure") was the baseline treatment and was applied to all four stores. Hypothesis 2 can be assessed by the impact of this treatment. Treatments 2 ("identity") and 3 ("observability") aimed to explore conditions that might boost (or hinder) the impact of Treatment 1 and were applied to only two stores each. **Table 5-1** displays which store received which treatment. Each treatment profile was randomly assigned to the stores (i.e., the assignment of the columns of **Table 5-1**). Treatments 1 and 2 were within-store, while Treatment 3 applied to the whole store. This structure of treatments can help disentangle the mechanism through which the small groups exert their impact. If Treatments 2 and 3 do not add anything to Treatment 1, we can point to repeated interactions as the driving mechanism. If Treatment 1 by itself does nothing, and all the action is in the interaction with Treatment 2 (Treatment 3), the mechanism through which small group work is identity (social control).

Table 5-1. Distribution of treatments across sites

	Antofagasta Store	Temuco Store	Huechuraba Store	La Reina Store
T1: structure	X	X	X	X
T2: identity		X		X
T3: observability			X	X

Treatment 1: Structure. In each of the four sites, we generated structure specifying "who is to be observed by whom", in the following way. Suppose the starting team had f observers (excluding the enabler). Half of the observers were randomly chosen and then each received the random assignment of $1/(f+1)$ of the workers in the store in the form of a printed list. The selected observers were restricted to observing their assigned workers. This was the treatment group. The remaining observers, plus the enabler, could execute observations freely across all remaining workers not assigned to a specific observer (a list of these workers was provided to the non-selected observers). This was the control group, representing the standard BAPP, where no structure was imposed. Randomization of observers was produced by the consultant using a lottery box in a starting team meeting in the fourth month, before training on observations. In the case of an odd number of observers, the even number below the mid-range was used. Randomization of workers was done by the researchers beforehand, preparing the worker

lists ready for distribution to observers. Randomization of workers to observers was stratified by sex, age, tenure and task (e.g., cashier). Before or at his/her first observation, the selected observer handed a letter to his/her assigned workers. The letter, reproduced in Appendix A.10, briefly introduced BAPP and then indicated that he/she would be the assigned observer. Crucially, in order to avoid priming group identity (in contrast to Treatment 2), at no point was there any explicit mention of the notion of a "group". This was emphasized to the consultants. What about new observers? A new observer was instructed to execute observations of the workers on his/her list of origin (either a particular treatment group or the control group at large). For new observers under treatment, an updated letter was delivered to the workers informing them about the addition of the new observer(s). Online appendix A.12 provides details of the implementation of this treatment.[14] With 5 groups in Treatment 1, the logic explained above suggests that the likelihood of repeated interactions increased by a factor of 5 under treatment as opposed to the control.

Treatment 2: "Identity". In the "La Reina" and "Temuco" stores, we modified the letters that were given to the workers in Treatment 1 by adding three elements. First, we added the notion of a group of workers to the letter. Second, we assigned a simple name to each group: "Group 1", "Group 2", and so on. Third, 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 A.10.

Treatment 3: "Observability". In the "Huechuraba" and "La Reina" stores, we published on the bulletin board of the site the number of observations carried out by all the observers at 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 and publish it on the bulletin board of the site. We certified execution by requesting photographic evidence of the report's publication. In the online appendix A.13 we display the report.

Pre-experiment power calculations. Assuming power of 80% and significance of 5%, and using data on observations from the DEKRA data set and on workplace accidents from SODIMAC (we had access to data from 2014 on), we calculated the effect size that our experiment would allow us to detect. Intra-class correlation (i.e., within-store) is low, around 0.1 for both observations and accidents. We expected to have 70 observers on average,

[14] The summary is as follows. A store had on average 10 observers in the starting team and 250 workers. Thus, roughly 5 observers and 125 workers were randomly matched in treatment groups of 25 workers. The remaining 5 observers could freely observe the remaining 125 workers, as in a standard BAPP implementation. Across 4 sites, we had approximately 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. The sites grew steadily so that in May 2018 the total number of observers was 92.

which would allow us to detect a minimum effect size of 1.7 observations per month. The four stores have 1,000 workers, which would allow us to detect a minimum effect size of 0.015 workplace accidents per worker per month. However, there are power gains from having panel data (Mckenzie, 2012); this reduces the size of the minimum detectable effect by approximately 40% to roughly 1 observation (equivalent to 44% of a standard deviation) and 0.009 accidents (equivalent to roughly 12% of one standard deviation in workplace accidents).

Exit interviews. In June 2018, we visited the sites and executed exit interviews with the consultant, the enabler, a group of 3 observers and 3 workers in Treatment 1, and a group of 3 observers and 3 workers from the control group. We executed a structured interview format, avoiding leading questions. The objective of these meetings was to gather qualitative evidence on the mechanisms that might have generated the results.

5.3. Data

We used two data sets. The first is a panel data set of observers and months of BAPP implementation. We recorded the name of the observer, the number of observations, the information encoded in these observations (number of coached observations, number of CBI behaviors observed/reported, number of risky/safe behaviors), whether the observer was a member of a starting team or a new observer, and the treatment(s) that he/she was allocated to. In the second data set, we built a monthly panel of workers and accidents, from January 2016 to May 2018. From SODIMAC's personnel registers, we have information about all the workers in each month in each of the four participating stores, plus information about their age, tenure, sex and job title. A worker was assigned to a treatment or a control condition in a randomized fashion. Using the first data set, we assigned the status of active observer to the workers that were executing observations. To study the impact on accidents, we merged our personnel data with the information that ACHS provided containing all the accidents that occurred at SODIMAC. Each accident was indexed by the time of the accident, the ID of the injured worker, the type of accident (e.g., with or without lost days), and the number of lost days due to the accident.

Balance of covariates. We executed two randomizations: workers to treatment groups or control groups (executed by the researchers), and observers of the starting team to treatment groups or control groups (executed by the consultant on the ground). **Table A-13** and **Table A-14** in the online Appendix A.14 show that the treatment groups and control are well balanced. This indicates that the randomizations were effectively executed.

Take-up. The lists of workers that we distributed to observers (plus the letters to workers) might not have been sufficient to secure compliance with the groups. As a consequence, we monitored the degree to which observers executed observations within their assigned group. We implemented a short survey to gather information about the treatment take-up. The enabler of the store conducted the survey on randomly drawn workers that had been assigned to Treatment 1. The survey was conducted between January 2018 and May 2018, after the store had reached an accumulated contact rate of one. **Table A-15** in the online appendix A.14 presents the results. Averaging across stores, 92% of the workers surveyed indicated that they knew about the implementation of BAPP in their store (8% had not yet received observations), and, of these, 92% knew they had an exclusive observer assigned to them. Of those who knew they had assigned observers, 78% remember having received the letter from their respective observer. We then asked for the number of observations and how many of these were made by their assigned observers: we found that 85% of the observations were realized by their assigned observer. This indicates that Treatment 1 was effectively implemented in stores, although not perfectly. Therefore, the impact of Treatment 1 needs to be interpreted as an intent-to-treat effect, a lower bound of the "real" effect with 100% compliance.

5.4. Results

5.4.1. Impact on observations, coaching and worker behavior

To study the impact of the treatments on the observations per observer, we use the following model:

$$OBS_{ijt} = b_1 + b_2 \times TREAT1_{ij} + b_3 \times TREAT1_{ij} \times TREAT2_{ij} + b_4 \times TREAT1_{ij} \times TREAT3_{ij}$$

$$+ b_5 \times NEW_{ijt} + b_6 \times ENA_{ijt} + b_7 \times TEN_{ijt} + b_8 \times TEN_{ijt} \times NEW_{ij} + v_{jt} + u_{ijt} \quad (4)$$

In this model we regress the number of observations by observer i in store j in the month t on the treatment dummies. Treatment 2 and treatment 3 enter as interaction effects on treatment 1. We assess Hypothesis 2 by evaluating the coefficient b_2. To explore the underlying mechanisms we use the coefficients b_3 and b_4 (as well as other ad-hoc tests). If b_2 and b_4 equal zero and b_3 is positive, then small groups work through (or require) identity and not the promotion of dyadic repeated interactions. The driver of cooperation is observability if b_4 is the only positive and significant coefficient. Given that Treatment 1 by design only increases the expectation of dyadic interactions, if b_2 is the only significant coefficient, then interactions are the mechanism driving our results. Of course this doesn't completely rule out other alternatives. This identification of b_2 (and b_3 and b_4 equal to zero) with the mechanism of repeated interaction hinges on: i) other potential mechanisms are theoretically ruled out (not implausible since two major alternatives are already discarded), and ii) the degree to which Treatment 1 creates a basic and clean structure where the only mechanism being activated is the "primordial" mechanism of repeated interactions.

We control by tenure, or 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 dummy variable NEW takes the value of 1 if the observer is not part of the starting team. **Figure 4-2** shows 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 4-2**. We also control for store and month with dummies (v_{jt}), which is necessary 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 to these treatments. We also control for the enablers by identifying them with the dummy ENA. Enablers were not part of the randomization and were instructed to execute observations in the control group. This introduced a downward bias in b_2 because enablers typically execute more observations than the rest of the observers (excluding them from the sample yielded consistent results).

In the discussion of Hypothesis 1 (Section 4), we show that cooperation in BAPP breaks down in with diffusion because the newer observers, due to decreasing gains in status or career prospects, perform fewer observations (and have shorter tenure as such). Therefore, we expect that the impact of adding structure will have a larger impact on new observers than on observers in the starting team. To allow for this, we extend the model:

$$\begin{aligned} OBS_{ijt} = {} & b_1 + b_2 \text{ x } TREAT1_{ij} \text{ x } NEW_{ij} + b_3 \text{ x } TREAT1_{ij} \text{ x } START_{ij} \\ & + b_4 \text{ x } TREAT1_{ij} \text{ x } TREAT2_{ij} + b_5 \text{ x } TREAT1_{ij} \text{ x } TREAT3_{ij} + b_6 \text{ x } NEW_{ij} \\ & + b_7 \text{ x } ENA_{ijt} + b_8 \text{ x } TEN_{ijt} + b_9 \text{ x } TEN_{ijt} \text{ x } NEW_{ij} + v_{jt} + u_{ijt} \end{aligned} \tag{5}$$

Model (5) splits the impact of Treatment 1 into two components: the impact on new observers and the impact on observers that are part of the starting team (START, which is equal to 1 minus NEW).

We display the results in **Table 5-2**. Column (1) indicates that Treatment 1 generates an increase of 0.97 observations, significant at 90%. This impact is just below the minimum detectable effect of one observation (assuming power at 80% and significance at 5%). Column (2) shows that this impact is concentrated on the new observers. These observers conduct 1.38 more observations, significant at 95%.[15] Observers in the starting team 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 starting team member, an effect size that is very similar to the difference depicted in **Figure 4-2** with the DEKRA administrative data. This result suggests that Treatment 1 operated as intended: it reduced the breakdown of cooperative effort as the number of observers increased, particularly for new observers whose effort was most affected by size. This supports Hypothesis 2.

[15] Common shocks within a store can generate correlations in the standard errors. 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). In column (2), we obtain a p-value of 0.165 for the coefficient of T1 x NEW, and of 0.065 for T1 x START. However, it is not obvious that we need to correct. According to Abadie et al. (2017), on experimental design grounds, clustering by store is not necessary in our case: Treatment 1 is executed within stores. On sampling design grounds, we should not cluster either: we do not randomize stores for Treatment 1.

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 from Treatment 1 is eliminated if the groups have a name and the names of the group members are revealed in the letter. Exit interviews suggested a clear explanation of this: (partially) lifting the anonymity condition of BAPP by revealing names through letters generated a backlash from the workers; the interviewees indicated that providing the names of workers jeopardized the BAPP promise of "no spying, no naming, no blaming". This backlash translated into lower worker willingness to collaborate with observers, which in turn affected the observers' efforts. DEKRA's and ACHS's consultants concurred with this, pointing out that workers worried a lot about being "spied on" and "ratted out" by observers. This is precisely why BAPP implementations emphasize and protect anonymity, consistently using the motto "no spying, no naming, no blaming". This effect may even have been exacerbated in our setting: SODIMAC experienced a strike involving 30% - 40% of workers between November and December 2016. Labor relations within the company became quite tense after this strike. This intensified the feeling of being "spied on" or "ratted out".

This result of Treatment 2 suggests that a distaste for the violation of anonymity was stronger than any identity effects that might have been generated. This result is novel to the literature, where transparency (broadly defined) is generally advocated because it fosters identity-building or reputation dynamics. This is natural and appropriate when the cooperative act entails providing a "positive" signal to the recipient. In our case, the recipient was told to change an *erroneous behavior*, which generated a *negative* signal and imposed a cost on the recipient if anonymity was not secured, especially in this sensitive context.

We do not find an effect of Treatment 3 on the number of observations. In conjunction with the result in Treatment 2, this suggests that the mechanism driving the results of structure is repeated interactions, and neither identity nor social control.

An additional type of cooperative behavior that observers can engage in is "coaching". We explored the impact of the treatments on the amount of coaching that the observers received (the BAPP system registers the presence of coaching in an observation, but not who is the coach). In columns (3) and (4) of **Table 5-2**, we replicate the analysis using the number of coached observations as the dependent variable. We use a POISSON regression because this variable behaves as a count variable (no substantial changes occur if we use OLS). Column (3) shows that Treatment 1 increases the amount of coaching that the observers receive, and column (4) shows that this effect is concentrated on new observers. Assuming covariates are set to zero, the impact of being a new observer without Treatment 1 is $\exp(1.02)=2.77$ coached observations, whereas adding Treatment 1 generates $\exp(1.02+0.4)=4.13$ coached observations. Therefore, Treatment 1 generates 1.36 additional coached observations. By contrast, for the starting team members, having no Treatment 1 generates $\exp(0)=1$ coached observations, while adding Treatment 1 generates exp(0.44)=1.55. Therefore, Treatment 1 generates only 0.55 additional coached observations, which is much lower than for new observers. We do not find an impact of Treatments 2 or 3 in the amount of coaching, lending additional support to the idea that structure affects cooperation through repeated interaction, not identity or social control.

Column (5) explores whether coaching mediates the impact of Treatment 1 on observations by adding coached observations as a control. Coaching exerts a strong positive impact on the number of observations (this is robust to adding observer fixed effects). However, coaching captures only a marginal share of the impact of Treatment 1. The coefficient of "Treatment 1" drops from 0.58 in column 2 to 0.41 and the coefficient of "Treatment 1 x new observer" drops from 1.32 to 1.22. This indicates that the driving mechanism behind Treatment 1 is not help received as coaching. Thus, given that coaching is a cooperative act on its own, this result enhances the confidence in the pattern we are uncovering: Treatment 1 is effective by itself, not just in conjunction with Treatments 2 or 3, and especially in new observers.

Observers had to record on their observation sheets whether the observed behaviors were executed in a safe or a risky manner. In other words, this represented an observer-reported measure of how safely workers were executing their tasks.[16] Columns (6) and (7) of **Table 5-2** present the impact of the treatments on the number of risky behaviors recorded by the observers. We control for the number of observations and the total number of recorded CBI items, in order to not capture merely a "volume" effect (i.e., more sheets lead mechanically to more risky behaviors). The result shows that risky behavior is significantly lower in Treatment 1, and this effect is again concentrated on new observers. This shows that our treatment mattered: increased observations by adding structure translated into a change in worker behavior. Again, we find that Treatment 2 reverses the beneficial impact of Treatment 1 and we find no effect for Treatment 3. The consistent results across three dependent variables provide increased confidence in the support of Hypothesis 2 and its underlying mechanism of repeated interactions.

[16] Regarding the number of CBI behaviors observed and reported by the observer in the sheet, it could be argued that they also constitute a measure of observer effort. We analyzed the impact of the treatments on the total number of recorded CBI behaviors, conditional on the number of observations, but we did not find any significant impact.

Table 5-2. Impact of treatments on number of observations, coaching and risky behavior

	Observations (1)	Observations (2)	Coached observations (3)	Coached observations (4)	Observations (5)	Risky behaviors (6)	Risky behaviors (7)
Treat. 1	0.97* (0.53)		0.42** (0.19)			-0.99* (0.52)	
Treat. 1 x starting team observer		0.58 (0.66)		0.44*** (0.22)	0.41 (0.64)		-1.09 (0.70)
Treat. 1 x new observer		1.38** (0.57)		0.40** (0.21)	1.22** (0.52)		-0.89* (0.53)
Treat. 1 x treat. 2	-1.52** (0.67)	-1.56** (0.68)	-0.14 (0.22)	-0.14 (0.22)	-1.52** (0.63)	1.15* (0.68)	1.14* (0.68)
Treat. 1 x treat. 3	-0.74 (0.61)	-0.51 (0.64)	-0.27 (0.20)	-0.28 (0.21)	-0.43 (0.61)	0.14 (0.70)	0.20 (0.75)
Enabler	3.40** (1.37)	3.28** (1.34)	0.49*** (0.16)	0.49*** (0.16)	2.87** (1.19)	0.76 (0.71)	0.74 (0.73)
Tenure	0.12 (0.14)	0.12 (0.14)	0.02 (0.05)	0.02 (0.05)	0.11 (0.13)	-0.08# (0.13)	-0.07# (0.13)
Tenure x new observer	-0.04 (0.16)	-0.04 (0.16)	-0.38*** (0.09)	-0.39*** (0.09)	0.11 (0.15)	-0.16# (0.16)	-0.15# (0.16)
New observer	-1.17 (0.88)	-1.60* (0.91)	1.00*** (0.38)	1.02*** (0.39)	-2.17** (0.84)	0.62 (1.06)	0.51 (1.10)
Coached observations					0.59*** (0.11)		
CBI items						0.02 (0.01)	0.02 (0.02)
Number of observations						0.48*** (0.15)	0.48*** (0.15)
Store-month fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	585	585	585	585	585	585	585
R-squared	38.95%	39.33%	21.69%	21.69%	44.49%	49.73%	49.75%
Mean (Standard deviation)	5.02 (2.82)	5.02 (2.82)	1.15	1.15	5.02 (2.82)	3.47 (0.69)	3.47 (0.69)

All regressions are estimated with OLS, except for (3) and (4), which are POISSON regression. 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. Results are robust to the inclusion of the interaction of treatments 2 and 3 with new and starting team observer.

The fact we find no effect for Treatment 3 is consistent with Roberts' (2008) prediction that information coming from extensive personal experience (high repeated interaction in Treatment 1) tends to dominate the use of indirect information that is used in social control (Treatment 3), such as reputational standing or effort contrasted against a collective norm. This does not mean that social control in and of itself cannot have an independent and positive impact on cooperation (e.g., Bandiera et al, 2005; Khadjavi, 2016). The issue is that our design cannot detect this main or individual effect, only the interaction with Treatment 1; that is, we measure whether, in the context of small groups that facilitate repetition of contact, Treatment 3 is a necessary condition to generate cooperation.

5.4.2. Impact on the likelihood of becoming an observer

So far we have analyzed cooperative effort, contingent on becoming an observer. However, cooperation in BAPP also entails becoming an observer in the first place. We use the following model to study this:

$$OBSERVER_{ijt} = b_1 + b_2 \text{ x } TREAT1_{ij} + b_3 \text{ x } TREAT1_{ij} \text{ x } TREAT2_{ij} + b_4 \text{ x } TREAT1_{ij} \text{ x } TREAT3_{ij} + X_{it} + \tau_{tj} + u_{ijt} \quad (6)$$

To estimate equation (6), we use all BAPP eligible workers at the site, excluding those who are part of the starting team. $OBSERVER_{ijt}$ is a dummy variable that takes the value of 1 if a specific worker i in a store j is an active observer in month t, and zero otherwise. TREAT1 is a dummy that takes the value of 1 if that worker is under treatment 1, and zero otherwise. The same is true for TREAT2 and TREAT3. X_{it} is a vector of controls at worker level for each period (age, tenure, gender and job title). τ_{tj} are fixed effects at the store and the calendar-month level. **Table 5-3** presents the results.

Table 5-3. Impact of the treatments on the probability of becoming an observer.

	P(observer) (1)	P(observer) by May 2018 (2)
Treat. 1	0.019# (0.013)	0.054** (0.025)
Treat. 1 x treat. 2	-0.021* (0.012)	-0.072** (0.028)
Treat. 1 x treat. 3	-0.009 (0.010)	-0.006 (0.027)
Individual controls	Yes	Yes
Store-month fixed effects	Yes	No
Store fixed effects	No	Yes
Observations	10,879	1,072
R-squared	0.027	0.011
Mean	0.022	0.052

OLS. Errors in parentheses: Robust and clustered at worker level. # p<0.15, * p<0.1,** p<0.05, *** p<0.01. All regressions exclude starting team members. Sample restricted to months and stores with BAPP already implemented.

The sample in column (1) includes all the months of BAPP implementation. Given that this sample includes the initial months where recruiting was non-existent, in column (2) we consider the workers in May 2018 only. The results indicate that Treatment 1 increases the likelihood of becoming an observer by 1.9 percentage points over the timeframe of our experiment, which is almost equivalent to the mean likelihood of 2.2%. For May 2018, the results are equivalent but more precisely estimated: 5.4 percentage points increase over a mean of 5.2. Again, we find in both samples that Treatment 2 completely reverts the impact of Treatment 1. Treatment 3 is again not significant. As before, the consistent results across different dependent variables –which now amount to four– provides high confidence in the support of hypothesis 2 and its underlying mechanisms.

5.4.3. Impact on accidents

We study six different measures of accidents registered by ACHS: total accidents, and their breakdown into work accidents (i.e., accidents that take place at the workplace), commuting accidents (i.e., accidents that take place between home and the workplace) and quasi-accidents (incidents that do not meet the conditions to be attended to by ACHS, mostly because they are not a workplace incident, but also because they are not meaningful or real incidents). We further break down work accidents into two sub-groups: without lost working days and with lost working days. Finally, in the case of lost days, we also consider the length of leave.[17]

We first look at the impact of BAPP as a whole, replicating the test in Section 0. This allows us to evaluate the impact of the experimental treatments against the baseline impact of BAPP. We present the details of the analysis in the online appendix A.15. We find that BAPP reduces work accidents over time, and this effect is focused on work accidents without lost time.[18] (This is consistent with the safety literature, which suggests that more severe accidents might have a different data-generating process, less related to worker behavior –the lever that BAPP can affect– and more to investments in equipment and maintenance.) The impact is not small: BAPP is associated with a reduction of 0.0015 work accidents per worker per month in the first year, which is equivalent to 35% of the variable's mean. This effect size is similar to the one estimated with archival data in Section 0. This effect is not driven by observers having fewer accidents. Instead, we find that the observers, in addition to receiving the baseline benefit of BAPP, also experience fewer accidents with lost time (i.e., more severe accidents).[19]

Now we turn to the impact of our treatments. We use the following model:

[17] 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). 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.

[18] We find no impact on commuting accidents and quasi-accidents. This acts as a falsification test, as we would not expect BAPP to generate an impact in these types of accident.

[19] We explored whether this impact varied over four observer types (new/starting-team and treated/control). However, smaller cells imply a very small number of accidents, as these are infrequent. This precluded a meaningful analysis.

$$ACCIDENT_{ijt} = b_1 + b_2 \text{ x } TREAT1_{ij} + b_3 \text{ x } TREAT1_{ij} \text{ x } TREAT2_{ij} + b_4 \text{ x } TREAT1_{ij} \text{ x } TREAT3_{ij} + X_{it} + \tau_{tj} + u_{ijt} \quad (7)$$

Treatment dummies take the value of 1 if that worker is under treatment 1, and zero otherwise. 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. X_{it} and τ_{tj} are the same as above. **Table 5-4** presents the results. Consistent with our previous results, we find that Treatment 1 alone reduces workplace accidents, but this is reversed by Treatment 2.

Table 5-4. Impact of treatments on accidents

Panel a)	Total accidents		Workplace accidents		Workplace accidents without lost working days		Workplace accidents with lost working days	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treat 1	-0.0003 (0.0017)	-0.0031 (0.0026)	-0.0007 (0.0012)	-0.0030** (0.0015)	-0.0014* (0.0086)	-0.0022* (0.0012)	0.0069 (0.0080)	-0.0083 (0.0087)
Treat. 1 x treat. 2		0.0072** (0.0033)		0.0047** (0.0022)		0.0034** (0.0016)		0.0013 (0.0015)
Treat. 1 x treat. 3		-0.0035 (0.0034)		-0.0013 (0.0024)		-0.0030* (0.0018)		0.0016 (0.0017)
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Store-month fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	11,277	11,277	11,277	11,277	11,277	11,277	11,277	11,277
R-squared	0.0071	0.0076	0.0071	0.0075	0.0044	0.0051	0.0058	0.0059
Mean	0.0081	0.0081	0.0037	0.0037	0.0019	0.0019	0.0018	0.0018
Panel b)	Commuting accidents		Quasi-accidents		Length of leave	Length of leave		
	(1)	(2)	(3)	(4)	(5)	(6)		
Treat. 1	-0.0006 (0.0085)	-0.0024 (0.0016)	0.001 (0.001)	0.0022 (0.0018)	-0.056 (0.0347)	0.0098 (0.0264)		
Treat. 1 x treat. 2		0.0031* (0.0017)		-0.0006 (0.0019)		-0.103 (0.0678)		
Treat. 1 x treat. 3		-0.0001 (0.0016)		-0.0028 (0.0018)		-0.0107 (0.0549)		
Accident with lost time					12.978*** (4.438)	12.985*** (4.442)		
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes		
Store-month fixed effects	Yes	Yes	Yes	Yes	Yes	Yes		
Observations	11,277	11,277	11,277	11,277	11,277	11,277		
R-squared	0.0032	0.0035	0.0052	0.0054	0.1819	0.1821		
Mean	0.0019	0.0019	0.0026	0.0026	0.045 (12.97)†	0.045 (12.97)		

OLS regressions. The results are consistent if we use count models and drop the individual-level controls as independent variable errors in parentheses: robust and clustered at worker level. † 12.97 is the days of leave conditional on having an accident. The results do not change if we use only cases of accidents. * p<0.1,** p<0.05, *** p<0.01.

The impact is fully concentrated on accidents without lost working days, the type of accidents that BAPP affects (see above). The impact of Treatment 1 without Treatments 2 and 3 in column (3) is a decrease of 0.003 accidents per worker per month. This impact is equivalent to one-third of the overall BAPP impact, a sizeable effect.[20] As a novel result, we find that the effects 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 our treatments, but not BAPP as a whole (see **Table A-16**), can generate benefits beyond the work environment. Neither quasi-accidents nor length of leave are affected by the treatments.

Regarding Treatment 3, we find that it generates a significant boost to Treatment 1 in work accidents without lost working days. The size of the effect is large: Treatment 3 more than doubles the baseline effect of Treatment 1. This result is unexpected, as previous tests of Treatment 3 yielded no impact. Given to prior impact of observations or risky behavior, it must be that the effect of Treatment 3 is operating either through higher quality of observations or higher engagement-motivation-compliance from observed workers. Given that Treatment 3 provides observability on the number of observations, not their quality, the former is unlikely. Instead, higher motivation by observed workers is plausible. As Treatment 1 boosts effort by observers (i.e., more observations), and Treatment 3 makes this observable, a worker may now have more incentives to comply: the high effort of his/her assigned observer is now public, so the responsibility is shifted to him/her.

5.5. Additional Evidence on the Mechanism of Repeated Interaction

Our findings so far (Treatment 1 is impactful but not Treatments 2 and 3) provide evidence for the repeated interaction mechanism. We now provide three additional pieces of evidence: random coaching, finer grained analysis of peer pressure among observers, and consistent exit interviews.

First, we explored how much of the additional coaching received by a new observer documented in **Table 5-2** comes from observers within his/her own group, versus from other treatment groups or the control group.

[20] This is small compared to the pre-experiment minimum detectable effect (MDE) of 0.009 workplace accidents. However, the mean of workplace accidents in Sodimac decreased from 0.0055 in 2014 to 0.004 in 2018, reducing the MDE to 0.007. If one considers accidents without lost working days, the MDE is 0.005, which is closer to the estimated effect of 0.003. Nevertheless, considering the variance of accidents with lost working days in 2018 (and no gains from panel data), the ex-post power for the effect we estimated is 20%. This means that in our sample, there is a 20% chance of detecting the effect we observe if we assume that it is there to be found. Ioannidis (2005) showed that insufficient power can also cause high rates of false positives. Ioannides (2005) recommends calculating the positive predictive value (PPV), which reflects the likelihood that a statistically significant finding actually reflects a true effect. In our case, the PPV for “treat. 1” equals [0.2*R/(0.2*R+0.025)], where 0.2 is the power, 0.025 the statistical significance in **Table 5-4** and R is the ratio of “true relationships” to “no relationships” in the population of studies to this one (R can be very low in fully empirical and a-theoretical fields such as genome-disease association studies). Given that all the previous findings in the paper provide a decent prior for the analysis on accident, we set R to 0.5. This yields a PPV of 0.8, meaning that there is an 80% chance that the statistically significant finding we uncover actually reflects a true effect (if R is set to 0.25, PPV is equal 0.66).

For all the coaching events for new observers, we hand-collected the name of the observer that executed the coaching. **Table A-17** of the online appendix A.16 provides detailed analysis. We find that the coaching of observers in Treatment 1 was not performed preferentially by members of their respective groups. The coaching was just as likely to come from other groups of Treatment 1 or from the control. This suggests that "active help" among observers within groups of Treatment 1 was weak. As it is reasonable to expect identity and social control to lead observers within a group to help one another, this represents evidence against these mechanisms. Instead, new observers seem to be more motivated by participating and becoming better in BAPP, which is consistent with reciprocal behavior spurred by the repeated interactions that only Treatment 1 can generate.

Second, we discuss peer pressure among observers. The number of observations executed by each observer was frequently displayed and discussed at the monthly meetings of the starting team. Our interviews suggest that this monthly discussion generated ample peer pressure on those observers that did not execute their share. To explore this, we executed the regressions displayed in **Table 5-5**. 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.

Table 5-5. Impact of observation ranking and its interaction with treatment 1 and 3

	Observations (1)	Observations (2)	Observations (3)	
			Starting team observers	New observers
Low rank in last month	0.56 (0.54)	2.11*** (0.78)	2.21** (1.08)	1.38*** (0.42)
Treat. 1 x low rank in last month		-2.19** (0.76)	-2.82*** (0.96)	-0.06 (0.62)
Treat. 3 x low rank in last month		-1.30† (0.86)	-1.28 (1.13)	-0.51 (0.63)
Tenure	Yes	Yes	Yes	
Tenure x new observer	Yes	Yes	Yes	
Observer fixed effects	Yes	Yes	Yes	
Store-month fixed effects	Yes	Yes	427	
Observations	585	585	585	
R-square (adjusted)	63.98% (47.98%)	65.51% (49.69%)	66.03% (50.01%)	

Errors in parentheses: robust and clustered at the observer level. † p<0.15 / * p<0.1 / ** p<0.05 / *** p<0.01. Parameters in column 3 are estimated in the same regression; we display them in parallel for presentation convenience. The results are robust to: i) adding lagged observations as a control (this controls for a possible "reversion-to-the-mean" effect); ii) inclusion of treatment 2 and its interactions; and iii) a continuous variable of ranking (instead of a dummy).

This variable displayed plenty of within-observer variance, which allowed us to add observer fixed effects. Column (1) indicates that a low rank in the previous month did not generate a significant change in observations. Column (2) shows that low rank *did* incentivize observers to increase observations if they were *not* part of Treatment 1. This again suggests that Treatment 1 did not operate through social control, in this case the data displayed at the starting team's meetings. Instead, the exit interviews suggest that the observers under Treatment

1 became responsible for their own group of workers, not sharing responsibility with other starting team members, and thus became "liberated from peer punishment" and insensitive to peer pressure. This liberation of peer responsibility effect appears only for observers in the starting team, as the only group that met regularly and could apply exert pressure among its members. Column (3) tests this: in one model, we disaggregate three low-rank variables into types of observer by multiplying them with the starting team and new observer dummies. While we confirm that the negative effect is concentrated on starting team observers, we fail to find a positive effect on new observers. This provides more convincing evidence that Treatment 1 did not operate via social control (in this case social control during meetings of the starting team).

For Treatment 3, we find a similar negative interaction effect, but weaker and less statistically significant. This is consistent with the substitution of "public" social control (reputation effect of the public display on the bulletin board) for "private" social control (display of observer statistics in the monthly meetings of the starting team) by which observers "externalize" the cost of punishing. However, this substitution is only partial, as the "low-rank" dummy remains significant and larger than the interaction term. This partial substitution, with the persistence of this "private" social control, can help explain why Treatment 3 did not show results: the manipulation by Treatment 3 was not strong enough to overtake the "private" social control mechanism.

Third, the exit interviews provide compelling accounts from workers and observers in favor of the repeated interaction interpretation. Workers from Treatment 1 said that having the same person coming over and over for observations created a higher level of commitment because "you cannot hide", as an interviewee put it. Another worker commented, "It is like being counselled by your father, and not any random guy… you will meet you father again, so you better comply". Observers from Treatment 1 mentioned that after a few interactions with the same person, they became more invested, caring more about really helping the person; "It created a kind of a bond", an interviewee indicated.

5.6. Robustness checks

There are three main candidates for alternative explanations of our findings. We explore each in turn. After that, we explore whether we can find consistent results using the administrative dataset.

Self-selection. The positive impact of BAPP and our treatments might reflect that the workers that become observers are not randomly selected. In **Table A-18** of the online appendix we evaluate the extent of these problems by comparing observables. We find that observers are older and have a higher tenure than the rest of the workers at the site, but they are not different in terms of gender or type of job. Interestingly, we find that this difference is generated exclusively by the observers that form the starting team. New observers are no different to the workers of the site in terms of tenure, age, sex and type of job. This indicates that the results we document for new observers are not driven by selection issues. **Table A-19** of the online appendix compares starting team observers and new observers using a survey filled out by observers.[21] We do not find any differences in terms of personality traits (big 5), altruism (dictator game) and size of social network. This suggests that the criteria for the selection of starting team members are age and experience, and not personality, behavioral or social traits. Therefore, barring tenure and age for starting team observers, the differences between observers and workers are not likely to be driving our results.

Leadership. Although our treatment protocol avoided tagging any role of "guide" and "leadership" to the starting team observer under Treatment 1 (and explicitly instructed the consultants not to emphasize it), these observers might still have adopted a "leadership" role towards new observers. Two pieces of evidence argue against this alternative explanation. First, the results for coaching indicate that starting team observers assigned to Treatment 1 did not help the new observers in their group disproportionately more than new observers outside their group. Second, we executed a robustness check where we controlled for starting team observer quality. We executed a two-stage model where in the first stage we use fixed effects to obtain a proxy for the quality of the observers in the starting team before the entry of new observers, and then we plugged these fixed effects into the regression of column (2) of **Table 5-2**.[22] Including this control did not alter our conclusions; if anything, the results became stronger. Furthermore, using interaction analysis, we find that having a better starting team observer is beneficial for new observers, but this is much more the case for the control group. This is consistent with the fact that the quality of the leader is more important when a new observer comes less motivated into BAPP, that is, in the control group (in Treatment 1, new observers came motivated with the desire to reciprocate the higher one-to-one effort they had received as workers). Overall, this suggests that starting team member quality (or "leadership" capacity) played a role, but did not substitute for Treatment 1.

The negative impact of treatment 2 is treatment 1 badly implemented. Regarding the negative impact of Treatment 2, an alternative mechanism could lie in the behavior of the consultants. Given that Treatment 2 is an addition to Treatment 1, it could be that the two consultants that executed it – one in Temuco and one in La Reina – executed Treatment 1 in a way that led to a negative outcome, and this "consultant effect" was picked up by Treatment 2. However, several arguments and tests indicate that this is not the case. First, the consultant in La Reina also executed BAPP in Huechuraba, a store that had Treatment 1 but not Treatment 2. Thus, if the executing consultants were the issue, we would find a negative impact of Treatment 1, because in three of four

[21] We sent an online survey to all observers immediately after the observer entered BAPP. The survey was voluntary and confidential. The survey was sent by the research team and included a terse explanation of the research project (revealing neither the topic nor the purpose of the research).

[22] In the first stage, we restricted the series of the starting team observers to the months before the entry of new observers into their specific group and we computed their fixed effects. Then, we computed a continuous variable where the fixed effects were orderly assigned, which was then plugged as a control in the second stage, which was estimated using the remaining data. The assignment of the fixed effects was as follows: a new observer in group "w" was assigned the fixed effect of the starting team member of group "w"; new observers in the control group were assigned the average of the fixed effects of the starting team observers in the control (the results did not vary if we added median or the percentiles 25 and 75); and starting team observers were assigned their own fixed effect (as expected from the addition of the new variable, the coefficient for the dummy of starting team observers was non-significant and close to zero in the second stage).

stores it would have been implemented in a “negative” way. However, we did not find this to be the case. Second, following the previous point, we executed a regression restricting the sample to the consultant in La Reina and Huechuraba (adding Antofagasta does not change the results). The results do not change: Treatment 1 increased observations and Treatment 2 decreased them; therefore, the result of Treatment 2 also occurs within one of the “suspect” consultants. Third, we executed a regression interacting Treatment 2 with the condition of being a new observer. If Treatment 2 is negative because of a workers’ backlash to “being listed”, there shouldn’t be any difference between the starting team and new observers in the negative coefficient of Treatment 2; in contrast, if bad implementation of Treatment 1 is the driving force, then the negative effect might be concentrated on new observers because this is the channel where Treatment 1 exerts its impact. We found the former to be the case: Treatment 2 is not affected by the type of observer. Fourth, Treatment 2 has a negative impact on dependent variables that capture observed workers’ outcomes (i.e., risky behavior, accidents and the likelihood of becoming an observer) or is influenced by it (i.e., observations) but a null impact on coaching, the dependent variable that exclusively captures observer behavior. This is consistent with workers being the driving force behind the negative effect of Treatment 2, and therefore closer to our proposed mechanism of a “workers’ backlash”. If the influence of Treatment 2 had come from idiosyncrasies of the consultant, the impact would also have been felt in coaching. Fifth, we explored the effect of time on the impact of Treatment 2. Treatment 2 was particularly detrimental at the start of BAPP implementation, generating a backlash of approximately two and half observations in the first couple of months. After that, the negative effect was gradually reduced, down to almost zero by the end of the experiment. This pattern is consistent with a backlash at the start, and then workers realizing that the list of names was not ill-intended and restoring effort.

“Natural” groupings in the archival data and generalizability. We return to the archival data analyzed in Section 3. As discussed, BAPP provides freedom for the site to try different implementation tactics and strategies. From DEKRA, we learned that some sites ensured that their observers specialized in different areas of a site[23] (e.g., production line, warehouse), and that, even without an area policy, some observers naturally do this anyway. This has two main effects: i) a “learning effect”: the observer learns about the tasks being performed in the area and can therefore provide better and deeper feedback to workers; ii) a “repeated interaction effect”: the observer now interacts with a reduced set of workers and this increases the frequency of interaction. In our experiment we can focus on ii) by shutting down i) via randomization. With archival data, we can measure area specialization and gauge its impact while controlling for learning. We measure area 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.[24]

[23] The observation sheet displays the different areas of the site where the observer can execute a particular observation. The set of areas is pre-defined by the starting team and stays fixed throughout implementation (typically 5 to 10 areas).

[24] We also used a measure that computes the HHI monthly, and the results did not change; if anything, they became stronger. We prefer to use 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.

Then, we average this for a site for every month (this generates some variation over time as the pool of observers changes on the site). This variable displays plenty of variance (see **Figure A-5**). More importantly, at the low end of the distribution we observe an HHI of 0.1 to 0.2, which is consistent with random observations across 5 to 10 areas, the typical number of areas in BAPP.[25]

We estimate a model analogous to Equation (3), where we interact BAPP with experience, controlling for the interaction of BAPP with effort (as a dummy), diffusion (as a dummy), observers' tenure (measured as the number of months elapsed since the observer's first observation, averaged across the site's observers for each month) and observers' experience (measured as the cumulative number of observations up to month t-1 for each observer and then averaged across the site's observers for each month). Experience is meant to capture the impact of the "learning affect" that area specialization can foster. The results are displayed in **Table A-10**. We find that the area specialization greatly enhances the impact of BAPP. The interaction between BAPP and experience, and the triple interaction between BAPP, experience and area specialization and experience, are mute. This strongly indicates that the findings of our experiment – the benefit of repeated interactions via structure – also hold true in the administrative data set. This reduces generalizability concerns.

6. Conclusion

This paper studies cooperation in large groups. Cooperation represents a public good, 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. In our empirical setting, the host firms implemented a safety methodology whereby a small group of workers was trained to advise 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 weakening of cooperative effort as the number of cooperators increased over time, as well as potential solutions to the challenges.

Fine-grained archival data and experimental interventions in the field allowed us to dissect the anatomy of cooperation. Using a large-scale data set of previous implementations of the methodology, we first document that cooperation is beneficial: indeed, it is associated with a reduction in accidents. We also document that cooperation suffers from scale: as the number of cooperators grows, the additional cooperators display lower and less sustained cooperative effort, thereby decreasing the capacity of cooperation to reduce accidents.

We then experimentally modified the methodology, applying three treatments. The first treatment added structure to who advised whom by creating smaller groups within the site (grouping is the essence of structure, see Puranam, 2018). This added structure boosted the degree of repeated interactions between observers and

[25] We know from our conversations with DEKRA consultants that at some sites observers might be pushed to be random across areas in order to avoid what is known as "developing a blind eye", that is, observers do not see (or don't want to see) the unsafe behavior after becoming "too" familiar with the tasks of a particular area.

observed workers by a factor of five, and therefore was expected to foster self-enforcing cooperation (Axelrod, 1981; Dal Bo and Frechette, 2018; Nowak, 2006; Gibbons and Henderson, 2012). Accordingly, we found that this treatment enhanced cooperative effort and the diffusion of cooperation (i.e., more workers enrolled to provide advice), as well as reducing the incidence of risky behavior and workplace accidents.

The second and third treatment evaluated whether group identity and social control (observability of effort) might further increase the impact of the structure created by Treatment 1. In our second treatment, we added names 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 the group (Tajfel, 1982), which research has shown to act better in supporting in-group cooperation. However, we found the opposite to be true: Treatment 2 reverted the impact of Treatment 1. Exit interviews and supplementary tests indicated that the methodology's motto of "no spying, no naming, no blaming" was deeply internalized by the workers and, in their view, compromised by this treatment. They displayed a strong distaste for being "listed" or "under surveillance", generating a cost that weighed against cooperation. This finding suggests two insights. First, any improved group identity was outweighed by the violation of valued anonymity. Second, when cooperation requires the correction of erroneous behavior, and this carries a cost, anonymity might be necessary for cooperation to thrive.

In our third treatment, we explored the idea that social control – peer pressure, targeted punishment, reputation concerns – affect cooperation, and it does so better in small groups (Bandiera et al, 2005; Boyd and Richerson, 1988; Suzuki and Akiyama, 2005). Given that these social control mechanism rely on observability, we created list of observers' efforts, which we publicly displayed in the site. We found that this treatment had a negligible effect. This result is consistent with theory that indicates that conditioning on extensive prior interactions (between observers and workers) dominates conditioning on simple forms social control (Roberts, 2008).

The main contributions of our study are two. First, we confirm in a field study that cooperation easily breaks down with organizational size, which informs a literature which has found contradictory evidence for and against the claim that cooperation falters with scale (Barcelo and Capraro, 2015; Pereda et al., 2019; Zhang and Zhu, 2011). Also, we show that this breakdown is due to a decreasing marginal benefit of cooperation, in line with recent theoretical models (Pereda et al., 2019). Second, we show that adding structure to a population can be a good remedy and that this happens mainly through the repeated interaction it fosters. This informs the nature of organizational structure: it shows that its function is not only separating groups so that gains from the division of specialized labor can be achieved (Puranam, 2018), but that it also fosters cooperation levels that otherwise would be difficult to achieve.

As any empirical study, our study has limitations. First, the archival data-set findings only use sites that were selected to implement the methodology that we were studying. Although we showed that causality within the sample is likely, this might not be generalizable. Second, power in our experiment is not ideal for the accident regressions. Even though the converging results across many dependent variables increase the confidence of having detected a true effect on accidents (Ioannidis, 2005), replication of our findings is necessary. Third, although we present a plausible interpretation for the negative impact of Treatment 2, we cannot definitively rule out alternative explanations. Fourth, the null findings around Treatment 3 might have been dampened by the presence of "private" social control that we document already existed among observers.

7. 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.

Aghion, P., & Tirole, J. (1997). Formal and real authority in organizations. Journal of political economy, 105(1), 1-29

Akerlof, George A. and Kranton, Rachel E. 2005 "Identity and the Economics of Organizations." Journal of Economic Perspectives, Vol. 19, 1: 9–32.

Alchian, A. A., Demsetz, H. 1972. Production, information costs, and economic organization. The American economic review, 62(5), 777-795

Allen, B., Lippner, G., Chen, Y. T., Fotouhi, B., Momeni, N., Yau, S. T., & Nowak, M. A. (2017). Evolutionary dynamics on any population structure. Nature, 544(7649), 227

Argyres, N. S., & Zenger, T. R. (2012). Capabilities, transaction costs, and firm boundaries. Organization Science, 23(6), 1643-1657

Axelrod, R. and Hamilton, W.D., 1981. The evolution of cooperation. science, 211(4489), pp.1390-1396.

Barcelo, H., Capraro, V. 2015 Group size effect on cooperation in one-shot social dilemmas. Scientific Reports, 5: 7937.

Barnard, C. 1938. *The functions of the executive*. Harvard University Press.

Bandiera, O., Barankay, I., & Rasul, I. 2005. Social preferences and the response to incentives: Evidence from personnel data. The Quarterly Journal of Economics, 120(3), 917-962

Bernhard, H., Fehr, E. and Fischbacher, U., 2006. Group affiliation and altruistic norm enforcement. American Economic Review, 96(2), pp.217-221.

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.

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.

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.

Capraro, V., Barcelo, H. (2015). Group size effect on cooperation in one-shot social dilemmas II: Curvilinear effect. PloS one, 10(7)

Carpenter, J. P. 2007. Punishing free-riders: How group size affects mutual monitoring and the provision of public goods. Games and Economic Behavior 60(1): 31-51.

Charness, G., Rigotti, L. and Rustichini, A., 2007. Individual behavior and group membership. American Economic Review, 97(4), pp.1340-1352.

Clement, J. and Puranam, P., 2017. Searching for structure: Formal organization design as a guide to network evolution. Management Science.

Colombo, M. G., Grilli, L. (2013). The creation of a middle-management level by entrepreneurial ventures: Testing economic theories of organizational design. Journal of Economics & Management Strategy, 22(2), 390-422

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.

Davila, A., Foster, G., Jia, N. (2010). Building sustainable high-growth startup companies: Management systems as an accelerator. California Management Review, 52(3), 79-105.

Fehr, E. 2018. Behavioral foundations of corporate culture. UBS Center Public paper #7.

Fehr, E., Gachter, S. (2000). Cooperation and punishment in public goods experiments, American Economic Review, 90(4), 980-994.

Garicano, L., & Wu, Y. (2012). Knowledge, communication, and organizational capabilities. Organization Science, 23(5), 1382-1397

Gibbons, R., 2006. What the folk theorem doesn't tell us. Industrial and Corporate Change, 15(2), pp.381-386.

Gibbons, R., Roberts, J. 2013 *Handbook of Organizational Economics*, Princeton University Press.

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? Handbook of Organizational Economics, Eds. R. Gibbons, J. Roberts. Princeton University Press.

Goette, Lorenz; Huffman, David and Meier, Stephan. 2006 "The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups." American Economic Review, 96(2), pp. 212–16.

Graham, J., Grennan, J., Campbell, H., Shivaram, R. 2018. Corporate Culture: Evidence from the Field. Working paper.

Grennan, J., 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.

Haan M, Kooreman P. 2002 Free riding and the provision of candy bars. Journal of Public Economics 83: 277–291

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.

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

Ioannidis, J.P., 2005. Why most published research findings are false. PLoS medicine, 2(8), p.e124.

Isaac, R. M., Walker, J. M., & Williams, A. W. 1994. Group size and the voluntary provision of public goods: Experimental evidence utilizing large groups. Journal of public Economics, 54(1), 1-36

Kandel, E., Lazear, E. P. (1992). Peer pressure and partnerships. Journal of political Economy, 100(4), 801-817

Khadjavi, M. 2016. "Indirect reciprocity and charitable giving—evidence from a field experiment." Management Science 63, no. 11: 3708-3717

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.

Loch, C.H. and Wu, Y., 2008. Social preferences and supply chain performance: An experimental study. Management Science, 54(11), pp.1835-1849.

Mas, A., and Moretti, E. (2009). Peers at work. American Economic Review, 99(1), 112-45.

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.

Nosenzo, D., Quercia, S., Sefton, M. (2015). Cooperation in small groups: the effect of group size. Experimental Economics, 18(1), 4-14.

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

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.

Olson, M. 1965 *The Logic of Collective Action: Public Goods and the Theory of Groups* HUP.

Organ, D.W., Podsakoff, P.M. and MacKenzie, S.B., 2005. Organizational citizenship behavior: Its nature, antecedents, and consequences. Sage Publications.

Ostrom, E. 2000. *Governing the Commons: The Evolution of Institutions for Collective Action*. New York: Cambridge University Press.

Pereda, M., Capraro, V., Sánchez, A. (2019). Group size effects and critical mass in public goods games. *Scientific reports*, 9(1), 1-10.

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., & Nowak, M. A. 2013. Human cooperation. Trends in cognitive sciences, 17(8), 413-425.

Raymond, E. (1999). The cathedral and the bazaar. Knowledge, Technology & Policy, 12(3), 23-49.

Rayo, L. 2007. Relational incentives and moral hazard in teams. The Review of Economic Studies, 74(3), 937-963

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

Suzuki, S., Akiyama, E. (2005). Reputation and the evolution of cooperation in sizable groups. Proceedings of the Royal Society B: Biological Sciences, 272(1570), 1373-1377

Suzuki, S., Akiyama, E. (2007). Evolution of indirect reciprocity in groups of various sizes and comparison with direct reciprocity. Journal of Theoretical Biology, 245(3), 539-552.

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, 1–39.

van Veelen, M., Garcia, J., Rand, D., Nowak, M. 2012. Direct reciprocity in structured populations PNAS, 109 (25) 9929-9934

Wichardt, P. C. (2008). Identity and why we cooperate with those we do. Journal of Economic Psychology, 29(2), 127-139

Yamagishi, T., Mifune, N. 2008 Does shared group membership promote altruism? Rationality and Society, 20 (2008), pp. 5-30

Yang, W., Liu, W., Viña, A., Tuanmu, M. N., He, G., Dietz, T., & Liu, J. (2013). Nonlinear effects of group size on collective action and resource outcomes. PNAS, 110(27), 10916-10921

Zhang, X. M., & Zhu, F. (2011). Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. American Economic Review, 101(4), 1601-15

Zelmer, J. Linear public goods experiments: A meta-analysis. Experimental Economics. 6, 299–310 (2003).

A. Appendix for online publication

A.1. Descriptive statistics of the DEKRA administrative data

In the **Table A-1** we compare the sample and the population. Except for year of start of BAPP –where the sample has newer projects, all the other variables are not statistically different at 95%. Categorical variables were tested using a chi-square test; the rest using a t-test.

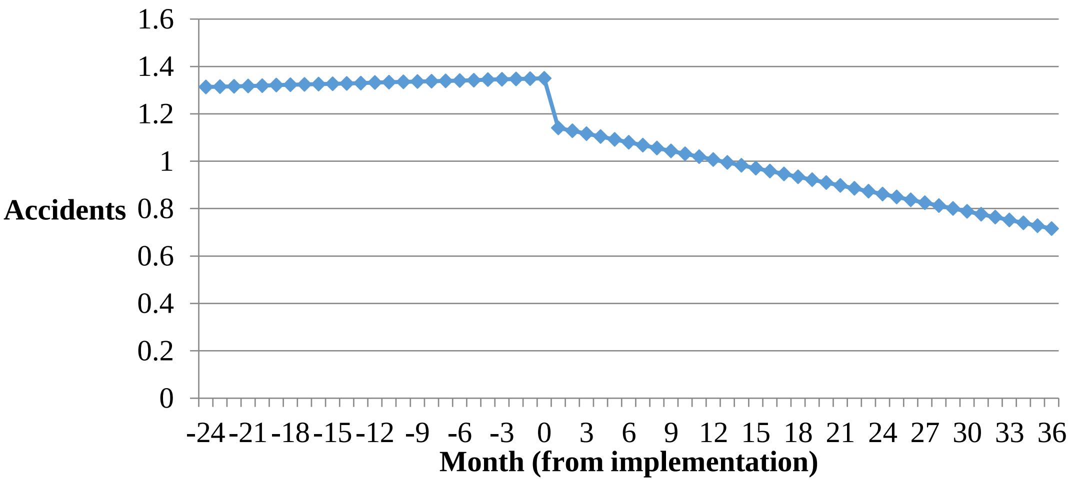
Table A-1. Comparison of population and sample of sites

	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

A.2. Identification of the impact of BAPP

In the **Figure A-1** we display the impact of BAPP using the column (3) of **Table 4-1**.

Figure A-1. Impact of BAPP over time



To probe on the causality of BAPP, we first do a flexible placebo test using the following model:

$ACCIDENTS_{it} = b_1 + \sum_j (\pi_j \times YEAR\_BAPP\_P_j \times BAPP\_P_{it}) + b_3 \times TREND_{it} + \sum_j (\rho_j \times YEAR\_BAPP\_P_j \times BAPP\_P_{it} \times TREND_{it}) + b_5 \times \ln(WORKERS_{it}) + U_i + ERROR_{it}$ (8)

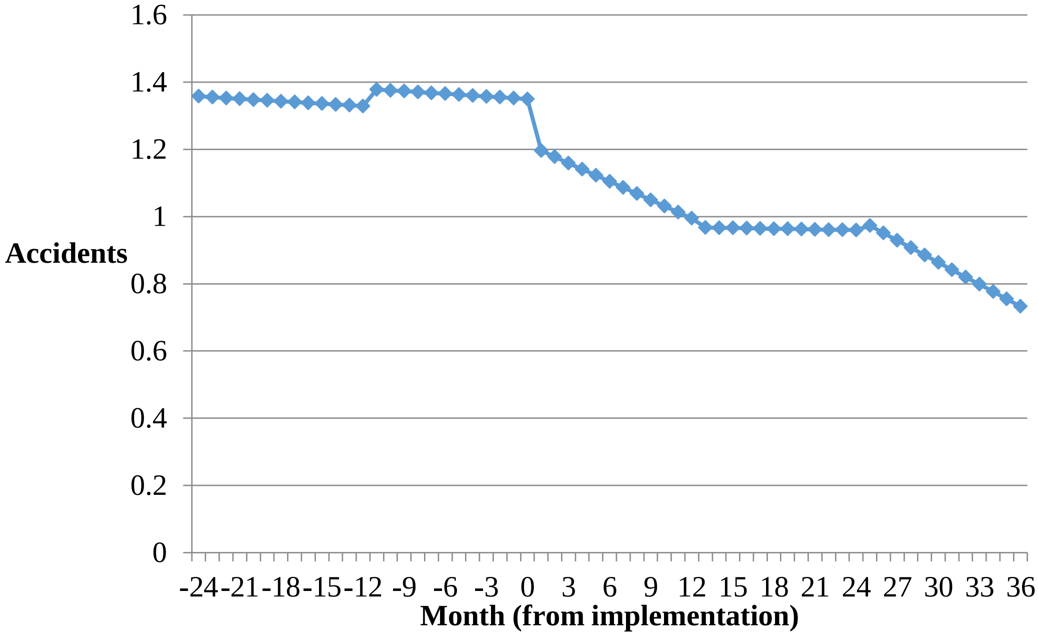
In this model, BAPP_P is the "placebo BAPP" and takes the value of 1 after the 12th month preceding the real start of BAPP (i.e., BAPP start in month -11). YEAR_BAPP is a dummy set that identifies the year preceding the real start of BAPP (from -11 to 0, where 0 is the month preceding the start of observations), 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 A-2** presents the estimates of equation 2. Interpreting this table can be tricky, so we graph the result in **Figure A-2**. This figure shows that there is no effect in the year before BAPP, neither at the level or slope.

Table A-2. 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 A-2. 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:

$$ACCIDENTS_{it} = b_1 + b_2 \times BAPP_{it} + b_i \times TREND_{it} + b_4 \times (BAPP_{it} \times TREND_{it}) + b_5 \times \ln(WORKERS_{it}) + U_i + ERROR_{it} \quad (9)$$

To estimate this model we use first differences and a fixed effect technique:

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

$$+ \Delta\, ERROR_{it} \quad (10)$$

The results are displayed in the **Table A-3**. 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) (as compared to **Table 4-1**). Statistical significance suffer in these models, as models in difference are noisier (see the r-square).

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 unobservables are tilted toward the cases that are not favorable to safety; for example, more extreme cases of low implementation quality than high. This resonates with intuition and with the distribution of contact rate in the **Figure 4-1** in the main body.

Table A-3. 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.				

Another way to assess the credibility of the estimates is to assess mechanisms. The mechanism we explore is how the pre-implementation culture of the site affects the impact of BAPP. As indicated above, DEKRA surveys the culture of the site (see the **Table A-4** in the online appendix), measuring 10 cultural in three buckets: organizational factors (i.e., relation between the firm and the workers), Teamwork factors (i.e., relations between workers), Safety factors (i.e., value of safety, communication of safety issues). In non-reported regressions on subsample of roughly 50 projects, 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, this suggest a substitution effect. BAPP operates by improving group relations and teaching workers how to approach co-workers. 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. Also, using a separate sample of 78 implementations, we find that BAPP is associated with a significant improvement in culture over time. BAPP improved directly the safety factors of the sites, which in turn improved organizational and teamwork factors.

A.3. Culture survey

Table A-4. 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

A.4. How contact rate affects the impact of BAPP

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

$$ACC_{it} = b_1 + b_2 \text{ x } BAPP_{it} + b_3 \text{ x } TREND_{it} + \sum_j b_{4j} \text{ x } BAPP_{it} \text{ x } QUINT\_CR_{jt} + b_5 \text{ x } \ln(WORKERS_{it}) + U_i + ERROR_{it} \quad (11)$$

In this model, QUINT_CR captures the quintiles of the contact rate, thus J = 5. In **Table A-5** we present the results. The joint t-test indicates that BAPP as a whole is significant. In column (2) we add as a control the interaction between BAPP and TREND, and the coefficients do not change.

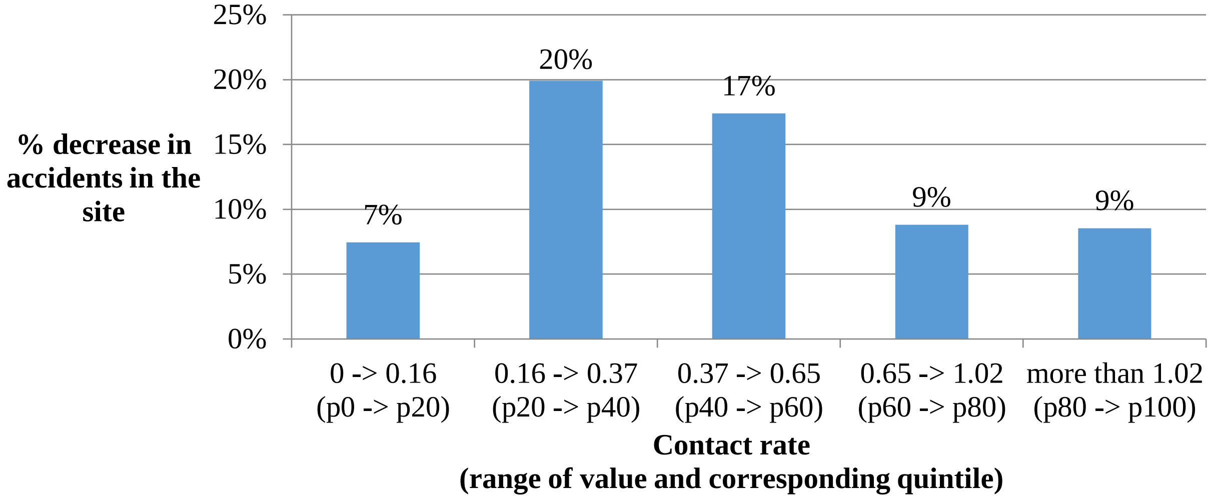
Table A-5. Role of contact rate on the impact of BAPP

	Accidents (1)	Accidents (2)
BAPP	-0.124† (0.191)	-0.123‡ (0.192)
BAPP X 1^{ST} QUINTILE OF CONTACT RATE	0.018† (0.145)	-0.029‡ (0.142)
BAPP X 2^{ND} QUINTILE OF CONTACT RATE	-0.155† (0.175)	-0.188‡ (0.175)
BAPP X 3^{RD} QUINTILE OF CONTACT RATE	-0.125† (0.125)	-0.148‡ (0.121)
BAPP X 4^{TH} QUINTILE OF CONTACT RATE	(Omitted)	(Omitted)
BAPP X 5^{TH} QUINTILE OF CONTACT RATE	-0.004† (0.109)	-0.020‡ (0.106)
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 fifth quintile kept, then the interaction with the second and third quintile would display p-values of 0.014 and 0.034; the same is true for column (2).

Figure A-3 displays the non-linear impact of contract rate 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.

Figure A-3. The impact of BAPP varies according to contact rate



A.5. Cohorts of observers

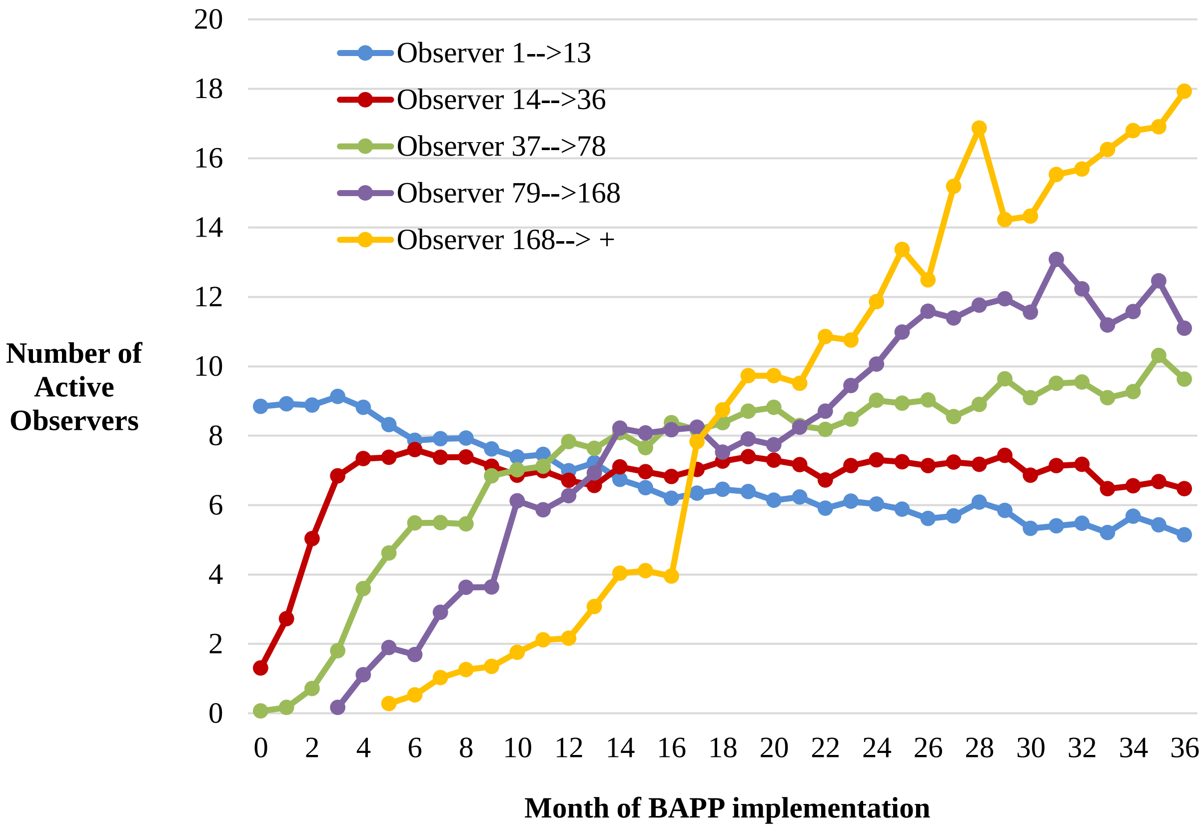
To generate the cut-offs of the quintiles/cohorts, we use the information at the observer-month level[26]. For example, at the period 12 there are, on average, 30 active observers per site, coming from the following cohorts:

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),

This data suggests that rotation of observers increases with the cohorts. At the 12th month, the first and second quintile have roughly 7 active observers but the pool of the former is much smaller, 13 observers compared to 23 (36-14+1). The same happens as we move further up. This suggests that newer observers might be leaving BAPP at a quicker rate than first cohorts. Cooperation seems to turn shakier with size.

[26] 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. 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.

Figure A-4. Number of observers per quintile of entry (or cohort)



A.6. Impact of newer observers on effort and rotation

We use the following regression model:

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

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 (as defined in the main body of the manuscript), the number of observers in the site (which captures diffusion), 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 could not add observer fixed effects as the cohort of the observer is time invariant. The results are displayed in the **Table A-6**. The column (1) show that the detrimental impact of higher cohorts of entry is robust to the control variables we used. However, sites have different number of workers, and therefore, using quintiles that are defined across sites (and not within) is inexact. To accommodate this, in columns (2) and (3) we use the order of entry of the observer to the site, 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 50th observer in entry order within a site displays 0.95 less observations, whereas the 100th observer displays 1.8 less observations.

Table A-6. Regression of effort on entry order

	Effort (1)	Effort (2)	Effort (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.			

In **Table A-7** we use observer tenure as the dependent variable. Here 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.

Table A-7. 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.		

A.7. Diffusion suffers with size of the site

Table A-8. Impact of site size on effort and diffusion

	EFFORT (1)	DIFFUSION (2)
DIFFUSION	0.196 (2.771)	
EFFORT		0.000 (0.000)
Ln(WORKERS)	2.789 (1.88)	-0.166*** (0.043)
Month of implementation fixed effect?	Yes	Yes
Site fixed-effect?	Yes	Yes
Constant	-4.569*** (1.690)	-4.569*** (1.692)
Adjusted R-square	17.71%	63.09%
Observations	2,696	2,696
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. The sample is restricted to the period of BAPP implementation.		

A.8. The impact of effort is not decreasing on diffusion

Table A-9. Interaction effect of effort and diffusion

	Accidents (1)	Accidents (2)
BAPP	-0.163 (0.117)	-0.160 (0.119)
BAPP X HIGH EFFORT	-0.186** (0.089)	0.194* (0.103)
BAPP X HIGH DIFFUSION	0.169* (0.094)	0.160 (0.121)
BAPP X HIGH EFFORT X HIGH DIFFUSION		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.		

A.9. Impact of specialization

Figure A-5. Distribution of specialization

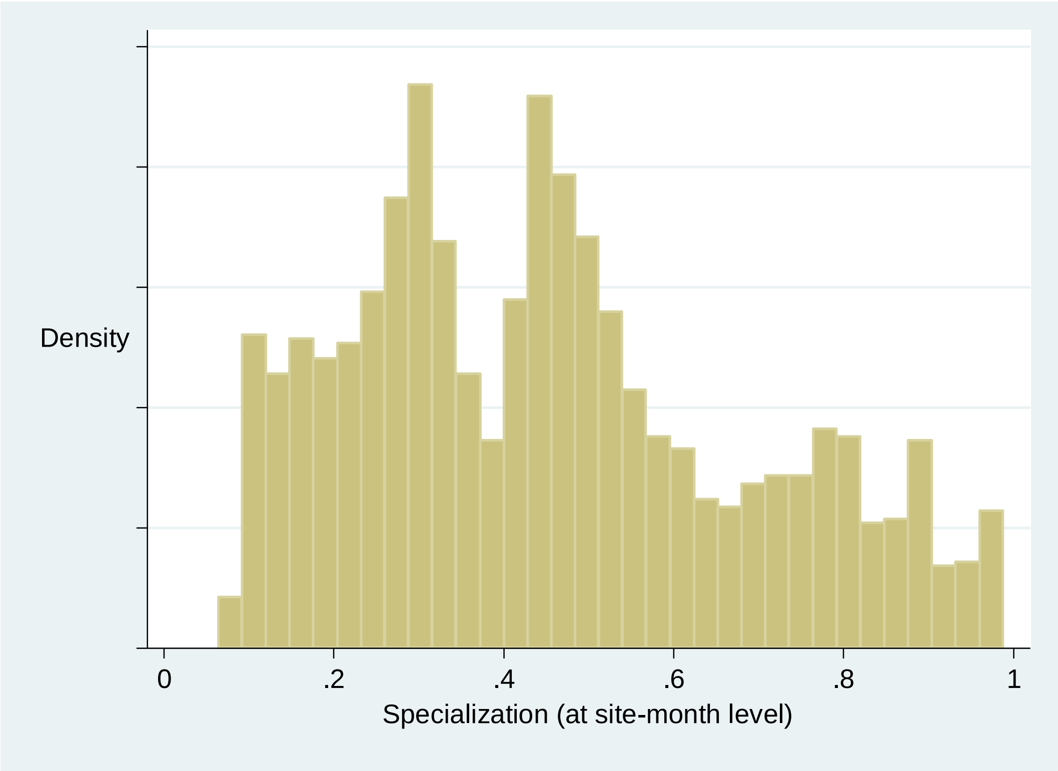


Table A-10. The role of specialization on the impact of BAPP

	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 SPECIALIZATION	-0.649** (0.212)	-0.655** (0.291)
BAPP x HIGH_EFFORT	-0.283*** (0.106)	-0.283*** (0.106)
BAPP x HIGH_DIFFUSION	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. High effort and High diffusion are dummies that use the 50^{th} percentile as cutoff. Tenure is measured as the number of months elapsed since the observer's first observation. Experience is measured using the cumulative number of observations of the observer up to month t-1. Tenure and Experience and then averaged to obtain a site level variable.

A.10. Impact of order of entry of observers in quality of observations

Table A-11. Impact of order of entry on the quality of information in observation sheets

Dependent variable:	Total SLV (per sheet)	Total Risk & Safes (per sheet)	Total Flags (per sheet)
Average	2.5	11.1	8.5
Order of entry	0.01 (0.2%)	-0.17 (-1.5%)	0.01 (0.1%)
Tenure as observer	0.01 (0.2%)	-0.16 (-1.4%)	0.07 (0.8%)
Specialization as observer	0.01 (0.2%)	0.14 (1.3%)	-0.18 (-2.1%)
Accumulated Experience	-0.01 (-0.5%)	-0.12 (-1.1%)	-0.10 (-1.2%)
Observations	353,637	348,439	123,962

Notes:
(1) No coefficient is significant in these regressions. The coefficients are presented as the impact of going from p25 to p75 in the respective dependent variable. In parentheses we represent the proportion of the effect with respect to average of the dependent variable (second row of the table).
(2) Dependent variables: "Sheet level variables" (SLV) is the heading of the sheet that contains typically around 5 to 7 fields, such as name of observer, date, time, place within the site, presence of a coach, etc. "Total Risks and Safes" is the number of CBI that the observer reported on in the sheet; this is the count of behaviors that either were deemed as executed in a safe or a risky way. "Total flags" records the quality of the written comment that the observer has to add to the observation for each risky behavior he reported; The comments should ideally comply with 8 criteria in the description of the behavior and the interaction and recommendation given by the observer (e.g., was intent discussed, did the observer provide a suggestion).
(3) Controls: Site specific month of implementation, Site unobservables, Year, Day of week, Day of Month, Sum of risks (only in the flags regression), Location within site.
(4) In the regression of total flags, we only use sheets that display at least one risk.
(5) The average of 8.5 Flags per sheet considers that every sheet with at least 1 risk has on average 1.5 risks. Thus, flags per risk is 5.5.

A.11. Letter handed out to workers

<u>Letter handed out under treatment 1</u>

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.

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]

<u>Letter handed out under treatment 2 (the areas highlighted in grey are added to the letter)</u>

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
…	…	…
…	…	…
…	…	…

A.12. Implementation details of treatments

Communication protocol. 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, which was sponsored by all three partners DEKRA, ACHS and SODIMAC. The same message was delivered to the enabler and the starting team of starting team observers, 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.

Treatment 1. First, in the 4th month of implementation, when the starting team was being trained to execute observations, 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). Randomization of observers and workers was done using a lottery box. Workers of the site had been pre-randomized and placed on lists that contained the names of the workers included in the treatment groups and the control group. These lists were prepared by the research team beforehand and sent to the consultant prior to his/her visit to the site. 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 explained randomization by indicating 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)[27]. In order to communicate to the workers in a treatment 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 online appendix A.10. 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. 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 A-12**). In these updates, the newly joining workers were randomly assigned to the groups or the control (again stratifying the assignment). The lists and letters were updated and distributed accordingly.

[27] 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.

Table A-12. 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 starting team observer appointed to it). The average includes the starting team observer.

A.13. Report used in treatment 3

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			
…			
…			

A.14. Balance and take-up

Table A-13. 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 A-14. Balance check of observer randomization

	Starting team members - All Stores			Starting team 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)

Table A-15. 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 respondents who acknowledge having received the letter.					

A.15. Impact of BAPP on accidents in Sodimac

We estimated the following model:

$$ACCIDENT_{ijt} = b_1 + b_2 \text{ x } BAPP_{ij} + b_3 \text{ x } BAPP_{ij} \text{ x } TIME\_ELAPSED_{ij} + b_4 \text{ x } OBS_{ijt} + X_{it} + \tau_t + \gamma_j + u_{ijt} \quad (15)$$

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 the analysis of probability of becoming observer. We control for month and store fixed effects to control for the common trend in accidents and store unobservables. Results do not change if we add worker fixed effects. We do not include them because rotation is 5% a month, and therefore, if we had included them, we would be measuring the impact only a subset of workers that are present before and after and not the whole population subject to BAPP. 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 behavior of observers. It could be that all the impact of BAPP on accidents is exerted through lower accidents of observers and not the general workforce. We estimate this model using the four sites of our experiment between January 2016 to May-2018, and we consider only workers that are subject of BAPP observations.

Table A-16. Impact of BAPP on accidents in Sodimac

Panel a)	Total accidents	Workplace accidents	Workplace accidents without lost working days	Workplace accidents with lost working days

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
BAPP	-0.0022 (0.0036)	-0.0022 (0.0036)	0.0000 (0.0023)	-0.0000 (0.0023)	-0.0014 (0.0019)	-0.0015 (0.0019)	0.0015 (0.0012)	0.0015 (0.0012)
BAPP x Time elapsed	-0.0016* (0.0008)	-0.0016* (0.0008)	-0.0015*** (0.0006)	-0.0015*** (0.0006)	-0.0011*** (0.0004)	-0.0011*** (0.0004)	-0.0004 (0.0003)	-0.0004 (0.0003)
Observer		-0.0007 (0.0031)		-0.0004 (0.002)		0.0011 (0.0019)		-0.0014*** (0.0004)
Ind. level Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Store FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	30,193	30,193	30,193	30,193	30,193	30,193	30,193	30,193
R-squared	0.0042	0.0042	0.0037	0.004	0.0025	0.0025	0.0018	0.0019
Mean	0.0094	0.0094	0.0043	0.0043	0.0023	0.0023	0.0020	0.0020
Panel b)	Commuting accidents		Quasi-accidents		Length of leave	Length of leave		
	(1)	(2)	(3)	(4)	(5)	(6)		
BAPP	0.00013 (0.019)	0.0001 (0.0019)	-0.0019 (0.0021)	-0.0018 (0.0021)	0.039 (0.036)	0.040 (0.036)		
BAPP x Time elapsed	0.0002 (0.0004)	0.0002 (0.0004)	-0.0004 (0.0005)	-0.0004 (0.0006)	0.001 (0.014)	0.001 (0.015)		
Observer		0.0008 (0.0019)		-0.0013 (0.0014)		-0.030 (0.027)		
Accident with lost time					13.382*** (2.905)	13.382*** (2.905)		
Ind. level Controls	Yes	Yes	Yes	Yes	Yes	Yes		
Store FE	Yes	Yes	Yes	Yes	Yes	Yes		
Month FE	Yes	Yes	Yes	Yes	Yes	Yes		
Observations	30,193	30,193	30,193	30,193	30,193	30,193		
R-squared	0.0013	0.0013	0.0029	0.0029	0.161	0.161		
Mean	0.0018	0.0018	0.0033	0.0033	0.049 (13.4)	0.049 (13.4)		

OLS regressions. Results are consistent if we use count models. Errors in parentheses: Robust and clustered at the worker level. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

A.16. Identity of coaches

Table A-17. Identity of coaches

	Number	Actual execution of coaching. Mean (St. dev.)	Theoretical benchmark of random coaching	Is the actual execution different then the benchmark? (p-value)
Panel a. Only for the coached observers of the treatment groups				
Percentage of coaching that was done by a member of the group	95	0.063 (0.245)	0.1	0.145
Percentage of coaching that was done by a member of the group (excluding coaching by the enabler)	72	0.083 (0.278)	0.1	0.613
Panel b. Only for the coached observers of the control group				

Percentage of coaching that was done by a member of the group	92	0.696 (0.462)	0.5	0.001***
Percentage of coaching that was done by a member of the group (excluding coaching by the enabler)	54	0.481 (0.504)	0.5	0.788
Notes: * p<0.1, ** p<0.05,*** p<0.01.				

We had 213 coaching events on new observers. We excluded 26 that were mainly done by consultants, leaving 187 coaching events. Out of these, in 95 cases the coached observer was a new observer that was part of the treatment (panel a), and in 92 it was part of the control group (panel b). For the first group, we computed a variable that took the value of 1 if the coaching event was executed by another observer of its treatment 1 group (and zero otherwise). For the second group, we computed a variable that took the value of 1 if the coaching event was executed by another observer of the control group or the enabler (and zero otherwise). The enablers executed plenty of coaching, 62 in total. To assess its impact we assigned them to the control group and then analyze the results with and without its inclusion. In panel a) we find that 6.3% and 8.2% of the coaching events (with and without the enabler, respectively) had a coach that was an observer of its own treatment group. Theoretically, if coaching was executed randomly, then the expected value for this percentage is roughly 10%. Either including or excluding the enabler, we cannot reject the hypothesis that the selection of the coached observer was done randomly. In panel b) the benchmark is 50%, as half of sites is assigned to control. Here we find that 48% of the coaching events (excluding the enabler), were done by another observer of the control group. (If we had included the enabler, the number goes artificially up, as it goes down artificially down in panel a). Again, we cannot reject the null that coaching was done randomly.

A.17. Difference between starting team observers, new observers and the rest of workers

Table A-18. 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). Starting team 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 starting team 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 starting team members that were supervisors. The result do not change if we include these back.

Table A-19. Difference between starting team members and new observers

	Observers members of the starting team 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	

*** 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).