# Reputation, Reciprocity and Large-Scale Cooperation: Evidence from the Field

Francisco Brahm*

Christoph Loch‡

Cristina Riquelme §

April 2024

**Abstract**

It is widely accepted that large-scale cooperation is crucial for organizational performance but is hard to attain. We study two drivers that harness contingent cooperation: reputation (I cooperate with you if I can tell that you have cooperated with others) and reciprocity (I cooperate with you if you also do so—which may occur as an equilibrium outcome among selfish agents, that is, instrumental reciprocity, or due to a social preference for "friends," that is, intrinsic reciprocity). We study workers volunteering to help colleagues avoid workplace accidents and document that (1) cooperation (in the form of volunteering and helping efforts) reduces accidents; (2) reputational benefits weaken when there are many volunteers, so it does not support large-scale cooperation; (3) modifying the formal structure of interactions among workers (via a pre-registered field experiment), so that reciprocity is boosted, effectively supports large-scale cooperation; (4) the latter result seems to be driven by both intrinsic and instrumental reciprocity. Our results contradict the received wisdom of reputation being a more effective driver of large-scale cooperation than reciprocity.

*Keywords*: Cooperation, Reciprocity, Reputation, Field experiment.
*JEL*: C93, D90, J28, M50, M10, M20

---

Acknowledgments: We are grateful for the comments offered by Bart Vanneste, Vincent Mak, Dmitry Sharapov, Jerker Denrell, Robert Gibbons, Sendil Ethiraj, Isabel Fernández-Mateo, Arianna Marchetti; participants at seminars in the London Business School, Rotterdam School of Management, LMU Org Seminar and the Business Economics Department at Pompeu Fabra University; and at the following conferences: London 50 Conference, the 2020 Berkeley Haas Culture Conference, Bocconi 2022 Assembly for Innovation and Cooperation (BAIC), the 2020 and 2023 Strategy Science Conference and the 2023 AOM annual conference (the usual disclaimers apply).

*London Business School. Email: fbrahm@london.edu
‡Cambridge Judge Business School, University of Cambridge. Email: c.loch@jbs.cam.ac.uk
§Department of Economics, University of Maryland, College Park. Email: riquelme@umd.edu

# 1 Introduction

Achieving and sustaining cooperation in large groups—exerting effort for the benefit of others and the group, even against private interest—is an essential ingredient in the success of organizations (Milgrom and Roberts 1995, Gibbons and Roberts 2013, Alchian and Demsetz 1972, Olson 1965, Fehr 2018, Podsakoff et al. 2009, Organ et al. 2005). However, large organizations find it hard to achieve: in a survey of CFOs/CEOs in 1348 large US Firms (Graham et al. 2022), 92% of responses indicate that cooperation among workers is the main antecedent to an effective culture, but only 16% believe their culture is where it should be.

Two conditions challenge large-scale cooperation. First, cooperation often involves a social dilemma: Individuals are tempted to free-ride on other group members' cooperation. The temptation to free-ride increases with the size of the group (Alchian and Demsetz 1972, Holmstrom 1982, Olson 1965). Group pay-for-performance, which may be used as an antidote, is nevertheless subject to free-riding issues as groups grow (Holmstrom 1982) and "unlike for the case of individual incentives, the jury on the effectiveness of team incentives is still out" (Friebel et al. 2017, p. 2169). Second, while cooperation can be enforced using formal managerial levers such as monitoring, job descriptions, formal control, or monetary incentives, these levers usually promote perfunctory cooperation (i.e., "to-the-letter" compliant effort) and not consummate cooperation (i.e., "above and beyond" voluntary effort) (Gibbons and Henderson 2012, Gibbons and Roberts 2013, Organ et al. 2005)[1]. Given that these managerial and organizational devices tend to lose efficacy with size (Zenger 1994, Rasmusen and Zenger 1990, Williamson 1967, Brahm and Tarziján 2012), consummate voluntary cooperation becomes increasingly important with size.

Consequently, firms need to rely on aspects of the informal organization to support large-

[1]For example, organizational citizenship behavior, which at its core deals with cooperative behavior, is "about the types of discretionary behavior and contributions that are not explicitly associated with specific job requirements [such as] job descriptions, specifications of employee rights and responsibilities, contracts, performance appraisals, and incentive plans" (Organ 2018, p. 7).

scale cooperation. Some of these "soft" aspects that research has explored are the role of leaders as guides and enforcers (Barnard 1968, Schein 2010, Kosfeld and Rustagi 2015, Hermalin 2013, Gartenberg and Zenger 2023); the identification of workers with the organization (Akerlof and Kranton 2005); punishment among peers (Fehr and Gächter 2000, Kandel and Lazear 1992); and corporate culture and purpose (Grennan 2020, Gartenberg et al. 2019).

In this paper, we focus on two crucial informal organization levers—reciprocity and reputation. These two levers harness contingent cooperation; that is, an individual's cooperation depends on whether others in the group behave cooperatively. Many fields recognize and study reciprocity and reputation as drivers of cooperation (Henrich and Muthukrishna 2021, Raihani 2021, Nowak 2006, Nowak and Sigmund 1998, 2005, Gächter and Herrmann 2009, Dal Bó and Fréchette 2018, Sobel 2005, Gouldner 1960, Emerson 1976, Cropanzano and Mitchell 2005, Sugden 1984, Bohnet and Huck 2004).[2] However, their relation to large-scale cooperation remains actively debated (Boyd and Richerson 1989, Allen et al. 2017, Van Veelen et al. 2012, Gächter and Herrmann 2009, Hauser et al. 2016, Raihani 2021, Henrich and Muthukrishna 2021, Boyd and Richerson 2005). We contribute to this debate by showing that i) against received wisdom (Raihani 2021, Henrich and Muthukrishna 2021), but consistent with recent research studying the vulnerabilities of reputation (Számadó et al. 2021, Giardini et al. 2022), reputation fails to sustain large-scale cooperation because benefits of reputation dwindle with scale, and ii) against received wisdom (Boyd and Richerson 1988, Raihani 2021) but consistent with a recent theoretical model (Allen et al. 2017), modifying the interaction structure of the group—a formal organizational lever—so that reciprocity is boosted in dyads, promotes large-scale cooperation.

Reciprocity refers to in-kind responses to interacting parties, that is, cooperating (defecting) in response to someone else's cooperation (defection) (Trivers (1971), Axelrod and

[2]While reputation is most often emerging out of informal interactions, sometimes it can be formally promoted–by those parties with authority/rights to do so–through "reputation systems", that is, systems that track and display past behavior with associated payoffs (e.g., stars in Uber) (Tadelis 2016)

Hamilton (1981), Nowak and Sigmund (1992), Sobel (2005), Rabin (1993), Dohmen et al. (2009)). This may have two motivations. Firstly, a "social preference" for the other person's well-being based on the other person's kind acts toward the focal person in the past (Bolton and Ockenfels (2000), Rabin (1993), Cabral et al. (2014), Sobel (2005) call this "intrinsic reciprocity"). Intrinsic reciprocity is an emotional "shortcut" (Trivers 1971, Uzzi 1997). Secondly, an "instrumental" reciprocity in expectation of valuable reciprocation from the other person in the future that more than compensates for the short-term payoff of defection (Cabral et al. 2014, Sobel 2005). Instrumental reciprocity involves calculation and doesn't require non-standard preferences.

Reputation refers to individual A cooperating with individual B only if A can tell that B has cooperated with others in the past (Ohtsuki and Iwasa 2006, Nowak and Sigmund 1998, Boyd and Richerson 1989). (Reputation is also referred to as "indirect reciprocity", usually in Biology and Evolutionary Anthropology.) While extensive lab evidence exists for reputation as a mechanism that supports cooperation (Kraft-Todd et al. 2015, Rand and Nowak 2013), as well as evidence for market-based interactions (Tadelis 2016, Elfenbein et al. 2015, van Apeldoorn and Schram 2016, Khadjavi 2017, Greif 1993, MacLeod 2007) and among community members (Ge et al. 2019, Yoeli et al. 2013), field evidence on the role of reputation in fostering cooperation inside organizations among its members is absent, as far as we can tell.

Despite their capacity to foster cooperation, both reciprocity and reputation may have problems sustaining large-scale cooperation. Received wisdom from evolutionary studies indicates that reciprocity becomes less effective in supporting cooperation in larger groups (Henrich and Muthukrishna 2021, Raihani 2021): when many individuals are attempting to cooperate, for example, in a public goods situation, reciprocity—not cooperating if some in the group don't—is too harsh as a dissuasive tool: cooperation withdrawal in response to a fraction of other players defecting hurts all group members, including those that cooperated,

and this can produce a downward spiral: cooperative people would be interpreted by others as selfish, and this would trigger more defection, more misjudgment of cooperative types as selfish, and so on (Boyd and Richerson 1988, Fischbacher et al. 2001, Hergueux et al. 2023)[3]. Without some punishment that is personally directed to defectors, large-scale cooperation is frail. Notwithstanding, a recent theoretical model suggests that reciprocity may be "rescued": Allen et al. (2017) study populations of agents interacting in prisoners' dilemma and shows that given *any* interaction network structure that a population may have, the best way to modify the structure to achieve widespread cooperation is to infuse the structure with strong pairwise ties that foster reciprocity.

Reputation is predicted to be much better suited to supporting large-scale cooperation (Henrich and Muthukrishna 2021, Raihani 2021). Evidence seems supportive (Yoeli et al. 2013, Ge et al. 2019, Greif 1993). However, in comparison to reciprocity, reputation involves more subtle and complicated machinery (Számadó et al. 2021, Giardini et al. 2022, MacLeod 2007): it requires a norm, a commonly agreed rule to translate behavior (concerning the norm) into a reputation, and the communication of that reputation. Furthermore, this cultural machinery co-evolves with other elements of culture and is, therefore, heterogeneous across groups (Henrich and Muthukrishna 2021). Recent research has identified limiting problems stemming from these complications, such as the dynamics of gossip (Wu et al. 2016), the complex rules that translate norm-related behavior into reputation (Santos et al. 2021), signals being affected by social context (Dumas et al. 2021), inequity aversion may limit the impact that the observability of one's action has (Bolton et al. 2021), old defection records might inefficiently trap some agents in defection cycles (Kamei and Putterman 2017) and players struggling to keep track of the reputation of many players for all periods and therefore relying just on the recent past (Baker and Bulkley 2014). These problems can limit

[3]Another problem is cyclical behavior: while direct reciprocity is a good catalyzer of cooperation, once it invades a group it can be displaced by cheaper unconditional cooperation strategies, which in turn are displaced by uncooperative strategy, leaving the group back at the starting point (Van Veelen et al. 2012).

the capacity of reputation to support large-scale cooperation inside organizations.

Our paper studies these issues—how reputation and reciprocity affect large-scale cooperation—in the field. We examine a workplace safety methodology implemented at sites such as plants, warehouses, or stores. It relies on workers' cooperation, particularly volunteering, to help coworkers conduct their tasks safely. Moreover, the group of volunteers is challenged to grow from a small size to a significant fraction of the organization's workforce; on average, sites have approximately 250 workers, and volunteers go from 10 at the start to 50 or more after a couple of years. Using a sample of 88 sites where this methodology was applied (comprising more than 1.2 million instances of a volunteer helping a worker), we show that volunteering and providing consummate helping effort reduces the incidence of workplace accidents; we also show that it is individually costly and thus, workers face free-riding incentives. Then, using the same dataset and guided by a straightforward formal framework incorporating reputational concerns, we document that cooperation weakens as more workers volunteer. This occurs because the reputational benefits a volunteer earns weaken as more workers volunteer. This result joins the research uncovering the limitations of reputation (Számadó et al. 2021, Giardini et al. 2022, Santos et al. 2021, MacLeod 2007).

Inspired by Allen et al. (2017), we use a pre-registered field experiment to show that reciprocity can sustain large-scale cooperation if supported by an interaction structure that promotes frequent interactions within the volunteer-worker dyad (i.e., how frequently a worker receives help from the same volunteer). This effect entirely counteracts the decay in cooperation and safety performance from weakening reputation benefits. Several tests support the reciprocity logic within volunteer-worker dyads. Our analysis suggests that both intrinsic and instrumental reciprocity are driving this result.

The rest of the paper is organized as follows: We describe the setting, then introduce our formal model and the predictions it generates. We then describe the data and field experiment, presenting our empirical analysis and results. In the final section, we provide a

discussion about the implications to the literature.

## 2 Setting

To measure and manipulate cooperation in a growing group, we collaborated with DEKRA Insight, a global company specializing in workplace safety services. One of its services is BAPP (Behavioral Accident Prevention Process). This methodology uses coworker feedback to improve workplace safety for employees at a site (such as a plant, a store, or a warehouse). The BAPP methodology starts with two months of assessment and establishing a team of 8 to 12 employees. The selection of employees does not follow predefined criteria other than focusing on front-line workers (supervisors or managers are not eligible) and is voluntary. One team member is chosen by consensus as the BAPP "enabler" and focuses 100% on the project, reporting directly to the site manager. The other members execute BAPP-related tasks in addition to their regular jobs. During implementation, the enabler and the team of volunteers meet monthly in the "BAPP committee" to monitor and manage progress. In the third and fourth months, the enabler and volunteers receive training on executing "observations" and becoming "observers."

An observation consists of approaching a worker, observing their behavior for 10 to 20 minutes with their consent, and filling out a detailed, itemized observation sheet. This sheet contains general information (e.g., the date, site area, and time of day) and a list of site-specific critical behaviors (e.g., driving a forklift or working at height) marked as being performed in a safe or risky manner. If a risky behavior is identified, the worker is given verbal feedback. Only front-line workers are observed—BAPP is a method "by the workers for the workers." BAPP does not establish predefined criteria about who observes whom, and the observed workers' identities are never recorded. This is captured by the motto, "No spying, no naming, no blaming" (diminishing workers' potential suspicions of being controlled

by management).

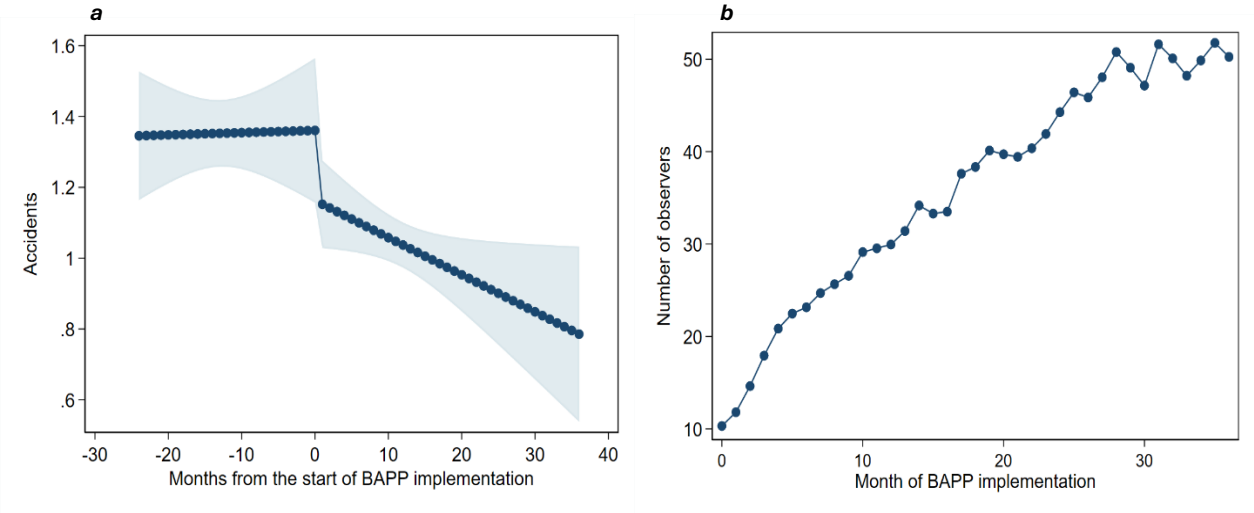
Further, BAPP tends to promote not repeating observations with the same set of workers to avoid a "blind eye" (or getting used to tasks or a set of workers and stopping spotting/calling on unsafe behavior). Crucial to our study, the initial observers are trained to enroll and train additional workers as observers in the fifth month. From month six, the enabler and observers execute observations and seek to expand the number of observers; again, enrollment is voluntary and limited to front-line workers. The new observers do not participate in the monthly progress meetings (the BAPP committee). In the twelfth month, the consultant performs a sustainability review of the program, after which the site employees are left to use their own devices.

BAPP involves two main instances of cooperation: first, a worker's decision to become (and remain) an observer (i.e., becoming a cooperator); second, an observer's decision about how many observations to execute in a month (i.e., how much cooperative effort to exert). Becoming an observer carries costs (such as time devoted to training and executing observations without any formal compensation). Still, it may also have informal benefits, mainly reputation (e.g., help from workers on other tasks or good standing with managers that may lead to future benefits). According to DEKRA consultants, these reputation benefits are crucial for workers adopting BAPP. A social dilemma may arise nonetheless if these reputation benefits are meager and the benefit to workers is larger from observations than the cost to observers. The second cooperation instance is how many observations to execute (which we label "effort"). A social dilemma also arises: everyone would be better off with high effort, but if the reputation benefits from exerting effort are insufficient, each observer is incentivized to free-ride on others' effort.

As the number of observers grows during BAPP implementation, we can explore *how these cooperation choices are affected by the number of observers*, that is, by the scale of cooperation. Thus, BAPP allows us to dissect the anatomy of cooperation as it expands.

Using econometric analysis of archival data from DEKRA (see subsection A.2.3), panel a of Figure 1 shows that BAPP effectively reduces accidents (the subsection A.2.3 provides evidence that this impact is causal). Panel b shows that the number of active observers grows to approximately fifty after three years (see the subsection A.2.1 for additional summary statistics).

Figure 1: Dynamics in BAPP implementation




**Notes: a.** The vertical axis represents accidents per site per month and the horizontal axis time since the start of BAPP. We plot predicted accidents (dark blue connected dots) based on a before-and-after regression model that includes site-fixed effects (see subsection A.2.3). We use our sample of 88 BAPP implementations (see subsection 4.2). The shaded area represents a 95% confidence interval. The graph shows that BAPP significantly reduces accidents. In the baseline econometric specification, this impact includes a step change at the start of implementation plus a downward trend afterward; other specifications yield similar results (see subsection A.2.3). **b.** Using the sample of 88 BAPP implementations, we plot the average number of observers per site across implementation months. Starting from the initial team of 10, BAPP observers grow to 50 in number, representing a 20% volunteering rate in the average site of 250 workers (see subsection A.2.1).

# 3 Theory and Hypotheses With a Simple Model

This section introduces the simplest possible model (sufficiently simple not to require formal propositions) customized to our setting. The model does not need to derive non-obvious optimal behaviors, as these are recognizable by verbal reasoning; nonetheless, it helps us to be transparent in our arguments and to identify which parameters drive behavior and need to be empirically estimated to explain i) the choice of becoming an observer, ii) the choice of the number of observations and, iii) safety outcomes.

## 3.1 Model Setup

The (monthly) payoff when not becoming an observer consists of receiving observations without performing them. Contact rate (the number of observations per worker) is the product of $\bar{b}$ (the average effort, or number of observations, per observer) and diffusion $d$ (the number of observers per worker) (see subsection A.2.1 for descriptive statistics of these variables). The contact rate drives the payoff (Figure 1a shows that this payoff, in the form of reduced worker accidents, is real and substantial).

$$\text{Payoff when not becoming observer} = \text{Contact rate} = \bar{b}d. \tag{1}$$

The payoff of becoming an observer is equal to the same contact rate, boosted by learning $p$ from being an observer, plus a reputational benefit $r(\cdot)$ and a reciprocity benefit $f(\cdot)$, minus the effort costs $c_0$ and $c_T$:

$$\text{Payoff of becoming the } k^{th} \text{ observer} = (1+p)\bar{b}d + r(b, \hat{d}_k) + f\left(\frac{b}{\widetilde{n}}\right) - c_0 b - c_T. \tag{2}$$

Here, $b$ is the focal worker's number of observations performed in the month. Given that observers are also observed, the parameter $p$ captures the additional rate at which observers

can learn from being observed (it is known and advertised in BAPP that observers enjoy an extra accident reduction benefit; in subsection A.2.4, we show that $p$ is statistically different from 0). $c_0$ is the effort cost per observation (observers are not paid and need to perform observations on top of their work duties), and $c_T$ is the fixed cost of training. These two costs typically represent together about 5% of a worker's time. We now turn to functions $r(\cdot)$ and $f(\cdot)$.

The function $r(\cdot)$ specifies the reputation benefit of being an observer (e.g., status, help received from colleagues, and the increased likelihood of future promotions) that the $k^{th}$ observer at the site enjoys. The variable $\hat{d}_k$ is the diffusion at the moment the worker becomes the $k^{th}$ observer at the site; for example, the $10^{th}$ observer receives reputational benefits that correspond to this "rank" of having joined, and this benefit is stable over time even when diffusion increases later. We assume that $\frac{\partial r}{\partial \hat{d}_k} < 0$ (we test this and other assumptions about $r(\cdot)$ below; see subsection 5.1). This means that managers and colleagues can roughly observe and recall the order in which workers joined the observer team and that they attribute more reputational benefits to observers who joined earlier than to those who joined later. Observers and BAPP consultants acknowledge that cooperating initially, when BAPP is still a risky proposition (as not all implementations succeed), represents a more credible signal of cooperative spirit to managers and coworkers and is better rewarded. The function $r(\cdot)$ is increasing and concave in $b$ ($\frac{\partial r}{\partial b} > 0$ and $\frac{\partial^2 r}{\partial b^2} < 0$), indicating that more observations boost the reputational benefit of a particular observer at a decreasing rate. The reputation increase from more observations is less pronounced for later joining observers ($\frac{\partial^2 r}{\partial b \partial \hat{d}_k} < 0$); as above, cooperative effort is perceived as more indicative of true cooperative spirit when it is early and risky.

Following Ashraf and Bandiera (2018), $f(\cdot)$ captures social incentives, that is, "any factor that affects the marginal benefit or cost of effort and that stems from interaction with others" (ibid, p. 440). In our case, we assume that workers obtain personal/private satisfaction from

helping others via observations. However, this satisfaction strongly depends on whether the observations are reciprocated by workers exerting effort in complying with the safety advice. Complying with an observer's advice is costly for workers—for example, changing the work routine/script or taking extra steps—and thus, it is not obvious that they will follow it. The personal satisfaction from helping others is captured by $f(\cdot)$ being increasing on the number of observations $b$ ($\frac{\partial f}{\partial b} > 0$); the contingency on reciprocal behavior by workers is captured by the number of workers that the observer targets to execute its observations, captured by $\widetilde{n}$. If this number is very large, then the observer is not repeating observations often with its targeted workers; in contrast, if $\widetilde{n}$ is small, then the observer repeats observations often, and this would trigger a relationship between observer and worker where both reciprocate each other's effort towards helping the other (i.e., intrinsic reciprocity); alternatively, a low $\widetilde{n}$ also means that the expectation of future interactions is high, and this would trigger effort by selfish observers and workers to sustain a profitable exchange (i.e., instrumental reciprocity) (Sobel 2005). To see this more clearly, notice that the term $\frac{b}{\widetilde{n}}$ represents the likelihood of observing the same worker in subsequent months; if $b = 3$ and $\widetilde{n} = 3$, then a worker is certain to be observed each month by the same observer. It is clear that $f(\cdot)$ decreases with $\widetilde{n}$ ($\frac{\partial f}{\partial \widetilde{n}} < 0$). Given that BAPP promotes random observation across the whole site to avoid the "blind eye" (see section 2), we assume for now that $\widetilde{n}$ is equal to the size of the site ($\widetilde{n} = n$), and thus benefit from reciprocity are small. In subsection 3.5, we will lift this assumption to address how changing the population structure increases reciprocity benefits. Finally, notice that $f(\cdot)$ doesn't depend on diffusion because it is based on the interaction of the focal observer with workers and not on the interactions between observers (we empirically explore such possibility in subsection 3.4).

## 3.2 The Choice of Becoming an Observer

A worker decides to become (or remain) the k$^{\text{th}}$ observer if the incremental payoff (the difference between Equations 1 and 2) is positive:

$$p\bar{b}d + r(b, \hat{d}_k) + f\left(\frac{b}{\widetilde{n}}\right) - c_0 b - c_T > 0. \tag{3}$$

For workers to be able to (approximately) evaluate this inequality, two conditions are required. The first is that workers can observe the variables, even if noisily. BAPP recommends that key aggregate metrics are made public to workers in a visible area of the site (e.g., "HR/ communication board"), particularly the contact rate ($\bar{b}d$). Regarding $p$, $c_0$, and $c_T$, workers are typically informed of these when they are approached to become observers or experience them as volunteers (while the learning benefits $p$ is more subjective and difficult to quantify, it is indeed communicated). The function $r(\cdot)$ is psychological and cultural; we rely on conversations with BAPP consultants as evidence that workers do reason about it. The function $f(\cdot)$ is subjective and private, and it is not wild to speculate that (some) workers consider how helping others and developing a reciprocal relation is inherently satisfactory. The second condition is that the worker translates these elements into a common (possibly monetary) value; for example, $\bar{b}$ translates into an accident reduction with (monetary) value for the worker (Lavetti and Schmutte 2016, Viscusi and Aldy 2003).

To study large-scale cooperation, we need to understand how inequality (3) is affected by diffusion $d$. When diffusion is still low, the first term is small, and thus $r(\cdot)$ needs to be sufficiently large for this inequality to be satisfied (given that we that $\widetilde{n} = n$, we assume $f(\cdot)$ to be small). As diffusion grows, $r(\cdot)$ will go down because $\dfrac{\partial r}{\partial \hat{d}_k} < 0$; if this derivative is sufficiently negative, in particular, if $\dfrac{\partial r}{\partial \hat{d}_k} < -(p\bar{b})$, there is a diffusion threshold after which the left-hand side of (3) becomes negative. The workers no longer choose to become observers. Online appendix A.1.1 offers evidence in favor of both conditions: of a sufficiently large $r(\cdot)$

when diffusion is low and the existence of a critical threshold after which becoming an observer is no longer attractive for workers considering it. This is a relaxed social dilemma (Hauert et al. 2006): cooperation is dominant up to a threshold diffusion, after which collective and private benefits conflict. Therefore, this analysis suggests that we might plausibly observe that the likelihood of deciding to become an observer is reduced as diffusion increases.

### 3.3 Choice of the Number of Observations as a Function of Observer Entry Order

If equation (3) is positive for a given worker, s/he decides to become an observer and then needs to determine how many observations $b$ to execute monthly by maximizing the following expression.

$$\max_{b} \quad p\bar{b}d + r(b, \hat{d}_k) + f\left(\frac{b}{\widetilde{n}}\right) - c_0 b - c_T. \tag{4}$$

The optimal number of observations $b^*$ is found by equating marginal benefit and cost, that is, $\frac{\partial r}{\partial b} + f'\left(\frac{b}{\widetilde{n}}\right)\frac{1}{\widetilde{n}} = c_0$. The reputation benefit of further observations is lower for later joining observers than for earlier ones because the marginal reputational benefit of $b$ is lower for later entrant observers $\left(\frac{\partial^2 r}{\partial b \partial \hat{d}_k} < 0\right)$. The second term in the left-hand side of the first-order condition doesn't change with diffusion. This implies that later joining observers are expected to exert lower effort than earlier observers. (In subsection A.1.2, we provide evidence-based conditions for this choice to be a social dilemma.)

### 3.4 Impact of the Number of Observers on Safety Outcomes

A basic tenet of BAPP is that safety is an increasing function of the overall contact rate $\bar{b}d$, namely $Safety = S(\bar{b}d)$ and $\frac{\partial S}{\partial(\bar{b}d)} > 0$. We know from the previous sections that the average effort $\bar{b}$ is a decreasing function of diffusion $d$; that is, $Safety = S(\bar{b}(d)d)$ and $\frac{\partial \bar{b}}{\partial d} < 0$.

Using these simple relationships, safety displays an inverted-U relation with observer entry number $\hat{d}_k$ as long as the reduction in $\bar{b}$ is large enough. The derivative of safety is $\frac{\partial S}{\partial d} = \frac{\partial S}{\partial\ (\bar{b}(d)d)}\left(\bar{b}(d) + d\frac{\partial \bar{b}}{\partial d}\right)$. When $d$ is small, $\bar{b}(d)$ (which we know is positive when $d > 0$) is larger than $\left|d\frac{\partial \bar{b}}{\partial d}\right|$ and thus the term in parentheses is positive; and given that $\frac{\partial S}{\partial\ (\bar{b}(d)d)} > 0$, the total derivative $\frac{\partial S}{\partial d}$ is positive. In contrast, when $d$ is large, the second term in the parentheses, $d\frac{\partial \bar{b}}{\partial d}$, is negative and large, and the first term $\bar{b}(d)$ shrinks; therefore, provided that the reduction in effort captured by $\partial \bar{b}$ is sufficiently large, the parenthesis becomes negative, and, given that $\frac{\partial S}{\partial\ (\bar{b}(d)d)} > 0$, the total derivative $\frac{\partial S}{\partial d}$ is negative. This is plausibly the case in our situation.

Intuitively, at the start, the impact of adding observers exerting effort is beneficial (a "volume" effect), but later on, adding more observers depresses the average number of observations executed in the site (an "effort" effect). When diffusion is sufficiently large, the latter impact comes to dominate. Therefore, safety has an inverted-U relation with $d$ (the number of observers), and there is an optimal $d^*$ that maximizes safety. A good safety measure is the number of accidents; given that fewer accidents mean more safety, accidents have a U-shaped relation with $d$.

In summary, this simple formal model tailored to the BAPP setting suggests the following hypothesis about how the reputational benefit of cooperation may weaken with the number of observers and how safety is affected. This hypothesis is tested in subsection 5.1.

**Hypothesis 1:** The reputation benefit for an observer will be lower for later entrant observers. Therefore, increasing the number of observers will be associated with i) a reduction in the average number of observations for later-entrant observers, ii) a lower willingness to be an observer for later-entrant observers, and iii) a U-shaped relationship between the number of observers and accidents.

### 3.5 Adding Population Structure Supports Cooperation Via Repeated Interactions

Standard implementations of BAPP encourage random observations across the site, leading to very few interactions between observer-worker pairs. With $n$ workers at the site and an observer executing $\bar{b}$ random observations a month, a pair's average number of interactions per month is $\frac{\bar{b}}{n}$. In a typical BAPP implementation, $\bar{b} = 5$ and $n = 250$, the number of monthly interactions within a pair is 0.02. DEKRA's stated motive for a low interaction frequency is to avoid the "blind eye," i.e., observers getting used to a particular worker's tasks and thus overlooking unsafe behavior or becoming lenient due to familiarity.

This is where population structure comes in: if the site of 250 workers is divided, for example, into 10 groups, and observers have to execute observations within the group they are located in, the interaction frequency with workers increases tenfold to 0.2 per month.

Thus, it is easy to verify that as the population structure reduces $\widetilde{n}$ the observer chooses more observations $b$ to satisfy their first-order condition for effort (from Equation 4) and that this will increase safety in the site (from discussion in the previous section). The increase in observations $b$ will be more pronounced for later entrant observers because, given that $r(.)$ is concave in $b$, in the first-order condition of equation (4), the $\frac{\partial r}{\partial b}$ is larger for latter entrant observers. Further, an increase in $b$ will lead to more workers becoming observers as inequality (3) is easier to satisfy. This leads to Hypothesis 2, which guides the design of the field experiment in this paper. It is tested in subsections 5.2, 5.4 and 5.2.

**Hypothesis 2:** An interaction structure that results in a higher number of repeated interactions between observer and worker will lead to: i) an increase in the number of observations (and more so for later entrant observers); ii) an increase in the willingness to become an observer; iii) a reduction in accidents. This will counteract the weakening of cooperation proposed in Hypothesis 1.

# 4 Data and Field Experiment

## 4.1 Datasets

**DEKRA datasets.** DEKRA provided a dataset of 1,352 sites with BAPP implementations between 1989 and 2013. Each site and month included information on the BAPP implementation and accident records, which DEKRA carefully reconciled across countries' different rules for reporting accident data. We restricted the sample to those implementations with information on workplace accidents for at least two years before and three years after the starting month of BAPP observations. This left us with a sample of 88 sites. Using an array of variables (number of employees, country, year, industry, etc.), we found that the sample was not significantly different from the overall population. For example, our sample sites have an average size of 245 workers, which is not statistically (at the usual significance levels) different from the average of 279 workers at the population level. This site-month dataset was used to estimate the impact of BAPP (see Figure 1a and subsection A.2.3) and to test Hypothesis 1 iii) in subsection 5.1.

DEKRA also provided us with information on all the observations from the 88 implementations in our sample, consisting of 1,265,176 observation sheets, each indicating the site, date, name of observer, area of the site, and observation characteristics (number of behaviors that were observed and recorded, time of day, location within the site, etc.), we used these data to estimate Hypotheses 1 i) and 1 ii) in subsection 5.1.

**Sodimac datasets.** We constructed two datasets at Sodimac, a large Chilean retail chain of home-improvement stores and a client of DEKRA, where we executed our experiment (see **??** below). The first is a panel of observers and months of BAPP implementation. We recorded observer names, numbers of observations, and information encoded in these observations (the number of coached observations, the number of critical behaviors observed and reported, and

the number of risky and safe behaviors), both for starting team observers and later observers, and the treatments they were allocated to. This dataset was used to study the impact of the treatments on cooperative effort and diffusion, testing Hypothesis 2 i).

The second dataset contained a monthly panel of workers and accidents from January 2016 to May 2018. From Sodimac's personnel registers, we had information about the workers in each month and participating store, including age, tenure, gender, and job title. We merged this personnel data with the information on all accidents at Sodimac. Each accident was indexed by the time, the ID of the injured worker, the type (e.g., with or without lost days), and the number of lost days. This dataset was used to study the treatments' impact on willingness to be an observer and accidents, testing Hypotheses 2 ii) and 2 iii). This dataset was also used to test the assumptions underlying the function $r(\cdot)$ in the simple model (see subsection 5.1).

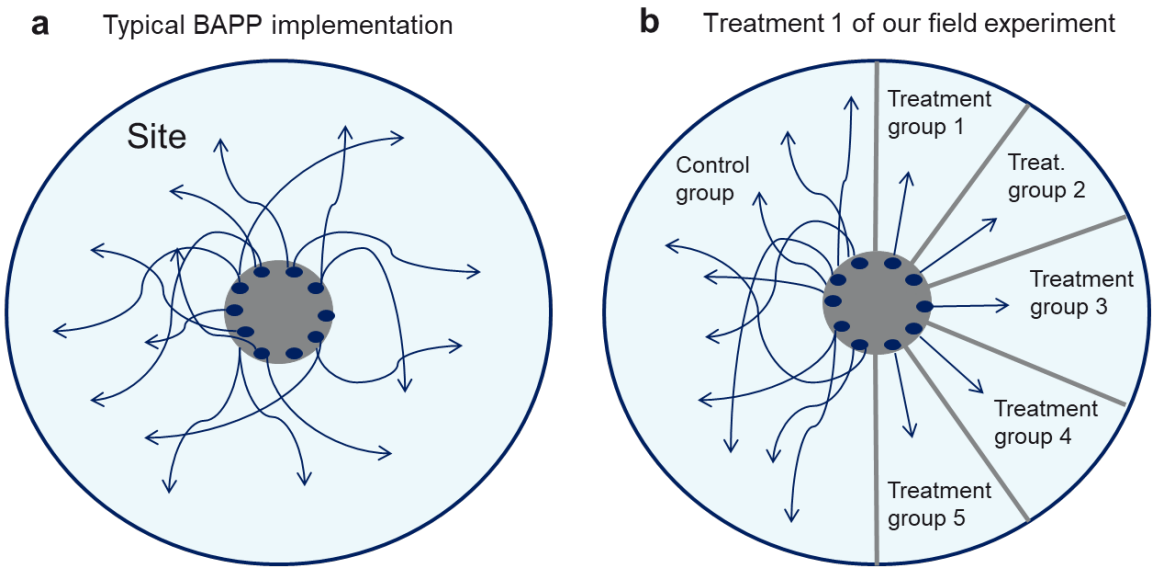
## 4.2 Pre-registered Field Experiment

We performed a pre-registered field experiment on four BAPP implementations within Sodimac. The experiment included four stores (Antofagasta, Temuco, Huechuraba, and La Reina), with BAPP-eligible workforces of 258, 268, 334, and 234 workers, respectively (excluding managers and supervisors).

We designed three treatments. Treatment 1 ("structure") was the baseline treatment applied to all four stores. It introduced repeated interactions by altering the observer assignment of BAPP; this treatment tested Hypothesis 2. Treatments 2 ("identity") and 3 ("reputation") aimed to explore mechanisms that might boost the impact of Treatment 1 and were applied to only two stores each. The resulting treatments per store are as follows: Antofagasta, Treatment 1; Temuco, Treatments 1 and 2; Huechuraba, Treatments 1 and 3; La Reina, Treatments 1, 2, and 3. We discuss the treatments 2 and 3 and their results in more detail in subsection 5.3 below.

Treatment 1 changed the interaction structure of BAPP by specifying who would be observed by whom. In standard implementations of BAPP, consultants encourage random observations across workers to avoid what DEKRA refers to as the "blind eye" mentioned earlier. Figure 2 shows how Treatment 1 departed from this.

Figure 2: Graphical Representation of BAPP implementation and Experimental intervention




**Notes:** This is the graphical representation of a standard BAPP implementation and our field experiment. Blue dots represent observers of the starting team (usually 10); the light blue area represents the site with 250 workers; the blue lines represent interactions (observations). **a.** A standard implementation of BAPP, in which observations are quasi-random, from starting team members to all workers. **b.** In our experiment, five observers and 125 workers were randomly matched in treatment groups of 25 workers. The remaining five observers, plus the enabler, could freely observe the remaining 125 workers (the left half of the circle), as in a standard BAPP implementation. (For simplicity, we do not consider the enabler; if we do, the figure and calculations change only slightly.)

Suppose the starting team had $k_s$ observers (excluding the enabler). Half of the observers were randomly chosen to receive the random assignment of $\frac{1}{(k_s+1)}$ 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 the remaining workers not assigned to a specific

observer (a list of these workers was provided to the unselected observers). This control group represented the standard BAPP with no structure imposed. As the implementation progressed, workers becoming observers could only observe workers of their group of origin (e.g., only observe workers in the control if the observer used to be a worker in the control; same for the groups created with the treatment). This randomization produced a balanced average number of workers assigned to observers: for example, in a site of 250 workers with ten observers in the starting team, our experiment created five groups of 23 randomly selected workers $\left(= \frac{250 * 1}{(10+1)}\right)$, each observed by a randomly selected observer; the control group had 135 workers, to be observed by the five remaining observers, plus the enabler (the ratio of workers per observer/enabler in the control group was $\frac{135}{(5+1)} = 23$, the same as in the treatment groups). Critically, compared with the control group, Treatment 1 increased the frequency of interactions between an observer and a given worker from $\frac{b}{135}$ per month in the control group to $\frac{b}{23}$ in the treatment groups. The theoretical increase in the likelihood of repeated interaction was $\frac{135}{23} = 5.9$. This improvement was slightly diluted because compliance with the assigned groups was 85%, not perfect (see subsection A.3.5), so the effective increase was a factor of 5.

The online appendix subsections A.3.5 to A.3.5 provide extensive detail on the firms involved in the experiment (Sodimac and ACHS), randomization and implementation protocols, treatment documents, balance checks, power calculations, and manipulation checks (or "take-up").

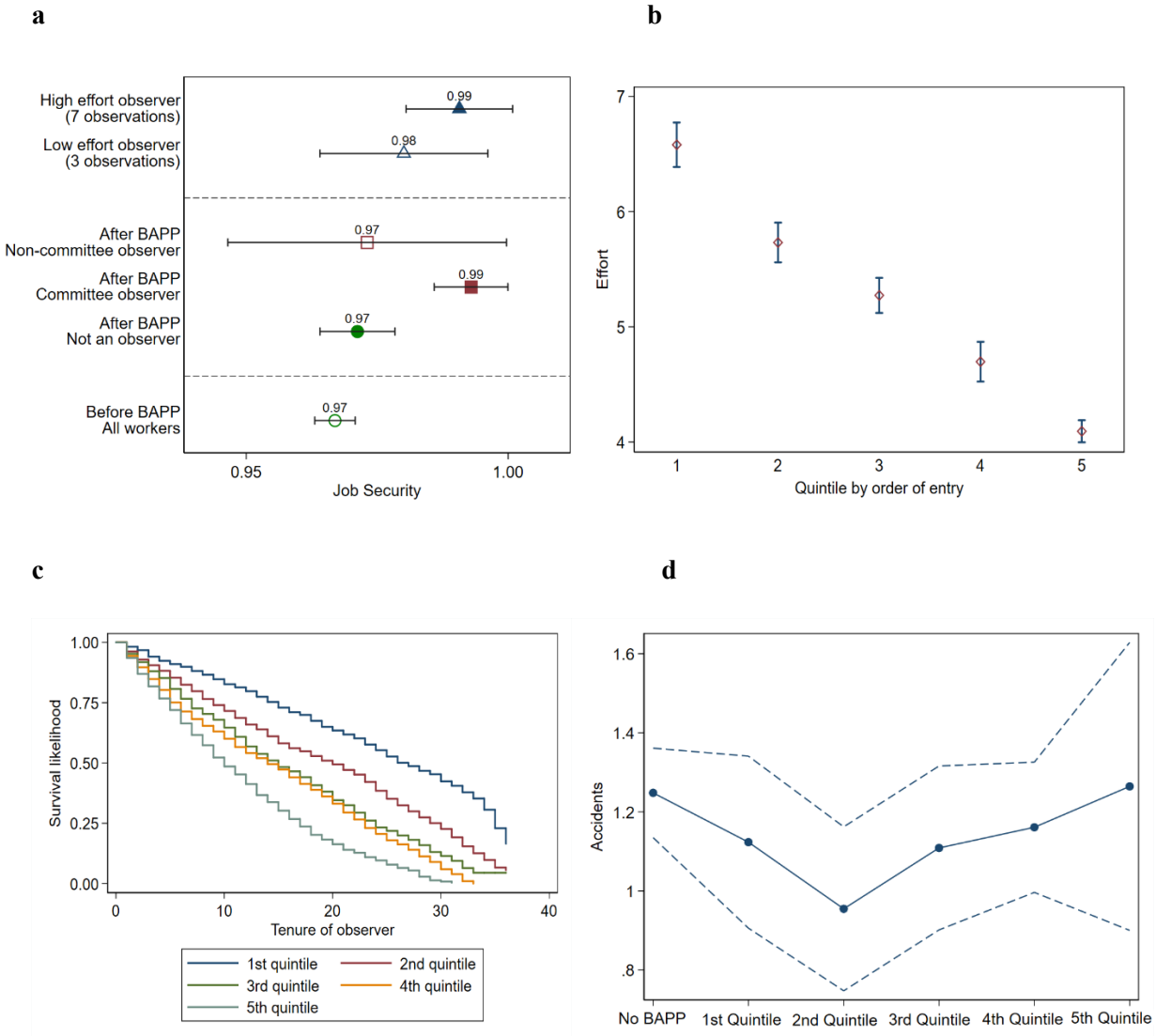
# 5 Results

## 5.1 Reputation Cannot Sustain Large-Scale Cooperation

This section tests Hypothesis 1. We provide evidence for the assumptions behind the reputational benefits function $r(\cdot)$. This function is crucial in generating the predictions of

H1.

Figure 3: Reputation benefits and reduction in cooperation and safety




**Notes: a.** Job security in a high-turnover environment is one measure of reputation: highly regarded workers are less prone to leave. The horizontal axis represents the probability that a worker remains in the next month. We plot the predicted likelihood of remaining an employee based on an econometric model and its 95% confidence interval (see subsection A.2.2 for details). **b.** The horizontal axis displays the five quintiles of when observers sign up; the vertical axis captures effort. We plot the mean effort and the 95% confidence intervals for each quintile using the raw data. (See subsection A.2.6 for a regression analysis). **c.** Kaplan-Meier survival estimates where the likelihood of remaining as an observer is plotted on the vertical axis and the months as an observer on the horizontal axis. (See subsection A.2.7 for a regression analysis using a Cox proportional hazard model.) **d.** The vertical axis displays the predicted accidents for each quintile from a regression model where other covariates are held at their means (see subsection A.2.8 for details); the horizontal axis displays different conditions: no BAPP and the five quintiles of observer sign up. The dotted lines depict a 95% confidence interval.

We use the "likelihood of remaining employed" (job security) as a proxy for reputational

benefits. In panel a of Figure 3 we use archival data from Sodimac to show that month-to-month job security is higher for earlier entrant observers (assumption $\frac{\partial r}{\partial \hat{d}_k} < 0$) as well as for those that do more observations (assumption $\frac{\partial r}{\partial b} > 0$). Evidence consistent with concavity (assumption $\frac{\partial^2 r}{\partial b^2} < 0$) and with effort being more valuable for early entrant observers (assumption $\frac{\partial^2 r}{\partial b \partial \hat{d}_k} < 0$) is provided in subsection A.2.2.

To test H1 i), we divide active observers at a site into five quintiles (cohorts) by order of entry (the date of an observer's first observation) (see subsection A.2.5 for details of the cohorts). Panel b in Figure 3 plots the mean effort per quintile and their respective 95% confidence intervals. Effort decreases significantly as we move to the higher quintiles. The first quintile executes 6.5 monthly observations, while the fifth only executes 4. This provides supportive evidence for Hypothesis 1 i). Regression analysis confirms these results in subsection A.2.6 while controlling for site-specific confounding factors.

Panel c of Figure 3 provides evidence supporting Hypothesis 1 ii). We display the Kaplan-Meier survival functions for each quintile of entry order. The graph shows that the willingness to remain an observer is reduced with the order of entry. For example, an observer of the first entry quintile has an 85% likelihood of remaining an observer after ten months of tenure, while an observer of the fifth quintile has only a 50% likelihood of remaining an observer after the same tenure. A log-rank test indicates that these survival functions are statistically different. subsection A.2.7 estimates a Cox proportional hazard model that confirms these descriptive patterns. In subsection A.2.9, we provide additional evidence in support of hypothesis 1 ii): using a regression that controls for site fixed effects, we find that diffusion has a concave relation with time: diffusion increases first, but it slows down, reaching a maximum at 30 months (and a "plateau" around that number).

Panel d of Figure 3 examines element iii) of Hypothesis 1. Using regression analysis (see the subsection A.2.8 for details), we find that adding observers initially decreases the number of accidents but then increases them, displaying a clear U relationship with accidents:

as compared to no BAPP present, increasing observers is beneficial in the first and second quintile, up to 25 active observers, beyond which adding more observers becomes detrimental. Subsection 3.4 illuminates the reason behind this: adding observers improves safety, but the decaying effort in new cohorts reduces it. Initially, the first effect is stronger, but eventually, the latter effect dominates.

## 5.2 Smaller Groups Can Sustain Direct Reciprocity and Large-Scale Cooperation

This section presents evidence from the field experiment suggesting that adding a group structure to BAPP remedies the reduction in cooperation and safety resulting from the decaying reputational benefits. The mechanism lies in the group structure promoting direct reciprocity between the observer and the observed workers, which motivates more observations. To test element i) of Hypothesis 2, we use the following regression model:

$$\begin{aligned} \mathit{Effort}_{ijt} &= b_1 + b_2\, Treat1_{ij} + b_3\, Treat1_{ij}\, Treat2_{ij} \\ &\quad + b_4\, Treat1_{ij}\, Treat3_{ij} + Cont_{ijt} + v_{jt} + u_{ijt}. \end{aligned} \tag{5}$$

The number of observations by observer $i$ at store $j$ during the month $t$ is regressed on the treatment dummies and controls (see Methods for details on the controls). Hypothesis 2 is evaluated by estimating $b_2$. Treatment 2 and Treatment 3 enter as interaction effects to explore whether Treatment 1 can be boosted by identity or reputation (the next section discusses these two treatments).

We controlled by tenure, measured as the number of months the observer had been active (TEN), to capture the ramp-up in observations occurring naturally when observers enter BAPP. The dummy variable NEW took the value one if the observer was not a starting team member. We controlled for the interaction between TEN and NEW, as the dynamics are

different (if we add a time dimension to panel b of Figure 2, we observe some convergence over time between the starting team and the new). We also added store-and-month dummies ($v_{jt}$) to control for differences in starting dates across stores (see subsection A.3.4); given the ramp-up of sites, not controlling for heterogeneous starting dates could introduce bias. We controlled 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. Not controlling for their presence would have introduced a downward bias in $b_2$ because enablers typically execute more observations than the rest of the observers (however, excluding them from the sample yielded consistent results).

Table 1: Impact of treatments on number of observations and worker behavior

| | Effort (observations) | | Risky behavior | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Treat. 1 | 0.97* (0.53) | | -0.99* (0.52) | |
| Treat. 1 x starting team observer | | 0.58 (0.66) | | -1.09 (0.70) |
| Treat. 1 x new observer | | 1.38** (0.57) | | -0.89* (0.53) |
| Treat. 1 x Treat. 2 | -1.52** (0.67) | -1.56** (0.68) | 1.15* (0.68) | 1.14* (0.68) |
| Treat. 1 x Treat. 3 | -0.74 (0.61) | -0.51 (0.64) | 0.14 (0.70) | 0.20 (0.75) |
| Enabler | 3.40** (1.37) | 3.28** (1.34) | 0.76 (0.71) | 0.74 (0.73) |
| Tenure | 0.12 (0.14) | 0.12 (0.14) | -0.08# (0.13) | -0.07# (0.13) |
| Tenure x new observer | -0.04 (0.16) | -0.04 (0.16) | -0.16# (0.16) | -0.15# (0.16) |
| New observer | -1.17 (0.88) | -1.60* (0.91) | 0.62 (1.06) | 0.51 (1.10) |
| Critical behaviors observed | | | 0.02 (0.01) | 0.02 (0.02) |
| Effort | | | 0.48*** (0.15) | 0.48*** (0.15) |
| Store-month fixed effects? | Yes | Yes | Yes | Yes |
| Observations | 585 | 585 | 585 | 585 |
| R-squared | 38.95% | 39.33% | 49.73% | 49.75% |
| Mean (standard deviation) | 5.02 (2.82) | 5.02 (2.82) | 3.47 (0.69) | 3.47 (0.69) |

**Note:** All regressions are estimated with OLS. Errors in parentheses are 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 including Treatments 2 and 3 interaction with new and starting team observers. "Critical behaviors observed" measures how many predefined critical behaviors the observer observed and reported in the observation sheet (independent of whether the worker was performing in a safe or risky way). The observed critical behaviors could be fewer than in the predefined set because some workers did not engage in some tasks (e.g., cashiers do not drive forklifts or climb ladders).

Table 1 displays the results. Column (1) indicates that Treatment 1 generated an increase of 0.97 observations per month (p-value = 0.068). Thus, Treatment 1 operated as intended: it counteracted the reduction in cooperative effort as the number of observers increased. This supports element i) of Hypothesis 2. Column (2) refines the test by splitting the impact of Treatment 1 into two components: the effect on new observers (captured by the dummy variable NEW) and the effect on observers in the starting team (START = 1 - NEW). The impact is concentrated on the new observers, who conduct 1.38 more observations (p-value = 0.016) (this effect is similar in size to the reduction in observations documented in panel b of Figure 2). Observers in the starting team display 0.58 additional observations under Treatment 1 (p-value = 0.380). This result suggests that the higher effort (observations) from direct reciprocity affects observers asymmetrically; in particular, following our model, this means that later entrant observers get a more considerable boost than early entrant observers.

Columns (3) and (4) provide evidence that the mechanism underlying this result lies in the repeated interactions between workers and observers, leading to workers changing their behavior. During an observation, the observer recorded whether or not the worker was performing a predefined list of critical behaviors in a risky manner. Using the number of risky behaviors as the dependent variable, workers under Treatment 1 performed 0.99 fewer risky behaviors (column (3), p-value = 0.057), which is economically significant compared with the mean of 3.47. This shows that workers reciprocate the observations by following the advice given to them by the observers. Moreover, column (4) indicates no difference between starting or new observers; this is consistent with our theory because, from the worker's perspective, repeated interactions increased with both types of observers. Below, in subsection 5.4, we summarize additional evidence that supports repeated interactions and direct reciprocity as the mechanism behind the impact of Treatment 1.

To test element ii) of Hypothesis 2, we study the likelihood of becoming an observer:

$$\begin{aligned} Observer_{ijt} &= b_1 + b_2\, Treat1_{ij} + b_3\, Treat1_{ij}\, Treat2_{ij} \\ &\quad + b_4\, Treat1_{ij}\, Treat3_{ij} + X_{it} + \tau_{tj} + u_{ijt}. \end{aligned} \tag{6}$$

This model uses all BAPP-eligible workers at the site, excluding those who belong to the starting team, because these workers become observers before treatments are assigned. $Observer_{ijt}$ is a dummy variable set to 1 if the worker $i$ in store $j$ is an active observer in month $t$, and 0 otherwise. Treatment variables do not have time indices because we estimate this model using the BAPP implementation period, where every worker is assigned to a particular treatment. $X_{it}$ is a vector of worker-level controls for each period (age, tenure, gender, and job title). $\tau_{tj}$ are store and month fixed effects.

Table 2 presents the results. The sample in column (1) includes all months of BAPP implementation, including the initial months where recruiting did not occur. Since this may bias the results downwards (the first few months are very slow in recruiting), column (2) restricts the analysis to May 2018. The results of column (1) indicate that Treatment 1 increases the likelihood of becoming an observer by 1.9 percentage points (p-value = 0.144), which is almost the same size as the average likelihood of 2.2%. For May 2018, the results are also substantial and more precisely estimated: a 5.4% increase (p-value = 0.031) over the average of 5.2%. This supports element ii) of Hypothesis 2: Treatment 1 increases the entry likelihood of late-entering observers.

Finally, we measure the impact of Treatment 1 on safety (element iii) of Hypothesis 2), reporting three safety measures: workplace accidents and two subgroups, workplace accidents without lost working days, and workplace accidents with lost working days. We use the following model:

Table 2: Impact of the treatments on the probability of becoming an observer and the likelihood of experiencing an accident

| | P(observer) | P(observer) by May 2018 | Workplace accidents | | Workplace accidents without lost working days | | Workplace accidents with lost working days | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Treat. 1 | 0.019 | 0.054** | -0.0007 | -0.0030** | -0.0014* | -0.0022* | 0.0069 | -0.0083 |
| | (0.013) | (0.025) | (0.0012) | (0.0015) | (0.0086) | (0.0012) | (0.0080) | (0.0087) |
| Treat. 1 × Treat. 2 | -0.021* | -0.072** | | 0.0047** | | 0.0034** | | 0.0013 |
| | (0.012) | (0.028) | | (0.0022) | | (0.0016) | | (0.0015) |
| Treat. 1 × Treat. 3 | -0.009 | -0.006 | | -0.0013 | | -0.0030* | | 0.0016 |
| | (0.010) | (0.027) | | (0.0024) | | (0.0018) | | (0.0017) |
| Individual controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Store-month fixed effects | Yes | No | Yes | Yes | Yes | Yes | Yes | Yes |
| Store fixed effects | No | Yes | No | No | No | No | No | No |
| Observations | 10,879 | 1,072 | 11,277 | 11,277 | 11,277 | 11,277 | 11,277 | 11,277 |
| R-squared | 0.027 | 0.011 | 0.0071 | 0.0075 | 0.0044 | 0.0051 | 0.0058 | 0.0059 |
| Mean | 0.022 | 0.052 | 0.0037 | 0.0037 | 0.0019 | 0.0019 | 0.0018 | 0.0018 |

**Note:** All regressions are estimated with OLS. Errors in parentheses are robust and clustered at the worker level. * p < 0.1, ** p < 0.05, *** p < 0.01. All regressions exclude starting team members. Samples are restricted to months, and stores with BAPP have already been implemented. If we use count models in columns (3) to (8) (e.g., Poisson regression), the results do not change.

$$\begin{aligned} Accident_{ijt} &= b_1 + b_2\, Treat1_{ij} + b_3\, Treat1_{ij}\, Treat2_{ij} \\ &\quad + b_4\, Treat1_{ij}\, Treat3_{ij} + X_{it} + \tau_{tj} + u_{ijt}. \end{aligned} \tag{7}$$

We use the BAPP implementation period, and $X_{it}$ and $\tau_{tj}$ are the same as in Equation (6). Columns (3) to (8) of Table 2 present the results. Treatment 1 reduces workplace accidents (column (4)). This impact is concentrated on accidents without lost working days (column (6)), which is reassuring because—as confirmed by BAPP consultants—BAPP is focused on these types of accidents (more serious accidents, although affected by the behavior of workers, also are determined by the age/quality of technology used in production, the availability and maintenance of machinery/equipment, and other organizational level variables). The impact of Treatment 1 is sizeable: it eliminates 0.0022 accidents without lost time per month (p-value

= 0.067), which is 40% of the pre-BAPP mean average and accounts for a large proportion of the overall impact of BAPP on accidents in Sodimac. To measure the latter, we replicated in subsection A.2.4, the before-and-after analysis performed the DEKRA dataset (which was displayed in Figure 1a and subsection A.2.3). These results support element iii) of Hypothesis 2.

## 5.3 Group Identity and Reputation Do Not Boost the Impact of Repeated Interactions

We added two treatments to test whether the effect of dyadic interactions might be boosted by group identity (Treatment 2) and reputation (Treatment 3). Sodimac's sample size limitations prevented our research design from measuring the direct impact of identity and reputation, so we focused on their interactions with Treatment 1.

For Treatment 2 (group identity), we added three elements to the letters of Treatment 1 (subsection A.3.2 displays the letters). 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. These two elements leverage "the minimal group paradigm" well known in social psychology (MGP) (Tajfel 1970): groups that are tagged even at a trivial or arbitrary label exhibit greater help for in-group members (Tajfel 1982). The third element is that, at the end of the letter, we added a list with the names of all group members (i.e., part of the group created by Treatment 1) and their job role in the store (e.g., cashier). This last element builds on recent findings that the positive effect of the MGP on help might not materialize unless groups are provided with a joint history (Bernhard et al. 2006, Buchan et al. 2006, Charness et al. 2007, Goette et al. 2006), even if this is a minimal introduction (Loch and Wu 2008) or mere common knowledge of group affiliation (Guala et al. 2013).

The results of Tables 1 and 2 highlight that the effect of group identity depends on the context: adding Treatment 2 to Treatment 1 reduces the number of observations and the

willingness to be an observer and increases accidents so much that it nullifies the baseline impact of Treatment 1. The context has been alluded to and is explained by the exit interviews: partially removing the anonymity of BAPP by revealing names jeopardized the BAPP motto of "no spying, no naming, no blaming," which generated a worker backlash. This backlash reduced worker willingness to collaborate, affecting the observers' efforts. DEKRA's consultants and Sodimac executives concurred with this interpretation, noting that in a unionized and recently contentious working environment (Sodimac had experienced several strikes in the year before our experiment), worries about being "spied on" and "ratted out" by observers were real, and that the distaste for the violation of anonymity was stronger than any identity effects that the treatment might have generated. The subsection A.3.7 discards several alternative explanations to bolster this interpretation.

Inspired by the reputation treatment implemented by Yoeli et al. (2013), Treatment 3 was executed as follows: at the beginning of each month, the research team would generate a report that included the name of the observer, their starting date, the accumulated number of observations until the previous month, and the monthly average number of observations. The resulting list was ranked by the average monthly observations, sorted from highest to lowest, and then published on the site's bulletin board. This board was widely read at Sodimac, included all BAPP public communications, and was in a frequently visited/transited location. We certified execution by requesting photographic evidence of the report's publication (see subsection A.3.3 for the report and a sample photograph of its implementation). Predictions for Treatment 3 are mixed. Earlier theoretical research suggested that when direct information from repeated interactions is available, reputation might be less impactful as agents prioritize first-hand information (Roberts 2008); however, a recent model suggests that this relation might be more complex without a preeminence of one type over the other (Schmid et al. 2021). Moreover, evidence from the lab shows that reciprocity and reputation are robust to each other's presence (Melamed et al. 2020).

The results of Tables 1 and 2 show that Treatment 3 had a null effect on the number of observations, the risky behavior of workers, accidents, and the willingness to become an observer. This supports the theories that indicate the importance of directly obtained information via repeated interaction over information obtained by third parties in the form of public reputation (Roberts 2008).

## 5.4 Additional Evidence for Dyadic Repeated Interactions as the Causal Mechanism

It is more challenging in field experiments than in the lab to fully uncover the mechanisms at play. We executed several additional analyses to increase our confidence that repeated interactions and reciprocity between observers and workers drove the results of Treatment 1. Here we provide a one-sentence summary of each (see subsection A.3.6 for details): i) workers in Treatment 1 also experienced fewer commuting accidents (between home and work), a sign of that workers not only absorbed observers' advice and changed their behavior inside the workplace (see columns (3) and (4) of Table 1 and associated discussion) but also beyond it; ii) we discarded three alternative explanations: self-selection (i.e. the decay in cooperation of hypothesis 1 might be driven by early observers being more prosocial or altruistic than later ones), quality of observers (i.e., randomization might lead to unbalanced observers in terms of unobserved quality which in turn drive the results of treatment 1), and observers under Treatment 1 creating a team spirit (i.e., observers within each treatment group might have generated a strong bond among themselves and this motivated them to exert effort); iii) the exit interviews confirm that a reciprocal relationship was fostered between observers and workers in Treatment 1; and iv) we replicate the results of Treatment 1 with the DEKRA archival data by exploiting variance in observers' naturally occurring within-site repeated interactions (i.e., naturally occurring variation in $\widetilde{n}$ and thus, in the capacity for reciprocity to flourish).

## 5.5 Evidence for Intrinsic Reciprocity

Reciprocity can be either "instrumental" or "intrinsic" (Sobel 2005, Cabral et al. 2014). In the former, observer and worker cooperate with one another based on the expectation of continuing with a fruitful exchange and is based on purely selfish motives (not prosocial preferences); in the latter, cooperation is driven by responding to past cooperation (defection) with cooperation (defection) and is not driven by foresight and calculation, but by preferences that display an "intrinsic desire to reciprocate" (Sobel (2005); p. 432).

To test which type of reciprocity might be driving our results, we collected information regarding the social preferences of a sample of observers; in particular, we measured their altruism (using a dictator game) and their willingness to punish selfishness (using a third-party punishment game). Agents possessing both traits have been labeled as "strong reciprocators" in the literature (Gintis 2000). If instrumental reciprocity is driving our impact of treatment 1, then this should not be affected by whether the observer is a "strong reciprocator" or not. In contrast, if intrinsic reciprocity is the driving mechanism, then the impact of treatment 1 should be concentrated on observers who possess these two traits.

We sent an online survey to every observer immediately after s/he signed up. The survey was voluntary and confidential and was answered by 57 observers. The survey included a terse explanation of the research project (revealing neither the topic nor the purpose of the research). The survey comprised a Dictator game, a Third-party punishment game, a "Big 5" personality traits questionnaire, and questions about the social network of observers (we used the "Big 5" and social networks are used to test selection issues between observers in the starting team and those that signed latter; see subsection A.3.6). For the dictator game, we asked employees to imagine receiving an endowment of 10,000 pesos and decide how much to give a stranger $(0, 1, 2, \ldots, 10)$ thousand pesos. For the Third-party punishment game, we asked employees to imagine they received an endowment of 5,000 pesos, asked them to observe a dictator game being played by two other players (with the same endowments as

the DG they played before), and decide whether to sacrifice 1 or 2 thousand Pesos to reduce the dictator's endowment by 3 and 6 thousand respectively; we used the strategy method, namely, they chose different choices of the dictator $(0, 1, 2, 3, \ldots, 10)$.

Using this data, Table 3 replicates the analysis displayed in Table 2 but interacts with Treatment 1 with altruism and punishment from the survey. The variable "Altruism" is the donations from the dictator (in 000 pesos, going from 0 to 10) (mean=4.5, sd=2.6, min=0, max=10, percentile25=2, percentile 75=5, median=5). The variable "Punishment" is a dummy variable that takes the value of 1 when the observer is willing to sacrifice either 1 or 2 thousand of his endowment to punish the dictator that gives 0 (mean=0.69) (the results do not change if we expand to the cases when the dictator provides "only" 1 or 2 thousand Pesos).

In columns (1) and (2), we find similar results as in Table 2. Columns (3) to (5) have positive coefficients for the interactions between Treatment 1 and altruism and punishment, but they are not precisely estimated. Column (6) shows that while the triple interaction between treatment 1, altruism, and punishment is positive and statistically significant, the two-way interactions are negative. Thus, to assess how these traits affect the impact of treatment 1, we need to compute all four types of observers. The effect of treatment 1 for non-altruistic (setting altruism equal to zero) and not punishing observers (setting punishment equal to zero) is 2.59 additional observations (imprecisely estimated); when an observer is high on altruism (setting the variable equal to 5, or percentile 75) and high on punishment (setting the variable equal to 1), the impact of treatment 1 is 2.2 additional observations (=2.59 - 0.34 x 5 - 2.65 + 0.8 x 5) (statistically significant); when the observer is high altruism and low punishment the impact is 0.89 additional observations (= 2.59 - 0.34 x 5) (imprecisely estimated); and when the observer is low altruism and high punishment, the effect is 0.06 fewer observations (=2.59 - 2.65) (imprecisely estimated). Thus, this provides evidence in favor of both instrumental and intrinsic reciprocity: treatment 1 is concentrated

Table 3: Observer altruism and punishment increases the impact of Treatment 1 on observations

| | Dependent variable: Observations | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Treat. 1 | 1.26 | | 0.30 | 0.57 | 0.03 | 2.59 |
| | (1.03) | | (1.12) | (1.15) | (1.18) | (1.94) |
| Treat. 1 x starting team observer | | 1.10 | | | | |
| | | (1.09) | | | | |
| Treat. 1 x new observer | | 1.86* | | | | |
| | | (1.05) | | | | |
| Treat. 1 x Altruism | | | 0.26 | | 0.27 | -0.34 |
| | | | (0.17) | | (0.20) | (0.30) |
| Treat. 1 x Punishment | | | | 0.85 | 0.36 | -2.65 |
| | | | | (0.84) | (1.05) | (1.91) |
| Treat. 1 x Altruism x Punishment | | | | | | 0.80* |
| | | | | | | (0.42) |
| Altruism | | | -0.06 | | -0.07 | 0.13 |
| | | | (0.09) | | (0.09) | (0.20) |
| Punishment | | | | -0.40 | -0.40 | 0.73 |
| | | | | (0.48) | (0.49) | (1.21) |
| Altruism x Punishment | | | | | | -0.25 |
| | | | | | | (2.69) |
| Treat. 1 x treat. 2 | -1.90 | -1.97 | -2.06* | -1.79 | -2.10* | -2.85** |
| | (1.17) | (1.19) | (1.19) | (1.15) | (1.17) | (1.39) |
| Treat. 1 x treat. 3 | -1.02 | -0.91 | -1.26 | -0.93 | -1.17 | -1.37 |
| | (0.94) | (0.96) | (1.19) | (0.94) | (1.02) | (0.97) |
| Controls? | Yes | Yes | Yes | Yes | Yes | Yes |
| Store-month fixed effects | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 333 | 333 | 333 | 333 | 333 | 333 |
| R-squared | 43.05% | 43.25% | 43.92% | 43.35% | 44.09% | 45.10% |
| Mean (Standard deviation) | 5.42 (3.18) | 5.42 (3.18) | 5.42 (3.18) | 5.42 (3.18) | 5.42 (3.18) | 5.42 (3.18) |

**Note:** All regressions are estimated with OLS. Errors in parentheses: robust and clustered at the observer level. * p<0.1,** p<0.05, *** p<0.01. Due to missing responses on some variables, we used 55 observers.

both on observers that do not display prosocial preferences (and therefore reciprocate out of convenience and calculation of future benefits) and on observers that are who display prosocial preferences of the "strong reciprocity" type (Gintis 2000). While statistical power is low in this analysis, and thus only speculative, notice that the correlation between the variables punishment and altruism is low, only 0.11; this means that we have a balanced distribution in types of observers to draw inferences from (16% low-low; 38% high-high; 15%

high altruism and low punishment; 31% low altruism and high punishment).

Finally, notice that, as one would expect, strong reciprocators only exert an impact on observation when relationships can be established. An observer who is high on altruism and punishment in the control group generates 0.13 additional observations (=0.13 x 5 + 0.73 - 0.25 x 5), while an observer with similar traits in treatment 1 generates 2.2 additional observations (same as above).

# 6 Discussion and Conclusion

## 6.1 Main findings and contribution

This article adds field evidence to the long-discussed question of whether reciprocity (dyadic relationships) and reputation (indirect reciprocity) can support large-scale cooperation (Raihani 2021, Henrich and Muthukrishna 2021). Reciprocity has been believed to break down at scale, but recent theoretical models suggest that it can be maintained with the proper population structure (Van Veelen et al. 2012, Allen et al. 2017). In contrast, indirect reciprocity may sustain large-scale cooperation but requires more subtle and complex machinery that may curtail its power in real-world settings (Számadó et al. 2021, Santos et al. 2021). We analyze an empirical field setting where a small group of workers was trained to advise coworkers on workplace safety voluntarily without compensation, and the initial small group then expanded by enrolling new workers as additional advice providers. Thus, a scale-up of cooperation was tested.

This study makes two contributions to this literature. First, it shows that, in our setting, cooperation weakens as the number of volunteers increases, and this weakening is caused by decreasing returns of the reputation benefits. Early volunteers enjoy a substantial reputation benefit, but later, volunteers are seen as "normal." This adds to our understanding of the nuances of reputation (Számadó et al. 2021, Santos et al. 2021, Giardini et al. 2022, MacLeod

2007).

Second, using a pre-registered field experiment, this study demonstrates that changing the interaction structure between volunteers and workers increases the frequency of interactions. Thus, relationships are promoted, fully restoring voluntary cooperation lost from the dwindling reputational benefits. This provides empirical evidence in favor of recent formal models suggesting that the best modification of an interaction structure (network structure) in a population to promote large-scale cooperation might be to infuse it with strong dyadic relationships via repeated interactions (Allen et al. 2017). This model and our findings are consistent with the documented prevalence of strong dyadic relations in society (Peperkoorn et al. 2020).

## 6.2 Implications

Beyond the detailed field evidence of cooperation in large groups, our study contributes to the economics of organization (Gibbons and Roberts 2013). First, we illustrate the general idea of "interaction structure" as a mechanism that supports and sustains cooperation in large groups: who is paired with whom and how they interact plays a crucial role in generating cooperation. Appropriate structures, plus a replicator dynamic, ensure that cooperation can spread over time and resist invasion from defecting individuals.

Second, extensive group cooperation is at the base of critical phenomena of interest in these two fields. In organizational economics, there is a strong interest in understanding the root of persistent performance differences among seemingly similar enterprises (PPD among SSE) (Gibbons and Roberts 2013). We believe that extensive group cooperation, understood through the lenses of "interaction structures" and replicator dynamics, can complement the advancements generated based on rational actor models of instrumental reciprocity (as opposed to emotional reciprocity). Our results represent evidence that extensive group cooperation can be a fundamental ingredient for group members to come together, to share

ideas, knowledge, and goals, and to coordinate to produce valuable goods and services that no individual could make on their own. We propose that interaction structures facilitate cooperation in big groups, facilitating organizational performance.

Third, there is a fundamental literature on volunteering: individuals devoting time to help causes and initiatives (which are usually prosocial). This is relevant to firms because volunteering may also occur within firms, as our case of the BAPP methodology shows, but also because "recent years have seen a proliferation of corporate initiatives providing employees with an opportunity to be involved in projects with explicit social impact goals, often in partnership with nonprofit organizations" (Bode and Singh 2018, p. 1004). Given that volunteering is economically important (Linardi and McConnell 2011) and reputation concerns are crucial in motivating volunteering (Exley 2018, Carpenter and Myers 2010, Linardi and McConnell 2011, Exley 2016, Bode and Singh 2018), our findings may carry more general relevance, especially to settings where volunteering displays decreasing reputational benefits. One of those settings is emergency response organizations (e.g., firefighting), primarily supported by volunteering schemes worldwide, where reputational benefits are larger for the first responders, as they face the greatest danger and risk.

## 6.3 Limitations

As usual, not all aspects of a field study can be designed; there are limitations in our study. The first is that, while we study how cooperation affects an organizational outcome, that outcome is accidents and not profitability. Nonetheless, we performed a non-reported (but available upon request) analysis of how BAPP affects organizational culture at the site level. For a subset of 70 projects, DEKRA measured the culture of sites at two points in time, before BAPP implementation and at some point after it (on average three years after, with a range of 1 to 6 years); this allows us to include site fixed effects in the analysis. There are nine dimensions of culture being measured, which are grouped into three buckets: "safety

climate," "teamwork" (i.e., effective work and good relations with colleagues), and "leadership" (i.e., leader/organization that is fair, supportive, caring, and credible). We find that BAPP increases "safety climate," which in turn increases "teamwork" and "leadership." Arguably, these two factors improve performance; thus, BAPP and the cooperation it fosters generate not only improvements in safety but also in culture and, through that channel, in performance.

A second limitation is that we do not measure reputation directly in the non-experimental analysis; instead, we infer its presence based on clues, such as institutional knowledge (e.g., conversations with DEKRA consultants), impact on job security (especially the shape of this impact), and close correspondence with the model predictions.

We conclude by mentioning one managerial implication that we believe to be important. Teamwork and collaboration are widely cited as performance drivers, but know-how on fostering collaboration is not explicitly available. Our study points to the crucial importance of the interaction structure of workers in generating cooperation and, thus, to the role that the formal organization of companies, departments, and units can play in such a challenge.

# References


Akerlof, G. A. and Kranton, R. E. (2005). Identity and the economics of organizations. *Journal of Economic perspectives*, 19(1):9–32.

Alchian, A. A. and 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., and Nowak, M. A. (2017). Evolutionary dynamics on any population structure. *Nature*, 544(7649):227–230.

Ashraf, N. and Bandiera, O. (2018). Social incentives in organizations. *Annual Review of Economics*, 10:439–463.

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

Baker, W. E. and Bulkley, N. (2014). Paying it forward vs. rewarding reputation: Mechanisms of generalized reciprocity. *Organization science*, 25(5):1493–1510.

Barnard, C. I. (1968). *The functions of the executive*, volume 11. Harvard university press.

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

Bode, C. and Singh, J. (2018). Taking a hit to save the world? e mployee participation in a corporate social initiative. *Strategic Management Journal*, 39(4):1003–1030.

Bohnet, I. and Huck, S. (2004). Repetition and reputation: Implications for trust and trustworthiness when institutions change. *American Economic Review*.

Bolton, G., Dimant, E., and Schmidt, U. (2021). Observability and social image: On the robustness and fragility of reciprocity. *Journal of Economic Behavior & Organization*, 191:946–964.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American economic review*, 91(1):166–193.

Boyd, R. and Richerson, P. J. (1988). The evolution of reciprocity in sizable groups. *Journal of theoretical Biology*, 132(3):337–356.

Boyd, R. and Richerson, P. J. (1989). The evolution of indirect reciprocity. *Social networks*, 11(3):213–236.

Boyd, R. and Richerson, P. J. (2005). *The origin and evolution of cultures*. Oxford University Press.

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

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):373–398.

Cabral, L., Ozbay, E. Y., and Schotter, A. (2014). Intrinsic and instrumental reciprocity: An experimental study. *Games and Economic Behavior*, 87:100–121.

Carpenter, J. and Myers, C. K. (2010). Why volunteer? evidence on the role of altruism, image, and incentives. *Journal of Public Economics*, 94(11-12):911–920.

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

Cropanzano, R. and Mitchell, M. S. (2005). Social exchange theory: An interdisciplinary review. *Journal of management*, 31(6):874–900.

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):60–114.

Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2009). Homo reciprocans: Survey evidence on behavioural outcomes. *The Economic Journal*, 119(536):592–612.

Dumas, M., Barker, J. L., and Power, E. A. (2021). When does reputation lie? dynamic feedbacks between costly signals, social capital and social prominence. *Philosophical Transactions of the Royal Society B*, 376(1838):20200298.

Elfenbein, D. W., Fisman, R., and McManus, B. (2015). Market structure, reputation, and the value of quality certification. *American Economic Journal: Microeconomics*, 7(4):83–108.

Emerson, R. (1976). Social exchange theory. *Annual Review of Sociology*, 2:335–362.

Exley, C. (2018). Incentives for prosocial behavior: The role of reputations. *Management Science*, 64(5):2460–2471.

Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, 83(2):587–628.

Fehr, E. (2018). Behavioral foundations of corporate culture. *University of Zurich, UBS International Center of Economics in Society, Public Paper*, (7).

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics letters*, 71(3):397–404.

Friebel, G., Heinz, M., Krueger, M., and Zubanov, N. (2017). Team incentives and performance: Evidence from a retail chain. *American Economic Review*, 107(8):2168–2203.

Gächter, S. and Herrmann, B. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1518):791–806.

Gartenberg, C., Prat, A., and Serafeim, G. (2019). Corporate purpose and financial performance. *Organization Science*, 30(1):1–18.

Gartenberg, C. and Zenger, T. (2023). The firm as a subsociety: Purpose, justice, and the theory of the firm. *Organization Science*, 34(5):1965–1980.

Ge, E., Chen, Y., Wu, J., and Mace, R. (2019). Large-scale cooperation driven by reputation, not fear of divine punishment. *Royal Society Open Science*, 6(8):190991.

Giardini, F., Balliet, D., Power, E. A., Számadó, S., and Takács, K. (2022). Four puzzles of reputation-based cooperation: Content, process, honesty, and structure. *Human Nature*, 33(1):43–61.

Gibbons, R. and Henderson, R. (2012). Relational contracts and organizational capabilities. *Organization science*, 23(5):1350–1364.

Gibbons, R. and Roberts, J. (2013). *The handbook of organizational economics*. Princeton University Press.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of theoretical biology*, 206(2):169–179.

Goette, L., Huffman, D., and Meier, S. (2006). The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. *American Economic Review*, 96(2):212–216.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American sociological review*, pages 161–178.

Graham, J. R., Grennan, J., Harvey, C. R., and Rajgopal, S. (2022). Corporate culture: Evidence from the field. *Journal of financial economics*, 146(2):552–593.

Greif, A. (1993). Contract enforceability and economic institutions in early trade: The maghribi traders' coalition. *The American economic review*, pages 525–548.

Grennan, J. (2020). Communicating culture consistently: Evidence from banks. *Available at SSRN 3350645*.

Guala, F., Mittone, L., and Ploner, M. (2013). Group membership, team preferences, and expectations. *Journal of Economic Behavior & Organization*, 86:183–190.

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):195–202.

Hauser, O. P., Hendriks, A., Rand, D. G., and Nowak, M. A. (2016). Think global, act local: Preserving the global commons. *Scientific reports*, 6(1):36079.

Henrich, J. and Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annual review of psychology*, 72:207–240.

Hergueux, J., Henry, E., Benkler, Y., and Algan, Y. (2023). Social exchange and the reciprocity roller coaster: evidence from the life and death of virtual teams. *Organization Science*, 34(6):2296–2314.

Hermalin, B. E. (2013). Leadership and corporate culture. *Handbook of organizational economics*, 432478.

Holmstrom, B. (1982). Moral hazard in teams. *The Bell journal of economics*, pages 324–340.

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

Kamei, K. and Putterman, L. (2017). Play it again: Partner choice, reputation building and learning from finitely repeated dilemma games. *The Economic Journal*, 127(602):1069–1095.

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

Khadjavi, M. (2017). Indirect reciprocity and charitable giving—evidence from a field experiment. *Management Science*, 63(11):3708–3717.

Kosfeld, M. and Rustagi, D. (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., and Rand, D. (2015). Promoting cooperation in the field. *Current Opinion in Behavioral Sciences*, 3:96–101.

Lavetti, K. and Schmutte, I. M. (2016). Estimating compensating wage differentials with endogenous job mobility.

Linardi, S. and McConnell, M. A. (2011). No excuses for good behavior: Volunteering and the social environment. *Journal of Public Economics*, 95(5-6):445–454.

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

MacLeod, W. B. (2007). Reputations, relationships, and contract enforcement. *Journal of economic literature*, 45(3):595–628.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more t in experiments. *Journal of development Economics*, 99(2):210–221.

Melamed, D., Simpson, B., and Abernathy, J. (2020). The robustness of reciprocity: Experimental evidence that each form of reciprocity is robust to the presence of other forms of reciprocity. *Science Advances*, 6(23):eaba0504.

Milgrom, P. and Roberts, J. (1995). Complementarities and fit strategy, structure, and organizational change in manufacturing. *Journal of accounting and economics*, 19(2-3):179–208.

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

Nowak, M. A. and Sigmund, K. (1992). Tit for tat in heterogeneous populations. *Nature*, 355(6357):250–253.

Nowak, M. A. and Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, 393(6685):573–577.

Nowak, M. A. and Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437(7063):1291–1298.

Ohtsuki, H. and Iwasa, Y. (2006). The leading eight: social norms that can maintain cooperation by indirect reciprocity. *Journal of theoretical biology*, 239(4):435–444.

Olson, M. (1965). *Logic of collective action: Public goods and the theory of groups (Harvard economic studies. v. 124)*. Harvard University Press.

Organ, D. W. (2018). The roots of organizational citizenship. *The Oxford handbook of organizational citizenship behavior*, pages 169–184.

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

Peperkoorn, L. S., Becker, D. V., Balliet, D., Columbus, S., Molho, C., and Van Lange, P. A. (2020). The prevalence of dyads in social life. *PloS one*, 15(12):e0244188.

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):122.

Rabin, M. (1993). Incooperating fairness into game theory and economics. *The American Economic Review*, 83(5):1281–1302.

Raihani, N. (2021). *The social instinct: how cooperation shaped the world*. Random House.

Rand, D. G. and Nowak, M. A. (2013). Human cooperation. *Trends in cognitive sciences*, 17(8):413–425.

Rasmusen, E. and Zenger, T. (1990). Diseconomies of scale in employment contracts. *The Journal of Law, Economics, and Organization*, 6(1):65–92.

Roberts, G. (2008). Evolution of direct and indirect reciprocity. *Proceedings of the Royal Society B: Biological Sciences*, 275(1631):173–179.

Santos, F. P., Pacheco, J. M., and Santos, F. C. (2021). The complexity of human cooperation under indirect reciprocity. *Philosophical Transactions of the Royal Society B*, 376(1838):20200291.

Schein, E. H. (2010). *Organizational culture and leadership*, volume 2. John Wiley & Sons.

Schmid, L., Chatterjee, K., Hilbe, C., and Nowak, M. A. (2021). A unified framework of direct and indirect reciprocity. *Nature Human Behaviour*, 5(10):1292–1302.

Sobel, J. (2005). Interdependent preferences and reciprocity. *Journal of economic literature*, 43(2):392–436.

Sugden, R. (1984). Reciprocity: the supply of public goods through voluntary contributions. *The Economic Journal*, 94(376):772–787.

Számadó, S., Balliet, D., Giardini, F., Power, E., and Takács, K. (2021). The language of cooperation: reputation and honest signalling.

Tadelis, S. (2016). Reputation and feedback systems in online platform markets. *Annual Review of Economics*, 8:321–340.

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific american*, 223(5):96–103.

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual review of psychology*, 33(1):1–39.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly review of biology*, 46(1):35–57.

Uzzi, B. (1997). Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative Science Quarterly*, 42(1):35–67.

van Apeldoorn, J. and Schram, A. (2016). Indirect reciprocity; a field experiment. *PloS one*, 11(4):e0152076.

Van Veelen, M., García, J., Rand, D. G., and Nowak, M. A. (2012). Direct reciprocity in structured populations. *Proceedings of the National Academy of Sciences*, 109(25):9929–9934.

Viscusi, W. K. and Aldy, J. E. (2003). The value of a statistical life: a critical review of market estimates throughout the world. *Journal of risk and uncertainty*, 27:5–76.

Williamson, O. E. (1967). Hierarchical control and optimum firm size. *Journal of political economy*, 75(2):123–138.

Wu, J., Balliet, D., and Van Lange, P. A. (2016). Reputation, gossip, and human cooperation. *Social and Personality Psychology Compass*, 10(6):350–364.

Yoeli, E., Hoffman, M., Rand, D. G., and Nowak, M. A. (2013). Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the National Academy of Sciences*, 110(supplement_2):10424–10429.

Zenger, T. R. (1994). Explaining organizational diseconomies of scale in r&d: Agency problems and the allocation of engineering talent, ideas, and effort by firm size. *Management science*, 40(6):708–729.

# Online Appendix

## A.1 Numerical Checks on Conditions in our Simple Model

### A.1.1 Choice of Becoming an Observer

*Is $r(.)$ "sufficiently large" in our setting?* Given that costs are 5% of workers' time and thus 5% of wage (if the wage is assumed to be the relevant opportunity cost), then $r(.)$ needs to be higher than 5% (this is conservative because the learning safety benefit $p\bar{b}d$ may still be meaningful for the observer even when diffusion, and thus contact rate, is very low). Table A.1 of section A.2 shows that job security increases by 2.1 percentage points (pp) for observers that are part of the committee (p-value<0.001) (but it doesn't for later joining observers) and this could go up to 3 or 4 pp if they perform many observations. Considering that $r(.)$ has other benefits such as an increase in promotion likelihood, enhanced status among colleagues, and help from colleagues, among others, it is plausible for $r(.)$ to exceed 5% of wages for the first observers.

*Is the threshold likely to occur in our setting?* Let's estimate Equation (2). Assume a value of $p = 0.5$ (notice that in the subsection A.3.3, we estimate that $p = 0.64$ in our field data). Consider a worker at the end of the second year of BAPP when diffusion is considerable (close to 30%), the contact rate is close to 1 (see panel b of Figure 1, and subsection A.3.1), and the impact of BAPP is a 30% accident reduction (in line with the estimates in panel a of Figure 1). According to estimates of compensating wage differentials literature (Viscusi and Aldy 2003, Lavetti and Schmutte 2016), the average "value" of one accident in one year across studies is 200% of a yearly wage. Assuming a yearly accident rate of 5%, the first term of Equation (2) is equal to 1.5% of a monthly wage (=200% x 12 x 5% / 12 x 30% x 0.5). Given that cost is equivalent to 5% of wage, $r(.)$ needs to be smaller than 3.5% for this worker to abstain from becoming an observer; this is very plausible at 20% diffusion. For example, Table A.1 of section A.2 shows that improving job security from becoming an observer is not distinguishable from zero for later entrant observers.

*How is the choice of becoming an observer a social dilemma?* Cooperation is dominant for initial observers. However, it is not after a certain diffusion threshold. To show that after this threshold having collective cooperation would be desirable, it suffices to show that when diffusion is 100% (and thus $r(.)$ converges to 0), the benefits from receiving observations from others surpass the costs, that is, $(1 + p)\bar{b} - c_0 b - c_T > 0$ (for now we assume that $f(.)$ is negligible given that $\widetilde{n} = n$). Assuming symmetric agents and that the share of $c_T$ over total costs becomes small over time, then observations are collectively desirable as long as $(1 + p) > c_0$, that is, the benefits of being observed are higher than the cost of observing. Empirically, assume symmetric agents and suppose that the contact rate becomes 2 at full diffusion (that is $\bar{b} = b = 2$, meaning that the average effort goes down sharply in comparison

to the initially observed level of 5) so that the impact of BAPP jumps proportionately from the 30% used in the previous paragraph to 60%. Then, the expression $(1 + p)$, that is, the benefits of being observed, would be equal to 9% (=200% x 12 x 5% / 12 x 60% x 1.5), which is larger than the cost of 5% (actually, as long as $\bar{b} = b > 5/4.5 = 1.11$ collective cooperation remains beneficial). (If we add to this calculation the private benefits represented by $f(.)$, however small it might be given $\widetilde{n} = n$, the cost $c_0$ would further be surpassed.) However, given that the private incentives are to not become an observer after the critical diffusion threshold, collective cooperation is not reached. Thus, we have a relaxed social dilemma with cooperation dominant up to a point.

### A.1.2 Choice of the Number of Observations

*How is the choice of the number of observations a social dilemma?* Notice that even though $b^*$ is individually optimal, the presence of $p\bar{b}d$ in Equation (3) means that if the impact of the contact rate is large enough, then a higher $b$ is beneficial for all. Formally, assume symmetric agents and an exogenous increase of $\delta$ in the number of observations for all observers; if $pd\delta > [r(b^* + \delta, \hat{d}_k) - r(b^*, \hat{d}_k)] - c_0\delta$ for all $k$, then there is a social dilemma (we assume that $f(.)$ is negligible given that $\widetilde{n} = n$). If $p$ and $d$ are large enough, this inequality is satisfied for all observers. This inequality becomes easier to satisfy as diffusion increases and for later entrant observers because $r(b^* + \delta, \hat{d}_k) - r(b^*, \hat{d}_k)$ becomes smaller when $k$ is larger (we assumed that $\dfrac{\partial^2 r}{\partial b \partial \hat{d}_k} < 0$), and the left-hand side of the inequality grows larger with $k$.

## A.2 Archival Data Analysis

### A.2.1 Descriptive Statistics of BAPP

The following equation defines three terms:

$$\textit{Contact rate} = \frac{\text{observations}}{\text{workers}} = \frac{\text{observations}}{\text{observers}} \times \frac{\text{observers}}{\text{workers}} = \textit{“effort”} \times \textit{“diffusion”} \qquad \text{(A.1)}$$

"Contact rate" is the number of observations per worker at a site in a given month. The contact rate drives safety (measured as accident reduction). It can be broken down into two components representing the two cooperation instances: "effort," or the average cooperative effort by observers in the site, and "diffusion," or the share of workers that are observers (i.e., the expansion of cooperation).

Using archival data of 88 BAPP implementations, Figure A.1 displays some important descriptive statistics—the average and 25th and 75th percentiles for contact rate, effort, and diffusion at the site level—for the first 36 months of BAPP implementation. The contact rate (green) approaches 1 by the end of year 3, but there is considerable variation across sites (dotted green lines). Effort (red) shoots up and then slightly decays, with a 10% decrease

from ≈5.3 observations per observer month in the first year to 4.8 in the third year. Variation is also high (red dotted lines): the 25th percentile displays around three observations, while the 75th percentile achieves 6.5. Diffusion increases from 4% in the first couple of months to 21% in the last months of the third year. An average of 245 workers per site in our sample translates to a change from ≈10 observers to ≈50 observers over 36 months.

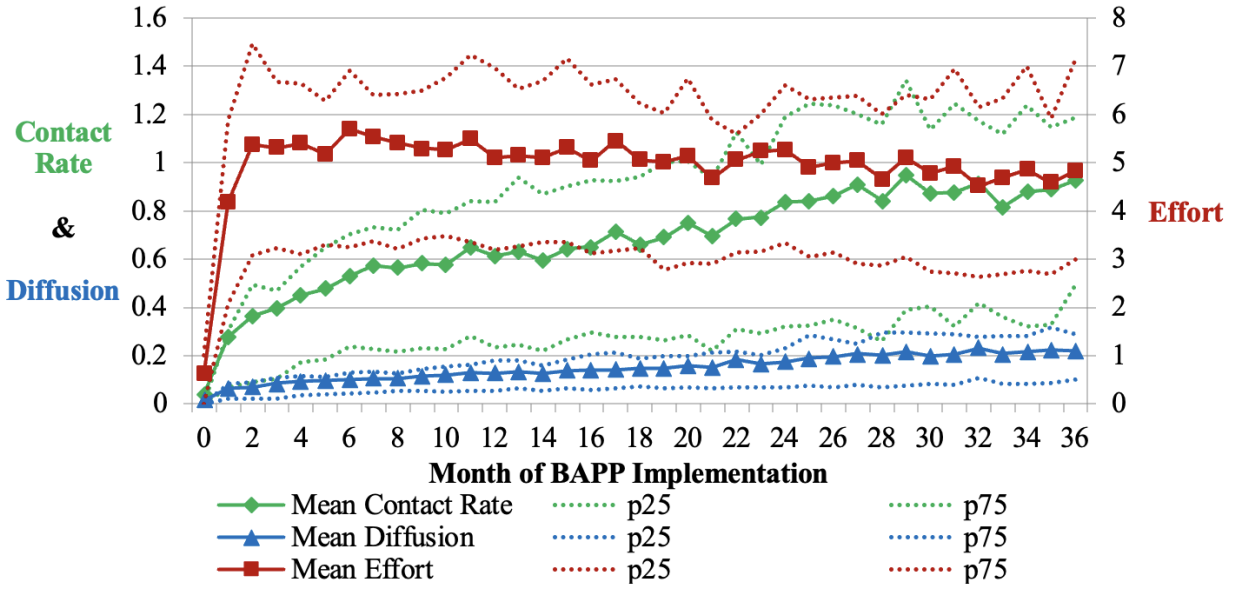



Figure A.1: Evolution of contact rate, effort, and diffusion over BAPP implementation. Note for figure: Dotted lines are $25^{th}$ and $75^{th}$ percentile.

### A.2.2 Verifying the Weakening Reputational Benefits Using Job Security

Hypothesis 1 depends on the assumptions of the function $r(b^*, \hat{d}_k)$. The assumptions are $\frac{\partial r}{\partial \hat{d}_k} < 0$ (reputation benefits are larger for early observers who sign up when BAPP is still risky), $\frac{\partial r}{\partial b} > 0$ and $\frac{\partial^2 r}{\partial b^2} < 0 < 0$ (there are positive but decreasing reputation benefits from effort), and $\frac{\partial^2 r}{\partial b \partial \hat{d}_k} < 0$ (the reputational return for effort is higher for early observers). We used the Chilean company Sodimac data to test these assumptions, where we executed our experiment (see the "Data and Field Experiment" section in the main body). We do this by estimating the following model for worker $i$ at store $j$ and month $t$:

$$Job\ Security_{ijt} = b_1 + b_2 BAPP_{jt} + \sum_j b_{3j} BAPP_{jt} \times Observer_{ijt} + X_{ijt} + U_i + T_t + \varepsilon_{ijt} \quad \text{(A.2)}$$

$Job\ Security_{ijt}$ is a dummy set to 1 if worker $i$ was working in any store in month $t+1$ (across all Sodimac stores) and to 0 otherwise. A reputation as a prosocial person who helps colleagues should be reflected by enhanced job security. $Observer_{ijt}$ is a dummy set to 1 if worker $i$ working at store $j$ is an observer in period $t$. $BAPP_{jt}$ is a dummy set to 1 when BAPP starts at store $j$ and to 0 before then. The $X_{ijt}$ vector of worker controls includes sex, age, tenure in the store, and a dummy for the job position (e.g., salesperson, cashier, or warehouse). We also include a worker fixed effect $U_i$ and month fixed effect $T_t$. Table A.1 presents the results.

Table A.1: Impact of becoming observer on job security

| | Dependent variable: Job security | | | | | | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| BAPP | 0.007 | 0.006 | 0.005 | 0.005 | 0.005 | 0.005 | 0.005 |
| | | (0.006) | (0.006) | (0.006) | (0.006) | (0.006) | (0.006) |
| BAPP × Observer | | 0.020*** | 0.014*** | | 0.001 | -0.007† | |
| | | (0.005) | (0.005) | | (0.012) | (0.022) | |
| BAPP × Observer × Starting team observer | | | | 0.021*** | | | 0.016*** |
| | | | | (0.002) | | | (0.003) |
| BAPP × Observer × New observer | | | | 0.003 | | | -0.037 |
| | | | | (0.013) | | | (0.047) |
| BAPP × Observer × Effort | | | | | 0.003* | 0.006† | |
| | | | | | (0.002) | (0.006) | |
| BAPP × Observer × Effort$^2$ | | | | | | -0.0002† | |
| | | | | | | (0.0004) | |
| BAPP × Starting team observer × Effort | | | | | | | 0.001** |
| | | | | | | | (0.000) |
| BAPP × New observer × Effort | | | | | | | 0.009 |
| | | | | | | | (0.009) |
| Time fixed effects? | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Worker fixed effects? | No | No | Yes | Yes | Yes | Yes | Yes |
| Store fixed effects? | Yes | Yes | No | No | No | No | No |
| Individual controls? | Yes | Yes | No | No | No | No | No |
| R-squared | 0.024 | 0.024 | 0.245 | 0.245 | 0.245 | 0.245 | 0.245 |
| Observations | 29,054 | 29,054 | 29,054 | 29,054 | 29,054 | 29,054 | 29,054 |
| Mean of dependent variable before BAPP | 0.969 | 0.969 | 0.969 | 0.969 | 0.969 | 0.969 | 0.969 |

**Note:** All regressions are estimated with OLS. Errors in parentheses: robust and clustered at the observer level. * p $<$ 0.1,** p $<$ 0.05, *** p $<$ 0.01. † is significant at p $<$ 0.001 in a joint t-test of $\partial(Job\ Security_{ijt})/\partial(Treatment1)$.

Column (1) shows that BAPP alone does not increase job security. Column (2) shows

that being an observer is associated with a statistically significant 2 percentage points (pp) increase in job security. In column (3), we refine the estimate by adding worker-fixed effects, finding that being an observer is associated with a 1.4 pp ($p < 0.001$) increase in job security. This effect is economically significant because the monthly turnover rate is 3.1% (= 1 – 0.969). Column (4) supports the assumption $\frac{\partial r}{\partial \hat{d}_k} < 0$. The coefficient of the interaction term suggests that the job security benefits are concentrated on the observers in the starting committee: they enjoy a 2.1 pp increase in job security while the remaining observers see no increase. Columns (5) and (6) support assumptions $\frac{\partial r}{\partial b} > 0$ and $\frac{\partial^2 r}{\partial b^2} < 0$, respectively: (5) shows that observations increase job security and (6) shows that this relationship is concave. Column (7) supports assumption $\frac{\partial^2 r}{\partial b \partial \hat{d}_k} < 0$ as it shows that effort is statistically related to increased job security only for initial observers and not for new observers that enter later.

### A.2.3 Verifying the Impact of BAPP on Accidents Using Archival Data

Our theory requires that BAPP observations actually generate a benefit to workers; in terms of our formal model, the term $\bar{b}d$ of Equations (1) and (2) (in the main body) has to generate safety benefits. In this section, we carry out this analysis using the archival panel data at the project level. We show that BAPP is related to a significant decrease in accidents during the first year of BAPP and that this impact is likely to be causal. We study BAPP's impact on accidents with the following model:

$$Accidents_{it} = b_1 + b_2 BAPP_{it} + b_3 Trend_{it} + b_4 \left(BAPP_{it} \times Trend_{it}\right) + b_5 ln(Workers_{it}) + U_i + \varepsilon_{it} \quad \text{(A.3)}$$

Equation A.3 models the accidents at site $i$ in month $t$. $BAPP_{it}$ is a variable that takes the value of 1 in the month when the first observation is executed at the site. $Trend_{it}$ equals $(t \dot{-} \theta_i)$, where $t$ is the month and $\theta_i$ is the month when 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 to control for time-invariant store unobservables. As a control, we add the natural logarithm of workers, as more workers cause more accidents.[4] 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.

Table A.2 displays the results. Column (1) indicates that BAPP is significantly associated with fewer accidents. Column (2) shows that the $Trend$ is negative and statistically significant. At the same time, BAPP now loses its statistical significance due to collinearity

[4]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; indeed, they became slightly stronger.

Table A.2: Impact of BAPP on accidents

| | Accidents – OLS (1) | Accidents – OLS (2) | Accidents – OLS (3) | Accidents - POIS (4) |
|---|---|---|---|---|
| BAPP | -0.357*** | -0.162† | -0.198*† | -0.156*† |
| | (0.087) | (0.104) | (0.115) | (0.085) |
| Trend | | -0.007*† | 0.001† | -0.001† |
| | | (0.004) | (0.007) | (0.005) |
| BAPP × Trend | | | -0.011† | -0.011† |
| | | | (0.009) | (0.007) |
| ln(Workers) | 1.030*** | 1.028*** | 1.028*** | 0.714*** |
| | (0.300) | (0.306) | (0.302) | (0.088) |
| Site fixed-effect? | Yes | Yes | Yes | Yes |
| Constant | -4.171** | -4.241** | -4.149** | |
| | (1.61) | (1.61) | (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 |

**Note:** 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 × Trend is also statistically significant at $p < 0.05$.

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 $BAPP \times Trend$ is below 0.001; a joint t-test for BAPP and $BAPP \times Trend$ is significant at 5% (the variance inflation factor is above 6 for these variables). Column (4) displays Poisson fixed effect estimates for 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 of 0.2 accidents and, regarding the slope, with a decrease of 0.132 accidents after 12 months. This is economically significant: at the end of the first year, BAPP is associated with an overall decrease of 30% in accidents.

These estimates are vulnerable to an 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 an out-of-sample analysis in Sodimac (the site of our experiment), where we estimate the analysis at the worker level (see the next section of the Appendix for this analysis). We execute the placebo test using the following model:

$$Accidents_{it} = b_1 + \sum_j (\pi_j Year\_BAPP\_P_j \times BAPP\_P_{jt}) + b_3 Trend_{it}$$
$$+ \sum_j (\rho_j Year\_BAPP\_P_j \times BAPP\_P_{jt} \times Trend_{it}) + b_5 ln(Workers_{it}) + U_i + \varepsilon_{it} \quad (A.4)$$

In this model, we force BAPP to start one year earlier and then flexibly estimate its impact in the four subsequent years. $BAPP\_P_{jt}$ 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 the month -11). $Year\_BAPP\_P_j$ 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.)

Table A.3: Placebo test on the impact of BAPP

| | Accidents – OLS (1) |
|---|---|
| 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 |

**Note:** 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 × Trend is also statistically significant at $p < 0.05$.

Essentially, this model breaks down the impact of BAPP on the level and slope into four parts, including the placebo year, one year before the actual start. 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 (and the year preceding that). The coefficient $b_3$ now identifies the trend from -24 to -12 in the months. Table A.3 presents the

estimates of Equation (A4). Interpreting this table can be tricky, so we graph the result in Figure A.2. This figure shows no effect in the year before BAPP, neither at the level nor slope.

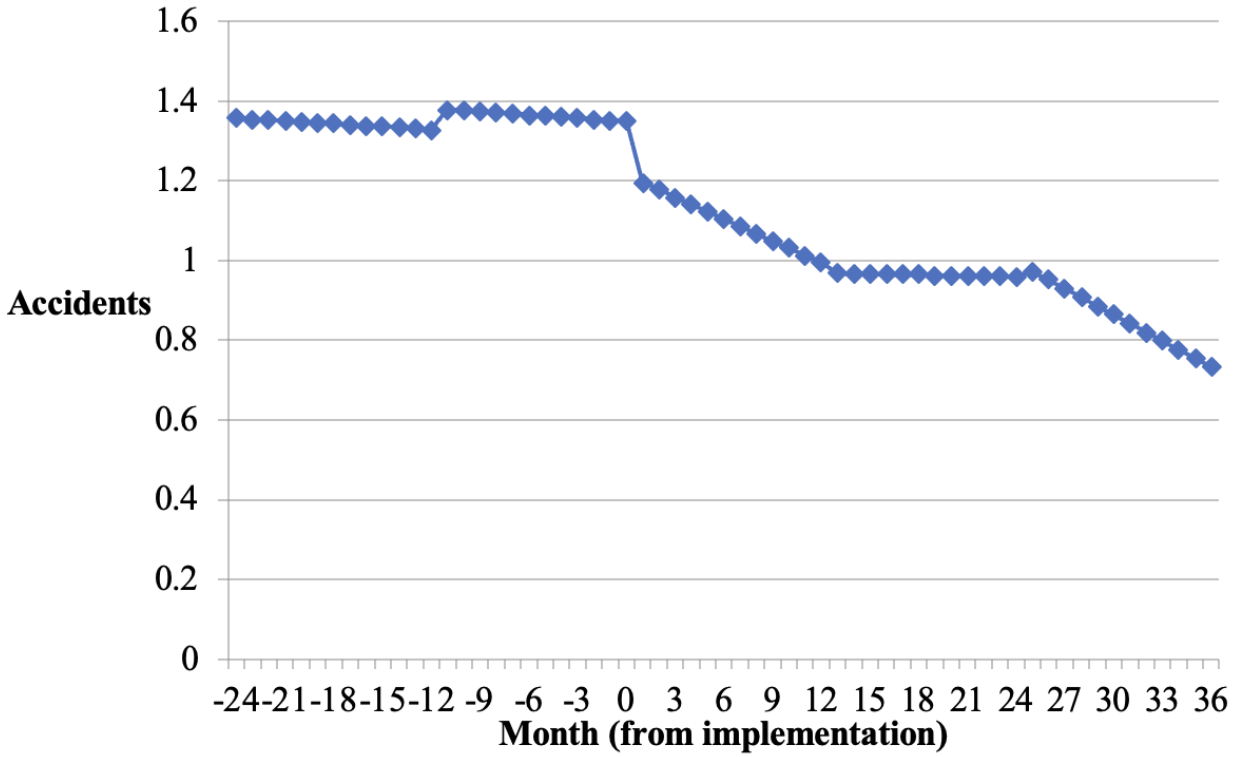



Figure A.2: Impact of BAPP in placebo year

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

$$Accidents_{it} = b_1 + b_2 BAPP_{it} + b_i Trend_{it} + b_4 \left(BAPP_{it} \times Trend_{it}\right) + b_5 ln(Workers_{it}) + U_i + \varepsilon_{it} \tag{A.5}$$

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

$$\Delta Accidents_{it} = a_1 + b_2 \Delta BAPP_{it} + b_i + b_4 \Delta \left(BAPP_{it} \times Trend_{it}\right) + b_5 \Delta ln(Workers_{it}) + \Delta\varepsilon_{it} \tag{A.6}$$

The results are displayed in Table A.4. In column (1), we find that BAPP decreases their coefficients, both at a similar level (from -0.198 to -0.056) and the slope (from -0.011 to -0.008) as in Table A.3. Statistical significance suffers in these models, as models in difference are noisier (see the R-squared).

Controlling for site-specific trends could also capture the quality of the BAPP implementation. The coefficients $b_2$ and $b_4$ capture the average impact of BAPP; thus, $b_i$ may capture the variation in the quality of the BAPP implementation. This implementation quality is a time-variant that is unobservable at the site level. Therefore, the estimates of (A5) could be biased depending on the rarity of the different extremes of implementation quality. Thus, columns (2), (3), and (4) 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 its statistical significance recovers. This is suggestive that the extreme values of time-variant unobservables are tilted toward cases that are not favorable to safety; for example, there may be more extreme cases of low implementation quality than high.

Table A.4: Impact of BAPP adding a site-specific trend as control

| | Δ Accidents (1) | Δ Accidents (2) | Δ Accidents (3) | Δ Accidents (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 $b_i$ |
| Δ BAPP | -0.056 | 0.066 | 0.197 | 0.065 |
| | (0.189) | (0.180) | (0.174) | (0.189) |
| Δ (BAPP x Trend) | -0.008 | -0.017 | -0.022* | -0.025** |
| | (0.013) | (0.014) | (0.013) | (0.009) |
| Δ ln(Workers) | 1.317** | 1.274* | 1.268 | 1.755* |
| | (0.609) | (0.719) | (0.799) | (0.971) |
| Site fixed-effect? ($b_i$) | Yes | Yes | Yes | Yes |
| Constant | -0.000 | 0.003 | 0.004 | 0.008 |
| | (0.008) | (0.008) | (0.007) | (0.006) |
| R-square | 1.54% | 1.44% | 1.45% | 5.9% |
| Observations | 4,748 | 4,199 | 3,776 | 2,773 |

**Note:** 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 the OLS panel fixed effect.

### A.2.4 Verifying the Impact of BAPP on Accidents Using Archival Data from Sodimac

We replicated the analysis of Table A.2, using archival data on BAPP implementations at Sodimac, the setting of our field experiment. We estimated the following model:

$$Accidents_{ijt} = b_1 + b_2 BAPP_{ij} + b_3 \left(BAPP_{ij} \times Time\_Elapsed_{ij}\right) + b_4 Observer_{ijt} + X_{it} + \tau_t + \gamma_j + u_{ijt} \quad \text{(A.7)}$$

$Accidents_{ijt}$ is a dummy that takes the value of 1 if worker $i$ at store $j$ experienced an accident in month $t$, and 0 otherwise. The variable $BAPP_{ij}$ takes the value of 1 in the month when observations start and zero before that. The variable $Time\_Elapsed_{ij}$ is a count variable with value 0 before BAPP and then 1, 2, 3, etc., for each month elapsed in the BAPP implementation of a site. Coefficient $b_2$ captures the impact on the level at time 0, while $b_3$

captures whether the impact of BAPP builds up over time. $X_{it}$ is the same vector of controls as in the analysis of the probability of becoming an 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 turnover is 5% a month. Therefore, if we had included them, we would have measured the impact only on a subset of workers present before and after and not the whole population subject to BAPP. $Observer_{ijt}$ is a dummy identifying that a worker is an observer after it becomes one: this variable captures the indirect impact of BAPP through observers' behavior. It could be that all the impact of BAPP on accidents is concentrated in lower accidents of observers and not the general workforce. We estimate this model using the four sites of our experiment between January 2016 and May 2018, and we consider only workers subject to BAPP observations. Table A.5 presents the results.

We find that BAPP reduces work accidents over time in Sodimac and that this effect is focused on work accidents without lost time, as expected.[5] The impact is large: BAPP is associated with a total reduction of 0.0015 work accidents per month in the first year or 35% of the variable's mean. This effect size is similar to the one estimated with archival data in the previous section of this Appendix. We also find that being an observer by itself reduces accidents, supporting the parameter $p$ of our model. Being an observer is associated with a reduction of -0.14% in the likelihood of an accident with lost working days; this is equal to 64% (= - 0.14 / (100 x - 0.0022) of the impact of Treatment 1 displayed the main body of the manuscript. Thus, one could approximate parameter $p$ by 0.6.

### A.2.5 Details of the Observers' Cohorts

We use the information at the observer-month level to generate the cut-offs of the quintiles/cohorts. This leads to Figure A.3 below. For example, in period 12, there are, on average, 30 active observers per site coming from the following cohorts:

1. 7 observers from the 1st quintile (observers with an entry order between 1 and 13),
2. 6.7 observers from the 2nd quintile (observers with an entry order between 14 and 36),
3. 7.8 observers from the 3rd quintile (observers with an entry order between 37 and 78),
4. 6.3 observers from the 4th quintile (observers with an entry order between 79 and 168), and

[5] Two expectations are met: 1) We find no impact on commuting accidents (i.e., accidents that take place between home and the workplace) and quasi-accidents (i.e., 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); this acts as a falsification test, as we would not expect BAPP to generate an impact in these types of accident. 2) Finding no impact of accidents with lost workdays 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 (expensive) investments in better machinery/equipment and maintenance.

Table A.5: 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 | -0.0022 | 0.0000 | -0.0000 | -0.0014 | -0.0015 | 0.0015 | 0.0015 |
| | (0.0036) | (0.0036) | (0.0023) | (0.0023) | (0.0019) | (0.0019) | (0.0012) | (0.0012) |
| BAPP x Time elapsed | -0.0016* | -0.0016* | -0.0015*** | -0.0015*** | -0.0011*** | -0.0011*** | -0.0004 | -0.0004 |
| | (0.0008) | (0.0008) | (0.0006) | (0.0006) | (0.0004) | (0.0004) | (0.0003) | (0.0003) |
| Observer | | -0.0007 | | -0.0004 | | 0.0011 | | -0.0014*** |
| | | (0.0031) | | (0.002) | | (0.0019) | | (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 | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| BAPP | 0.00013 | 0.0001 | -0.0019 | -0.0018 | 0.039 | 0.040 |
| | (0.019) | (0.0019) | (0.0021) | (0.0021) | (0.036) | (0.036) |
| BAPP x Time elapsed | 0.0002 | 0.0002 | -0.0004 | -0.0004 | 0.001 | 0.001 |
| | (0.0004) | (0.0004) | (0.0005) | (0.0006) | (0.014) | (0.015) |
| Observer | | 0.0008 | | -0.0013 | | -0.030 |
| | | (0.0019) | | (0.0014) | | (0.027) |
| Accident with lost time | | | | | 13.382*** | 13.382*** |
| | | | | | (2.905) | (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) |

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

5. 2.2 observers from the 5th quintile (observers with an entry order above 169).

These data suggest that the rotation of observers increases with the cohorts. In the 12th month, the first and second quintiles have roughly seven active observers, but the pool of the former is much smaller, with 13 observers compared to 23 (36-14+1). The same happens with higher cohorts: newer observers leave BAPP at a higher rate than the first cohorts—cooperation seems to turn shakier with size.

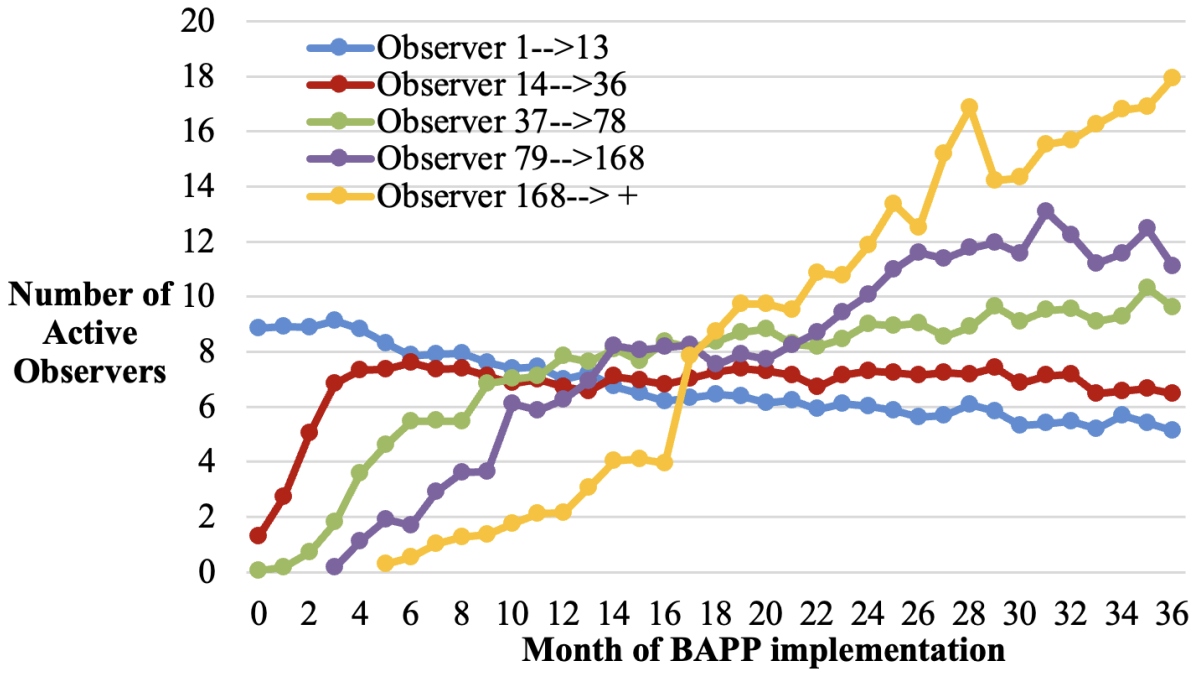



Figure A.3: Number of observers per quintile of entry (or cohort)

### A.2.6 Regression Showing that Effort Decays With the Number of Observers

The descriptive analysis displayed in Figure 1 of the manuscript indicating that effort goes down with diffusion is subject to confounders. For example, the lower effort of higher quintiles might be due to a higher diffusion rate: to achieve a predefined contact rate, low effort might be needed if diffusion is high. To check this, we use the following regression model:

$$Effort_{ijt} = b_1 + \sum_k b_{2k} Observer\_Q_{ik} + b_3 Total\_Observersjt + b_4 Tenure_{ijt} + T_t + U_j + \varepsilon_{ijt} \tag{A.8}$$

This model regresses the number of observations by observer $i$ at site $j$ in the month of implementation $t$(from 1 to 36) on the quintile of the observer (as defined earlier), the number of observers at the site (which captures diffusion), the tenure of the observer (measured as the months elapsed between the month of their first observation and the focal month) which controls for the impact of rotation (higher quintiles have higher rotation), and fixed effects of the 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 Table A.6. Column (1) shows that higher entry cohorts exert significantly lower levels of effort.

However, sites have different numbers of workers; therefore, using quintiles defined across

sites rather than within them is inexact. To accommodate this, columns (2) and (3) of Table A.6 use the observer order of entry within the site and this variable, in the presence of site (columns (2) and (4)) or site-month fixed effects (columns (3) and (5)) is not affected by such concerns. Column (5) suggests that the 50th observer by entry order within a site performs 0.95 fewer observations per month than the first, and the 100th observer performs 1.8 fewer observations.

Table A.6: Regression of effort on entry order

| | Effort | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| 1st quintile of entry order | 3.046*** | | | | |
| | (0.255) | | | | |
| 2nd quintile of entry order | 1.986*** | | | | |
| | (0.253) | | | | |
| 3rd quintile of entry order | 1.331*** | | | | |
| | (0.185) | | | | |
| 4th quintile of entry order | 1.082*** | | | | |
| | (0.127) | | | | |
| 5th quintile of entry order | - | | | | |
| | – | | | | |
| Order of entry | | -0.003*** | -0.006*** | -0.016*** | -0.02*** |
| | | (0.001) | (0.001) | (0.002) | (0.001) |
| Order of entry$^2$ | | | | 0.00002*** | 0.00002*** |
| | | | | (2.09e-06) | (2.34e-06) |
| Tenure | 0.023*** | 0.072*** | 0.045*** | 0.036*** | 0.006 |
| | (0.007) | (0.007) | (0.005) | (0.007) | (0.006) |
| Number of observers | -0.006*** | -0.006*** | - | -0.005*** | - |
| | (0.001) | (0.001) | - | (0.001) | - |
| Month of implementation fixed-effects? | Yes | Yes | No | Yes | No |
| Site fixed-effects? | Yes | No | No | Yes | No |
| Site X month of implementation fixed effects? | No | No | Yes | No | Yes |
| Constant | 1.921*** | 4.878*** | 1.021*** | 4.965*** | 1.052 |
| | (0.367) | (0.269) | (0.005) | (0.268) | Missing |
| R-square | 8.51% | 8.37% | 27.89% | 8.46% | 27.99% |
| Observations | 91,145 | 91,145 | 91,145 | 91,145 | 91,145 |
| Mean of dependent variable | 5.28 | 5.28 | 5.28 | 5.28 | 5.28 |

**Note:** Errors in parentheses are robust and clustered at the observer level. All models are estimated using OLS. $*\ p < 0.1,\ **\ p < 0.05, ***\ p < 0.01$.

### A.2.7 Willingness to be Observed Decays with Observer Number (Survival Analysis)

To study the willingness to be an observer, we study the likelihood that an observer remains as such after they become one. This type of analysis is known as a survival analysis. Using the observation level data, we constructed an observer-level data set that indicates whether the observer has resigned or whether s/he remains an observer by the end of the 36th-month windows (after the start of BAPP), and if it resigned, after how many months of tenure the resignation happened. With this data, we estimate the following econometric model, known as the Cox proportional hazards model:

$$H(n)_{ij} = H(0) \times Exp\{\sum_{jk} b_{2k} Observer\_Q_{ik} + U_j + \varepsilon_{ij}\} \tag{A.9}$$

$H(n)_{ij}$ is the hazard rate, that is, the likelihood that the observer $i$ in site $j$ resigns in the month of tenure n. $H(0)$ is a baseline likelihood of resignation. $Observer\_Q_{ik}$ captures the same as Equation (A8) above. Table A.7 displays the results. Column (1) indicates that the likelihood of resignation at any n grows with the quintiles of the order of entry (results are normalized to the 5th quintile being 1). In particular, it shows that the hazard ratio $\frac{H(n)}{H(0)}$ is multiplied by 1.25 (=exp(0.244)) when the observer is in the 1st quintile, by 1.53 (=exp(0.427)) when the observer is in the 2nd quintile, and so on, all the way to being multiplied by 2.72 (=exp(1)) in the 5th quintile.

Table A.7: Cox proportional hazard model

| | Likelihood of resignation (hazard ratio) | |
|---|---|---|
| | (1) | (2) |
| 1st quintile of entry order | 0.244*** (0.011) | 0.057*** (.004) |
| 2nd quintile of entry order | 0.427*** (0.017) | 0.132*** (0.007) |
| 3rd quintile of entry order | 0.609*** (0.021) | 0.265*** (0.012) |
| 4th quintile of entry order | 0.691*** (0.024) | 0.437*** (0.017) |
| 5th quintile of entry order | Base=1 (Omitted) | Base=1 (Omitted) |
| Site fixed-effects? | No | Yes |
| Observations | 9,196 | 9,196 |

**Note:** Errors in parentheses are robust. $***$ $p < 0.01$ in two-tailed test. We display the hazard ratio, that is, we estimate $\frac{H(n)}{H(0)} = Exp\{\sum_{jk} b_{2k} Observer\_Q_{ik} + U_j + \varepsilon_{ij}\}$.

### A.2.8 Accident Impact of Effort and Diffusion Estimated Using Archival Data

We examine how diffusion affects the impact of BAPP on accidents, exploiting the fact that a high contact rate can be achieved using different strategies: high effort and low diffusion, low effort, and high diffusion, or both at a medium level. BAPP does not impose an execution strategy: sites decide, leading to naturally occurring variance across implementations. Figure A.4 displays all the month-site combinations of diffusion and effort for the three years of BAPP implementation. In red, we display a site that achieved a high contact rate by increasing 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. We exploit this strategy variation to isolate the impacts of effort and diffusion. We use the following model:

$$\begin{aligned} Accidents_{it} &= b_1 + b_2 BAPP_{it} + b_3 Trend_{it} + \sum_j b_{4j} BAPP_{it} \times Q\_Effort_{jt} \\ &\quad + \sum_j b_{5j} BAPP_{it} \times Q\_Diff_{jt} + b_6 ln(Workers)_{it} + U_i + \varepsilon_{it} \end{aligned} \quad \text{(A.10)}$$

This model is the same as equation A.3, but we "break down" the impact of BAPP across two sets of five quintiles of effort and diffusion.

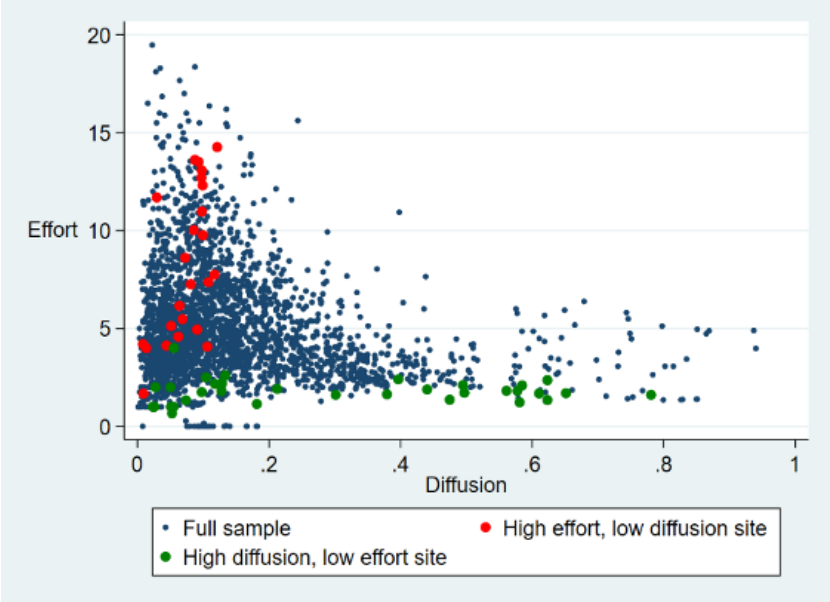



Figure A.4: Variation across effort and diffusion

Table A.8 presents the estimation of Equation (A10). The results indicate that increases in effort unambiguously decrease accidents (graphically shown in Figure A.5). 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 a joint t-test; see Table A.18 below). Adding the control of BAPP times TREND in column (2) does not change the results.

Table A.8: The role of effort and diffusion in the impact of BAPP

| | Accidents | |
|---|---|---|
| | (1) | (2) |
| BAPP | 0.016 (0.149) | -0.039 (0.152) |
| BAPP X $1^{ST}$ × Quintile of Effort | - | - |
| BAPP X $2^{ND}$ × Quintile of Effort | -0.113 (0.089) | -0.118 (0.091) |
| BAPP X $3^{RD}$ × Quintile of Effort | -0.144 (0.101) | -0.147 (0.103) |
| BAPP X $4^{TH}$ × Quintile of Effort | -0.218* (0.126) | -0.226* (0.130) |
| BAPP X $5^{TH}$ × Quintile of Effort | -0.267** (0.117) | -0.266** (0.119) |
| BAPP X $1^{ST}$ × Quintile of Diffusion | - | - |
| BAPP X $2^{ND}$ × Quintile of Diffusion | -0.169 (0.119) | -0.144 (0.113) |
| BAPP X $3^{RD}$ × Quintile of Diffusion | -0.016 (0.110) | 0.015 (0.116) |
| BAPP X $4^{TH}$ × Quintile of Diffusion | 0.037 (0.096) | 0.084 (0.094) |
| BAPP X $5^{TH}$ × 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 |

**Note:** 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 columns (1) and (2), respectively.

### A.2.9 Diffusion Slows Down Over Time

Table A.9 shows a regression using the DEKRA archival data where we explore how diffusion changes over the months of implementation. The quadratic specification shows that diffusion increases, but this increase slows down over time. This is consistent with Hypothesis 1 ii) in the main body of the paper. The results show that the maximum diffusion occurs at the month 29.5 months $\left(= \frac{0.0059}{(-2 \times 0.0001)}\right)$. After this point, diffusion is reduced over time.

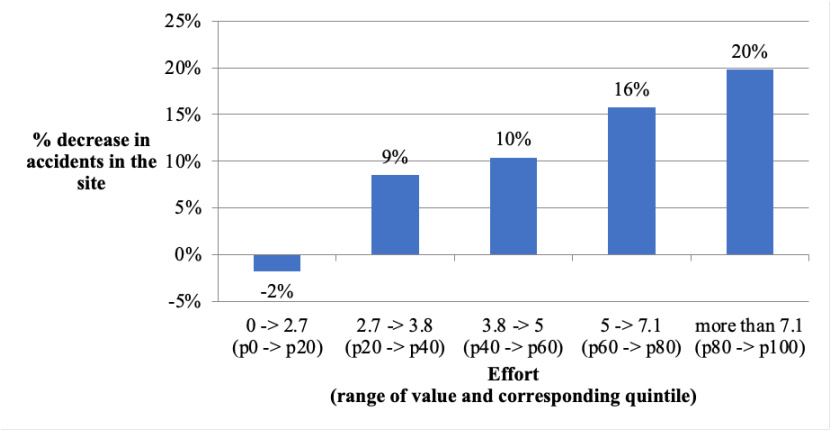



Figure A.5: The impact of BAPP varies according to Effort

**Note:** To build this graph, we plot the derivative of accidents over 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.

Table A.9: Diffusion slows down

| | Difussion |
|---|---|
| | (1) |
| Month of Implementation | 0.0059*** (0.0011) |
| Month of Implementation$^2$ | -0.0001** (0.0003) |
| Effort | 0.0002 (0.0005) |
| ln(Workers) | -0.166*** (0.043) |
| Site fixed-effect? | Yes |
| Constant | 0.9492*** (0.2290) |
| R-square | 64.12% |
| Observations | 2,696 |

**Note:** 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.3 Field Experiment

## A.3.1 Details of the pre-registered Field Experiment

**Review board and pre-registration.** The field experiment was revised and approved by the IRB of the Cambridge Judge Business School (while the first author was executing a

PhD in that institution). The experiment was pre-registered in July 2017 on the American Economic Association registry for randomized controlled trials (ID: AEARCTR-0002350).

**Date, firms, and location.** The three treatments were designed during the last quarter of 2016. We conducted the experiment in Chile in 2017 and 2018, collaborating with the Chilean Safety Association (ACHS) and Sodimac. ACHS is a large non-profit organization that provides occupational safety and health services (prevention, medical treatments, disability pensions, and subsidies). It partnered with DEKRA in 2012 to implement BAPP in its affiliated firms. DEKRA allocated permanent staff to ACHS to train and mentor a cadre of ACHS consultants, sharing handbooks, guidelines, IP, and software, thus enabling ACHS to deliver BAPP. Sodimac is a home-improvement-store company with operations across South America. Across Chile, it employs 20,000 employees and owns approximately 75 stores. A Sodimac store typically employs 200 to 350 workers. Sodimac had already implemented BAPP in five stores and a distribution center, starting in 2014. In 2017, they began implementation in four additional stores (Antofagasta, Temuca, La Reina, Huechuraba), staggering the start dates from June 2017 through October 2017 (see Table A.10 for details). We were allowed to modify these implementations from start to June 2018 experimentally.

**Implementation.** We created and followed a careful communication and implementation protocol regarding what to communicate, to whom, and when and how to implement Treatment 1 (which is the most on-site logistics-heavy treatment). Considering all four sites, we had approximately 20 observers in treatment and 20 observers in control before adding new observers, as well as 500 workers in treatment and 500 workers in control. The sites' observer groups grew until, in May 2018, there were 92 observers in total. See the SI for more details.

**Randomization.** The consultant performed a random selection of observers and their matching to groups using a lottery box in a starting team meeting in the fourth month. When there was an odd number of observers, the natural number below the median was used. The researchers randomized workers into groups beforehand, preparing the worker lists for distribution to observers upon lottery box results. The randomization of workers was stratified by sex, age, tenure, and task type. Before or during their first observation, each observer in this treatment handed a letter to their assigned workers. The letter, reproduced in the next section, briefly introduced BAPP and then suggested that the worker accept observations only from their assigned observer to enforce the groups. To avoid priming group identity (in contrast to Treatment 2), at no point was the notion of a group explicitly mentioned. This was emphasized to the consultants. Over time, as workers volunteered and became new observers, each was instructed to execute observations on the workers of their group of origin (either a specific treatment group or the control group at large). An updated letter was delivered to the workers informing them of the addition of the new observers.

### A.3.2 Treatment Documents

#### A.3.2.1 Letter handed out under Treatment 1

Estimado Colaborador,

En nuestra tienda estamos implementando la metodología BAPP cuyo propósito es ayudarnos a trabajar de forma segura, sin accidentes y enfermedades laborales.

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]

#### A.3.2.2 Letter handed out under Treatment 2 (the areas highlighted in grey are added to the letter)

Estimado Colaborador,

En nuestra tienda estamos implementando la metodología BAPP cuyo propósito es ayudarnos a trabajar de forma segura, sin accidentes y enfermedades laborales.

En esta metodología mi rol es ser tu "observador". Esto significa que de forma frecuente, por ejemplo una vez al mes, observaré cómo ejecutas tu trabajo, tomaré nota de lo observado y te entregaré retroalimentación. Si estás haciendo alguna tarea o actividad de forma insegura, intentaré hacértelo ver y podremos discutir cómo mejorar; si estás haciendo las tareas de forma segura, reforzaremos en conjunto la importancia mantener ese comportamiento en el futuro.

Todas las "observaciones" serán anónimas, tú nombre no quedará registrado en ninguna parte del proceso. Asimismo, yo seré tu único observador. Si algún otro observador se acerca por error a observarte, por favor indícale gentilmente que ya tienes un observador asignado.

Yo estaré haciendo observaciones a ti y a [NUMERO] otros trabajadores de la tienda. Más abajo encontrarás un listado con los trabajadores que forman parte este grupo. Hemos bautizado a este grupo con el nombre "[GRUPO NUMERO XX]".

Finalmente, es importante que sepas que TÚ también puedes ser un observador como yo. Si en el futuro decides serlo, yo te podré entrenar y podrás realizar observaciones a los mismos [NUMERO] trabajadores que yo observo (es decir, a los trabajadores del listado de abajo). Podremos trabajar codo a codo, ayudando a nuestro compañeros a trabajar de forma segura!

Si tienes cualquier duda o comentario, no dudes en contactarme.

Cordialmente,

[FIRMA DEL MIEMBRO DEL EQUIPO IMPLEMENTADOR]

[NOMBRE DEL MIEMBRO DEL EQUIPO IMPLEMENTADOR]

Observador asignado al "[GRUPO NUMERO XX]"

Integrantes del "[NOMBRE DEL GRUPO]"

| | NOMBRE COMPLETO | CARGO |
|---|---|---|
| 1 | xxx | xxx |
| 2 | xxx | xxx |
| 3 | xxx | xxx |
| 4 | xxx | xxx |
| 5 | xxx | xxx |
| ... | | |

### A.3.3 Report Used in Treatment 3

**Listado observadores y observaciones BAPP**

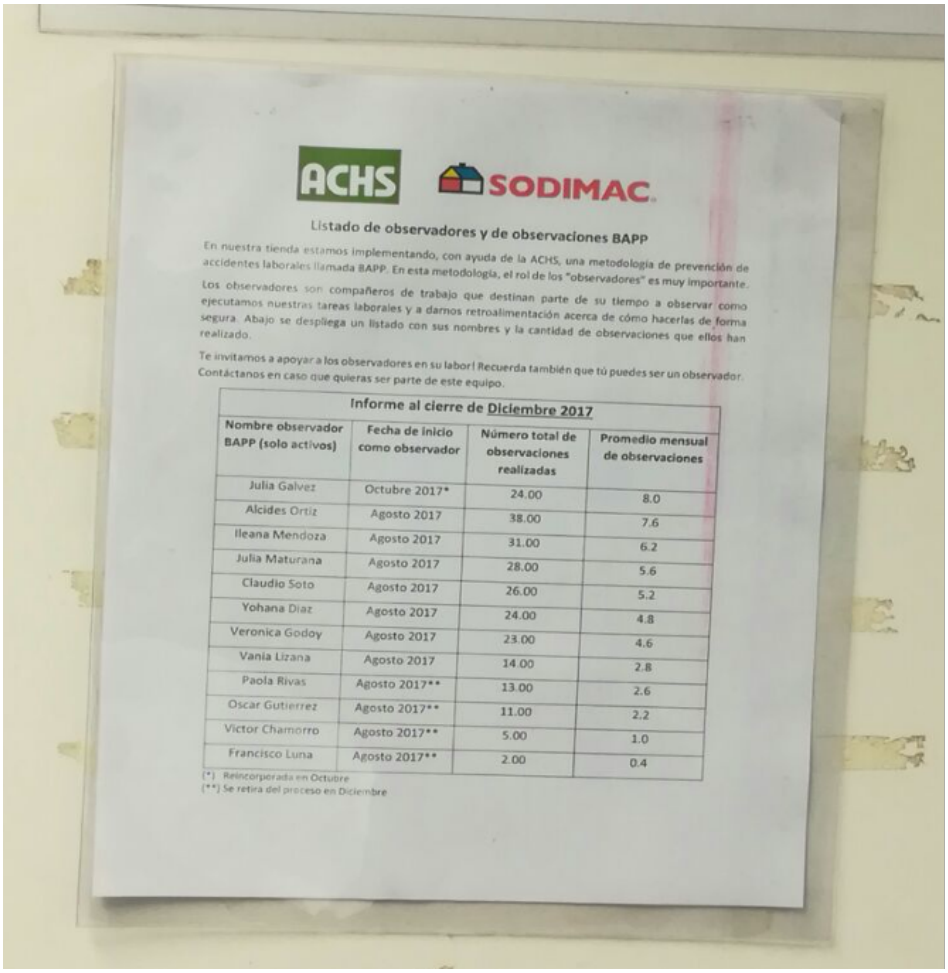
En nuestra tienda estamos implementando, con ayuda de la ACHS, una metodología de prevención de accidentes laborales llamada BAPP. En esta metodología, el rol de los "observadores" es muy importante.

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 |
|---|---|---|---|
| xxx | | | |
| xxx | | | |
| xxx | | | |
| ... | | | |

This is a sample picture that was used to certify the implementation of the treatment:



ACHS SODIMAC

Listado de observadores y de 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 de 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.

Informe al cierre de Diciembre 2017

| Nombre observador BAPP (solo activos) | Fecha de inicio como observador | Número total de observaciones realizadas | Promedio mensual de observaciones |
|---|---|---|---|
| Julia Galvez | Octubre 2017* | 24.00 | 8.0 |
| Alcides Ortiz | Agosto 2017 | 38.00 | 7.6 |
| Ileana Mendoza | Agosto 2017 | 31.00 | 6.2 |
| Julia Maturana | Agosto 2017 | 28.00 | 5.6 |
| Claudio Soto | Agosto 2017 | 26.00 | 5.2 |
| Yohana Diaz | Agosto 2017 | 24.00 | 4.8 |
| Veronica Godoy | Agosto 2017 | 23.00 | 4.6 |
| Vania Lizana | Agosto 2017 | 14.00 | 2.8 |
| Paola Rivas | Agosto 2017** | 13.00 | 2.6 |
| Oscar Gutierrez | Agosto 2017** | 11.00 | 2.2 |
| Victor Chamorro | Agosto 2017** | 5.00 | 1.0 |
| Francisco Luna | Agosto 2017** | 2.00 | 0.4 |

(*) Reincorporada en Octubre
(**) Se retira del proceso en Diciembre

Figure A.6: Sample picture of treatment 3

### A.3.4 Treatment Implementation

**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 to support a research project sponsored by all three partners: DEKRA, ACHS, and Sodimac. After each team was constituted, the same message was delivered to the enabler and the starting observers. In the 3rd month, the enabler and the team were also asked to answer a short and voluntary personality and social preferences survey (explained in subsection 5.5 of the main body). In the 4th month, treatments 1 and 3 were explained to them (the latter only to the two stores that received it). Importantly, the three consultants used the same PowerPoint slides for all these communications instances to convey the 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). The observers and workers were randomly allocated using a lottery box. Workers of the site had been pre-randomized and placed on lists containing the names of the workers in the treatment groups and the control group. These lists were prepared by the research team beforehand and sent to the consultant prior to their 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 particular set of workers to observe (i.e., groups were not biased).[6] 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 subsubsection A.3.2.1 above. 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; see subsubsection A.3.2.2.) This message of the letter also played a role in enforcing the groups' compliance as the implementation progressed. Each observer in the control group was also given a list containing all the workers who were not assigned to a group. The observers in the control group were supposed to observe workers only from this list. Stores experience a non-negligible turnover 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.10). 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. Table A.10 presents several statistics of the implementation of treatment 1.

[6]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), indicating 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, large provides low focus but high flexibility.

Table A.10: 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) The numbers of workers and observers we display are the averages of all the lists handed out on the implementation, including the observers in each group/control. (2) * After the starting team of observers was trained and assigned to a 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 quit (we probed whether the treatment caused this, but this was unclear as other elements were present in their decision). After it was clear who was not quitting, we corrected the lists as follows: if the observer who quit was part of a group, some of its workers were randomly assigned to the other groups; if the worker was part of the control, the control list was not changed. We did this to avoid excessive changes in lists and, given the enabler as a default in the control (who never quit), to be conservative on the sizing of groups (i.e., not to favor Treatment 1 with smaller groups). One example is Temuco. Originally, we had five groups and control and thus 11 observers (including enablers). 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 the treatment changed to 33.4 * 5 / 4 = 41.9 (3). ** If the update was in, for example, October, 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, for example, the 10th of November. (4) ‡ we compute the average without considering the months when the group had only one member (i.e., the starting team observer appointed to it). The average includes the starting team observer.

Each observer in the control group was also given a list containing all the workers who were not assigned to a group. The observers in the control group were supposed to observe workers only from this list. Stores experience a non-negligible turnover 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.10). 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. Table A.10 presents several statistics of the implementation of Treatment 1.

### A.3.5 Experimental Checks (power, balance, manipulation, exit interviews)

**Pre-experiment power calculations.** We calculated the effect size that our experiment would allow us to detect by assuming power of 80% and significance of 5%, using data on observations from the DEKRA dataset and on historical workplace accidents from Sodimac, including intra-class correlation estimates and power gains from having panel data (McKenzie 2012). We estimate that the minimum detectable effect is roughly 1 observation (equivalent to 44% of a standard deviation) and 0.009 workplace accidents (equivalent to roughly 12% of one standard deviation in workplace accidents).

**Positive predictive value (PPV).** In the main body of the paper, we found that Treatment 1 generated a reduction of 0.003 workplace accidents. This is small compared to the pre-experiment estimate of the minimum detectable effect (MDE) of 0.009 workplace accidents. To complement pre-experiment statistical power calculations, Ioannidis (2005) recommends calculating the positive predictive value (PPV). In our case, the PPV for Treatment 1 equals [0.2*R/(0.2*R+0.025)], where 0.2 is the power, 0.025 is the statistical significance associated with the estimate of 0.003 in accident reduction, and R is the ratio of "true relationships" to "no relationships" in the population of studies that test hypotheses such as the one we test in Treatment 1 (R can be very low in fully empirical and a-theoretical fields such as genome-disease association studies). Given that in the main body of the manuscript, we document an impact of Treatment 1 on the effort, which is above the minimum detectable effect, as well as an impact on behavior change; this provides a good *a-priori* credence for our hypothesis, and, thus, 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 to 0.66)

**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). Tables A.11 and A.12 show that the treatment groups and control are well balanced, indicating that the randomizations were appropriately executed.

Table A.11: 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.12: Balance check of observer randomization (starting team members), for all stores

| | Starting team members | | | Starting team members (excluding 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) |

**Exit interviews.** In June 2018, we visited the sites. We executed exit interviews with the consultant, the enabler, a group of three observers and three workers in Treatment 1, and a group of three observers and three workers from the control group. We executed a structured interview format, avoiding leading questions. The objective of these meetings was

to gather qualitative insights on the implementation of the treatments and the mechanisms that might have generated the results.

**Take-up survey (Manipulation check).** The lists of workers we distributed to observers (plus the letters to workers) might not have been sufficient to secure compliance. Therefore, 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 surveyed randomly drawn workers who 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.13 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 receiving 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 conducted by their assigned observer. This indicates that Treatment 1 was effectively implemented at stores, although not perfectly. Therefore, the impact of Treatment 1 represents a lower bound of the effect with 100% compliance. Regarding treatments 2 and 3, exit interviews indicated that there was awareness of the list of members and of the report on the board.

Table A.13: 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%)** |

**Note:** * Numbers in parenthesis restrict the count to respondents who acknowledge receiving the letter.

### A.3.6 Additional Evidence that Dyadic Repeated Interactions Drove Treatment 1

In the main body of the paper, we list several tests and arguments that enhance confidence that repeated interactions between observers and workers were the mechanism driving the results of Treatment 1. In this section, we provide the details of these tests and arguments.

**Workers' behavior.** In the main body of the paper, we show that workers display less risky behavior as a consequence of Treatment 1, as captured by observers in their observation sheets (in each observation, several behaviors are measured and assessed). This provides

evidence that workers reciprocate the observers' by complying with the advice provided.

**Fewer commuting accidents.** Table A.14 assesses the experiment's impact on other types of accidents. Regarding commuting accidents, Treatment 1 decreases these accidents, but the estimates are not very precise (p-value = 0.13 in column (2)). This also indicates that workers have internalized the observers' advice and changed their behavior even outside of work. This suggests that observations are reciprocated with the internalization of the advice and behavioral compliance.

We find no impact on quasi-accidents (columns (3) and (4)) and length of leave (columns (5) and (6)). These are two reassuring falsification tests as our treatment should not affect either: quasi-accidents are mostly non-accidents (usually injuries produced outside work, for example, due to sports activity) that are wrongly reported as workplace accidents, and length of leave which, like accidents with lost time, is not affected by behavioral interventions such as BAPP but by rather (expensive) investments in better machinery/equipment and maintenance.

Table A.14: Impact of BAPP on commuting, quasi-accidents and length of leave

| | Commuting accidents | | Quasi-accidents | | Length of leave | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Treat. 1 | -0.0006 | -0.0024‡ | 0.001 | 0.0022 | -0.056 | 0.0098 |
| | (0.0085) | (0.0016) | (0.001) | (0.0018) | (0.0347) | (0.0264) |
| Treat. 1 x Treat. 2 | | 0.0031* | | -0.0006 | | -0.103 |
| | | (0.0017) | | (0.0019) | | (0.0678) |
| Treat. 1 x Treat. 3 | | -0.0001 | | -0.0028 | | -0.0107 |
| | | (0.0016) | | (0.0018) | | (0.0549) |
| Accident with lost time | | | | | 12.978*** | 12.985*** |
| | | | | | (4.438) | (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 | 0.045 |
| | | | | | (12.97)† | (12.97) |

**Note:** Dependent variables: Commuting accidents occur between home and the workplace. Quasi-accidents are incidents that do not meet the conditions to be considered by ACHS, mostly because they are not workplace incidents but also because they are not meaningful or real incidents. Finally, in the case of workplace accidents with lost days, we also consider the length of leave.
We execute 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 the 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<0.15, * p<0.1,** p<0.05, *** p<0.01.

**Discarding self-selection.** The workers that become observers are not randomly selected, and this could introduce bias. There are two types of differences to consider: i) observers vs. workers and ii) starting team observers vs. new observers. The difference i) affects the interpretation of the impact of BAPP, whether it is a treatment effect or a selection effect driving the results. The difference ii) might be damaging to our analysis and interpretation of hypotheses 1 and 2. In particular, a strong alternative explanation to our findings would be that the starting observers have a higher $f(.)$ and that it is this asymmetry, not a reduction in $r(.)$ due to dwindling reputational benefits, that is driving the patterns we uncover from increased diffusion.

Table A.15 and Table A.16 evaluate the extent of this possible self-selection by comparing observables. Table A.15 compares observers and workers. We find that the 38 observers in the starting teams are older, have a longer tenure, and are more likely to be women than the rest of the workers at the site; however, we found no difference in terms of the type of job they do. Regarding the difference between the new observers and other workers, we found none across these four observables.

Table A.15: 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 | |

**Note:** *** p-value <0.01, *** p-value <0.05, *** p-value <0.1. We included all the workers employed while the experiment was being conducted. We lost three observers in the starting team, given that we filtered by the type of workers that were eligible for BAPP observations and to become new observers (not supervisors or managers). To make an apples-to-apples comparison, we dropped the cases of starting team members who were supervisors. The result does not change if we include these back.

Table A.16 compares observers of the starting team and new observers. Consistent with Table A.15, starting team observers are older, have a longer tenure, and are more likely to be women (but do not differ in the job). Using the survey mentioned introduced in subsection 5.5 of the main body, we found no differences between the starting team observers and the new observers across the "Big 5" personality traits: altruism, altruistic punishment, or the size of their social network. Given that new observers primarily drive our results, the absence of these differences mitigates self-selection concerns. In the main body, we explain how altruism and punishment are measured. "Big 5" questions are measured using a 1 to 5 Likert scale. For the social network, we asked workers to state with how many coworkers on the site they had a social relationship (i.e., acquaintance, friend).

Table A.16: Difference between starting team members and new observers

| | Observers members of the starting team Mean (S.D.) | New observers Mean (S.D.) | 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 (Altruism) | 4.68 (0.52) | 4.13 (0.40) | 0.450 |
| Third party punishment game (Punishment) | 0.69 (0.08) | 0.70(0.10) | 0.950 |
| Social network | 5.74 (0.86) | 6.09 (0.97) | 0.790 |
| Number | 34 | 23 | |

**Note:** *** p-value <0.01, *** p-value <0.05, *** p-value <0.1. S.D. stands for "standard deviation".

**Discarding "quality of starting team observers."** An alternative explanation is that the starting team observers assigned to Treatment 1 were simply of a "better quality." They generated more observations and guided the new observers. Out of chance, the random selection of observers might have ended up unbalanced in this unobservable aspect. To check for this, we controlled for the quality of starting team observers using a two-stage analysis.

We first computed fixed effects only for the starting team observers using the first five periods of implementation; then we plugged these fixed effects as control into a regression analogous to column (2) of Table 1 of the main body of the manuscript, but only using new observers from the sixth month of implementation onwards. Table A.17 shows the results. Column (1) reproduces results from the main body of the paper. Column (2) adds "observer quality," and column (3) shows the interaction between "observer quality" and Treatment 1.

Note that columns (2) and (3) only have new observers. Hence, the impact of Treatment 1 captures the impact of the experiment solely on new observers (in that way, it is comparable to the coefficient associated with "Treat. 1 x new observer" in column (1)). The results of columns (2) and (3) show that the effect of Treatment 1 on new observers is not driven by the heterogeneous quality of starting team observers.

Table A.17: Impact of observer quality on the effect of treatment 1

| | Results from column 2 of Table 1 of the manuscript | Results of the second stage plugging the observers fixed effects (Sample: only new observers, from 6$^{th}$ month of implementation onwards) | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Treat. 1 x Starting team observer | 0.58 (0.66) | | |
| Treat. 1 x New observer | 1.38** (0.57) | 0.77* (0.41) | 0.78* (0.42) |
| Treat. 1 x Observer quality | | | -0.14 (0.23) |
| Treat. 1 x Treatment 2 | -1.56** (0.68) | 0.35 (0.60) | 0.55 (0.67) |
| Treat. 1 x Treatment 3 | -0.51 (0.64) | -0.29 (0.55) | -0.34 (0.58) |
| Enabler | 3.28** (1.34) | | |
| Tenure | 0.12 (0.14) | -0.02 (0.07) | -0.02 (0.07) |
| Tenure x New observer | -0.04 (0.16) | | |
| New observer | -1.60* (0.91) | | |
| Observer quality | | 0.55** (0.28) | 0.67** (0.31) |
| Store-month fixed effects | Yes | Yes | Yes |
| Observations | 585 | 217 | 217 |
| R-squared | 39.33% | 11.52% | 11.17% |
| Mean (Standard deviation) | 5.02 (2.82) | 5.02 (2.82) | 5.02 (2.82) |

**Note:** The regressions is estimated using OLS. Errors in parentheses: robust and clustered at the observer level. * p<0.1,** p<0.05, *** p<0.01. In the first stage we used only committee observers and the first five months in each site. We regressed using a regression of monthly observations on the control variable "tenure", store-month combination fixed effects, and observers fixed effects. Then, we added these fixed effects as a continuous control variable labeled "observer quality" in the second stage, estimated using only new observers and data from the sixth month of implementation onwards. The assignment of the fixed effects was as follows: a new observer in group "w" of treatment 1 was assigned the fixed effect of the starting team observer that was in group "w"; the new observers in the control group were assigned the average of fixed effects of the committee observers that were in the control group (results do not change if we assigned the percentile 25th or 75th).

**Discarding "observers of Treatment 1 creating team spirit".** It could be that forming groups comprised of few observers automatically triggered a team spirit among them, which drove their increased effort (and not the repeated interactions between observers and workers). We provide evidence against this: a lack of "peer pressure" for observers in Treatment 1, and a lack of "within-group help in coaching".

Regarding *peer effects/pressure*, we find evidence against this possibility; we show that

"spontaneous" peer pressure was present in the control group but not in the treatment groups. As the implementation carried on, we learned that the number of observations executed by each observer was discussed at the monthly meetings of the starting team (regardless of Treatment 3). This generated peer pressure on observers who were lagging in their numbers. To explore this, Table A.18 regresses the number of observations over a dummy variable that captures whether the observer was below the median of the cumulative number of observations per observer up to the previous month. This variable displays plenty of within-observer variances, so we added observer-fixed effects. The dummy generated an increase in observations only in the control group (and reassuringly, it did so disproportionately for starting team observers, which are the ones that met). This suggests that the impact of Treatment 1 is likely not due to peer effects. The exit interviews suggest that being in a group of its own substituted the peer pressure going on in the control group: the observers under Treatment 1 became "responsible for their own group of workers, not sharing responsibility with others in the starting team members"; in short, Treatment 1 "cut-observers off" from the peer control. This again lends more credibility to the interpretation of the repeated interactions.

Table A.18: Impact of observation ranking and its interaction with treatment 1 and 3

| | Observations (1) | Observations (2) | Observations (3) | |
|---|---|---|---|---|
| | | | Starting team observers | New observers |
| Low effort in last month | 0.56 | 2.11*** | 2.21** | 1.38*** |
| | (0.54) | (0.78) | (1.08) | (0.42) |
| Treat. 1 x Low effort in last month | | -2.19** | -2.82*** | -0.06 |
| | | (0.76) | (0.96) | (0.62) |
| Treat. 3 x Low effort in last month | | -1.30† | -1.28 | -0.51 |
| | | (0.86) | (1.13) | (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%) | |

**Note:** 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 effort (instead of a dummy).

*Coaching* refers to the practice across all sites that, in addition to observations, an observer may also perform coaching, where s/he watches a fellow observer execute an observation and then provides feedback and suggestions for improvement. Coaching functions "on-demand"

at Sodimac; observers request to be observed, and fellow observers step up to the call. This is another type of cooperative behavior by observers because requesting and coaching are privately costly but collectively beneficial. It would be an alternative explanation of our proposed mechanism if observers under Treatment 1 were helping each other more than in the control, providing coaching to one another, especially to new observers, and this drove the increase in observations we document for Treatment 1. In Table A.19, we study how our treatments affect the amount of coaching and the extent to which this coaching drove the impact on observations.

Table A.19: Impact of treatments on coaching

| | Coached observations (1) | Coached observations (2) | Observations (3) |
|---|---|---|---|
| Treat. 1 | 0.42** | | |
| | (0.19) | | |
| Treat. 1 x Starting team observer | | 0.44*** | 0.41 |
| | | (0.22) | (0.64) |
| Treat. 1 x New observer | | 0.40** | 1.22** |
| | | (0.21) | (0.52) |
| Treat. 1 x Treat. 2 | -0.14 | -0.14 | -1.52** |
| | (0.22) | (0.22) | (0.63) |
| Treat. 1 x Treat. 3 | -0.27 | -0.28 | -0.43 |
| | (0.20) | (0.21) | (0.61) |
| Enabler | 0.49*** | 0.49*** | 2.87** |
| | (0.16) | (0.16) | (1.19) |
| Tenure | 0.02 | 0.02 | 0.11 |
| | (0.05) | (0.05) | (0.13) |
| Tenure x New observer | -0.38*** | -0.39*** | 0.11 |
| | (0.09) | (0.09) | (0.15) |
| New observer | 1.00*** | 1.02*** | -2.17** |
| | (0.38) | (0.39) | (0.84) |
| Coached observations | | | 0.59*** |
| | | | (0.11) |
| Store-month fixed effects | Yes | Yes | Yes |
| Observations | 585 | 585 | 585 |
| R-squared | 21.69% | 21.69% | 44.49% |
| Mean (Standard deviation) | 1.15 | 1.15 | 5.02 (2.82) |

**Note:** All regressions are estimated with OLS, except for (1) and (2), which are Poisson regressions. Errors in parentheses: robust and clustered at the observer level. * p<0.1,** p<0.05, *** p<0.01. Results are robust to the inclusion of the interaction of treatments 2 and 3 with new and starting team observers.

In columns (1) and (2), we replicate the analysis of Table 1 of the main body of the paper 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 (1) shows that Treatment 1 increases the amount of coaching that the observers receive, and column (2) shows that this effect is concentrated on new observers. Assuming covariates are set to zero, the impact of being a new observer in the control is exp(1.02)=2.77 coached observations, whereas adding Treatment 1 generates exp(1.02+0.4)=4.13 coached observations — Treatment 1 generates 1.36 additional coached observations. In contrast, starting team members in the control group have exp(0)=1 coached observations, while in treatment 1, they generate exp(0.44)=1.55, only 0.55 additional coached observations than for new observers. Treatments 2 or 3 have no impact on the amount of coaching.

Column (3) 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 on observations. The coefficient of "Treatment 1 x starting team observer" drops from 0.58 in column (2) of Table 1 of the main body to 0.41 in column (3), 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. In sum, coaching can be seen as a cooperative act on its own, and it explains only a minor part of the overall impact of Treatment 1 on observations.

To further dispel the coaching concern, we hand-collected data on who coached whom to study the source of the additional coaching in Treatment 1. Table A.20 presents the analysis. We had 213 coaching events for new observers (i.e., the new observer was coached in an observation). We excluded 26 that were done by consultants, leaving 187 coaching events. Out of these, in 95 cases, the coached observer was a new observer who 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. We assigned them to the control group to assess their impact and then analyzed the results with and without their 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 who 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 each site was 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 would have artificially increased, as it goes down artificially down in panel a). Again, we cannot reject the null hypothesis that coaching was done randomly.

Table A.20: Identity of coaches

| Coaching events done by enabler: | **Panel a.** Only coached observers of **treatment** groups Included | Excluded | **Panel b.** Only coached observers of **control** group Included | Excluded |
|---|---|---|---|---|
| Number | 95 | 72 | 92 | 54 |
| Percentage of events in which the coach was an observer of the respective group [Mean (S.D.)] | 0.063 (0.245) | 0.083 (0.278) | 0.696 (0.462) | 0.481 (0.504) |
| Theoretical benchmark of random coaching for the % executed by a coach of the same group | 0.1 | 0.1 | 0.5 | 0.5 |
| Is the actual execution different than the benchmark? (p-value) | 0.145 | 0.613 | 0.001*** | 0.788 |

**Note:** Total number of coaching events, excluding those done by consultants = 187. * p<0.1,** p<0.05, *** p<0.01. S.D. stands for "standard errors".

These findings indicate that coaching *was not preferentially executed by members of the same group. Rather than observers of group 'X' in Treatment 1 disproportionately coaching each other, we found instead that the coach source came randomly from different groups in the treatment and control groups.* This suggests that relationships *among observers* within groups of Treatment 1 were not the motivator of the additional coaching. Instead, this result is consistent with new observers being more motivated to improve the quality of their observation because, due to repeated interactions, they were more committed to providing high-quality training to the smaller set of workers they observed repeatedly.

**Exit interviews.** The exit interviews provided 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," "It is like being counseled by your 'father,' and not any random guy ... you will meet your 'father' continuously, so you better comply", or that "It created a kind of bond."

**Naturally occurring groupings in the archival data.** As already discussed, BAPP provides freedom for the site to try different implementation tactics. This generates heterogeneity in terms of how much observers specialize their observation on different areas of a site. Specialization in specific areas of the site 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 (note, however, that this effect is not encouraged by BAPP, to avoid the counteracting effect of the "blind eye," that is, getting too familiar with tasks or workers and therefore decreasing quantity and quality of observations); ii) a "repeated interaction effect": the observer now interacts with a reduced set of workers, and this increases the frequency of interaction. Our experiment focused on ii) by shutting down i) via randomization.

Using DEKRA's archival data, we can measure the extent of site-area specialization. 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.[7] Then, we average this for a site every month (this generates some variation over time as the pool of observers changes on the site). This variable displays plenty of variance (Figure A.7). More importantly, at the low end of the distribution, we observe an HHI of 0.1 to 0.2, consistent with random observations across 5 to 10 areas, the typical number of areas in BAPP implementations. The abundance of sites displaying behavior consistent with random observations is likely to be produced by the fear of the "blind eye" (see discussion above and in the main body).

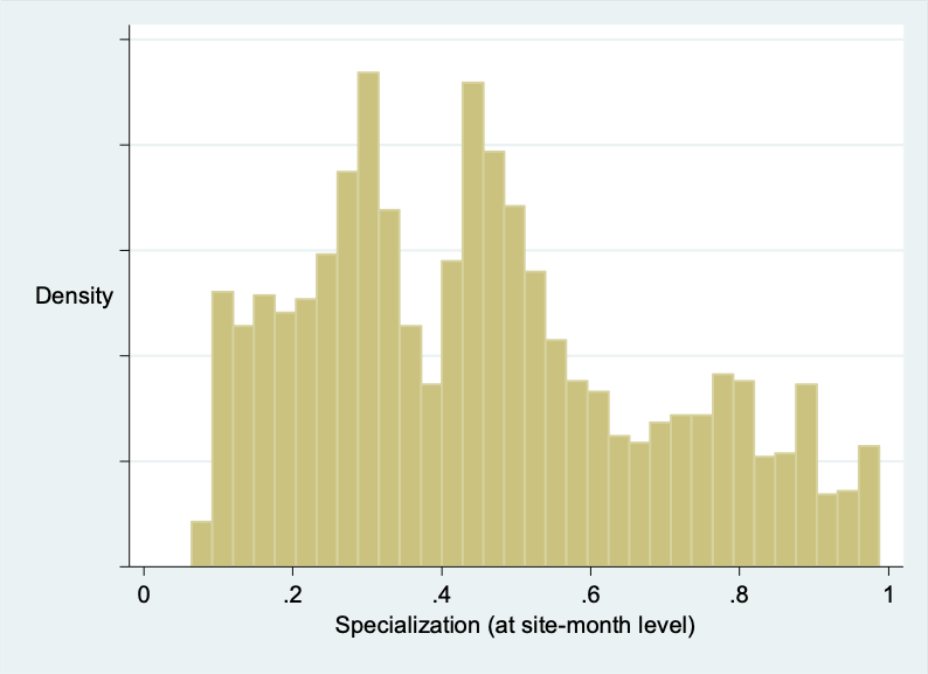



Figure A.7: Distribution of specialization

We then used this measure in an econometric analysis of the impact of BAPP on accidents. In this analysis, we "shut down" mechanism i) of "learning" by explicitly controlling for the variable "Experience," which captures the cumulative observations up to month t-1. We estimate a model analogous to Equation (A3) in subsection A.2.3 of this Appendix. The results are displayed in Table A.21. We find in column (1) that area specialization greatly enhances the impact of BAPP. Given that we control for "learning," this effect reflects mostly the impact of area specialization via repeated interactions. Further, the interaction between BAPP and experience in column (2), and the triple interaction between BAPP, experience and area specialization in column (3), are neutral. Overall, this strongly indicates that the impact of the HHI index of area specialization is not due to learning; instead, consistently with our experiment, it suggest that it is channeled through the mechanism ii), increased

[7] 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.

repeated interactions via structure between observer and worker.

Table A.21: The role of specialization on the impact of BAPP

| | Accidents - OLS (1) | Accidents – OLS (2) |
|---|---|---|
| BAPP | 0.210 | 0.212 |
| | (0.134) | (0.166) |
| Trend | -0.033** | -0.033** |
| | (0.014) | (0.014) |
| BAPP x Specialization | -0.649** | -0.655** |
| | (0.212) | (0.291) |
| BAPP x High Effort | -0.283*** | -0.283*** |
| | (0.106) | (0.106) |
| BAPP x High Diffusion | 0.226** | 0.226** |
| | (0.098) | (0.097) |
| BAPP x Tenure | 0.034** | 0.034** |
| | (0.014) | (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*** | 1.230*** |
| | (0.331) | (0.330) |
| Site fixed-effect? | Yes | Yes |
| Constant | -5.247*** | -5.246*** |
| | (1.757) | (1.755) |
| R-square | 43.30% | 43.30% |
| Observations | 4,447 | 4,447 |
| Mean of dependent variable before BAPP | 1.338 | 1.338 |

**Note:** 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 50th percentile ascut-offf. Tenure is measured as the 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 are then averaged to obtain a site-level variable.

### A.3.7 Treatment 2 Is Not Treatment 1 Badly Implemented

We can discard an alternative explanation for the negative impact of Treatment 2, namely that "its negative impact is simply Treatment 1 badly implemented". Given that Treatment 2 is an addition on top of Treatment 1, it could be that the two consultants that executed it – one in Temuco and one in La Reina – were less effective in executing Treatment 1, which led to the negative outcome of Treatment 2 rather than the "anonymity backlash." However, several arguments and tests indicate that this is not the case. First, the consultant in La Reina also executed BAPP in Huechuraba, which had Treatment 1 but not Treatment 2. Second, we ran a regression restricting the sample to include only La Reina and Huechuraba, who have the same consultant. The results did not change (see column (1) of Table A.22): Treatment 1 increased observations, and Treatment 2 decreased them; therefore, the result of Treatment 2 also occurred within the area of one of the "suspect" consultants. Third, we performed a regression interacting Treatment 2 with the condition of being a new observer or a starting committee observer (see column (2) of Table A.22). If Treatment 2 was 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 a deficient implementation of Treatment 1 was the driving force, then Treatment 2's negative impact should be concentrated on new observers, where Treatment 1 exerts its impact. We find the former to be the case: Treatment 2 was unaffected by the observer type.

Table A.22: Tests on Treatment 2

| Sample: | Effort | | |
|---|---|---|---|
| | Consultant in La Reina and Huechuraba | Full | Full |
| | (1) | (2) | (3) |
| Treat. 1 | | | 0.94*** |
| | | | (0.34) |
| Treat. 1 x Starting team observer | 0.66 | 0.60 | |
| | (0.52) | (0.50) | |
| Treat. 1 x New observer | 1.24*** | 1.36*** | |
| | (0.38) | (0.40) | |
| Treat 1 x Time elapsed | | | |
| Treat. 1 x Treat. 2 | -0.63** | | -2.04*** |
| | (0.67) | | (0.63)# |
| Treat. 1 x Treat. 2 x Starting team observer | | -1.60*** | |
| | | (0.56) | |
| Treat. 1 x Treat. 2 x New observer | | -1.52*** | |
| | | (0.49) | |
| Treat. 1 x Treat. 2 x Time elapsed | | | 0.085 |
| | | | (0.08)# |
| Treat. 1 x Treat. 3 | -0.74 | -0.51 | -0.50 |
| | (0.61) | (0.42) | (0.41) |
| Enabler | 1.84** | 3.28*** | 3.30 |
| | (0.65) | (0.62) | (0.62) |
| Tenure | 0.06 | 0.12 | 0.14 |
| | (0.12) | (0.13) | (0.12) |
| Tenure x New observer | 0.003 | -0.04 | -0.06 |
| | (0.161) | (0.15) | (0.14) |
| New observer | -1.69* | -1.58* | -1.08 |
| | (0.88) | (0.91) | (0.84) |
| Store-month fixed effects? | Yes | Yes | Yes |
| Observations | 390 | 585 | 585 |
| R-squared | 33.77% | 39.33% | 39,41% |
| Mean (Standard deviation) | 5.06 | 5.02 | 3.47 |
| | (2.76) | (2.82) | (0.69) |

**Note:** All regressions are estimated with OLS. Errors in parentheses: robust standard errors. * p<0.1,** p<0.05, *** p<0.01. In column (1), we excluded the site of Temuco, which was the other site with treatment 2; thus, in this regression, the interaction with treatment 2 only used the stores. (#) In column (3), a joint t-test is statistically significant at p<0.01.

Fourth, Treatment 2 has a negative impact on dependent variables that capture workers'

outcomes (i.e., risky behavior, accidents, and the likelihood of becoming an observer; see Tables 1 and 2 in the main body of the paper) but no impact on coaching. This dependent variable exclusively captures observer behavior (see Table A.19 above). This is consistent with workers being the driving force behind the negative effect of Treatment 2, and thus a "workers' backlash"; if bad implementation were the reason, coaching would also have suffered. Fifth, we explored the effect of time on the impact of Treatment 2 (non-reported analysis, but available upon request). Treatment 2 was particularly detrimental at the start of BAPP implementation, generating a backlash of approximately two observations in the first couple of months. After that, the negative effect was gradually reduced to one observation 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.