ECONtribute
MARKETS & PUBLIC POLICY

ECONtribute
Discussion Paper No. 304

Is This Really Kneaded? Identifying and Eliminating Potentially Harmful Forms of Workplace Control

Guido Friebel Matthias Heinz Mitchell Hoffman
Tobias Kretschmer Nick Zubanov

May 2024

www.econtribute.de

UNIVERSITÄT BONN
Universität zu Köln

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1-390838866 is gratefully acknowledged.

Is This Really Kneaded? Identifying and Eliminating Potentially Harmful Forms of Workplace Control*

Guido Friebel[†] Matthias Heinz[‡] Mitchell Hoffman[§]

Tobias Kretschmer[¶] Nick Zubanov[‖]

May 2024

Abstract

In a large German bakery chain, many workers report negative perceptions of monitoring via checklists. We survey workers and managers about the value and time costs to all in-store checklists, leading the firm to randomly remove two of the most perceivedly time-consuming and low-value checklists in half of stores. Sales increase and store manager attrition substantially decreases, and this occurs without a rise in measurable workplace problems. Before random assignment, regional managers predict whether the treatment would be effective for each store they oversee. Ex post, beneficial effects of checklist removal are fully concentrated in stores where regional managers predict the treatment will be effective, reflecting substantial heterogeneity in returns that is well-understood by these upper managers. Effects of checklist removal do not appear to come from workers having more time for production, but rather due to improvements in employee trust and commitment. Following the RCT, the firm implemented firmwide reductions in monitoring, eliminating a checklist regarded as demeaning, but keeping a checklist that helps coordinate production.

Keywords: Monitoring; checklists; respect; time use

*We thank Charlie Brown, Alessandra Fenizia, Michael Kosfeld, Rosario Macera, Axel Ockenfels, Paul Oyer, Andrea Prat, Jonah Rockoff, Kathryn Shaw, Chris Stanton, Lowell Taylor, John Van Reenen, and especially Wouter Dessein, as well as numerous seminar/conference participants. We thank the study firm and its management for their enthusiastic participation in this collaboration. Financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2126/1– 390838866) and SSHRC is gratefully acknowledged. The experiment was pre-registered on 04/14/2021 with the AEA RCT registry under ID AEARCTR-0007550. IRB approval was received from the University of Cologne. The Worker Council approved the project and was involved in all steps.

[†]Goethe University of Frankfurt and CEPR and IZA

[‡]University of Cologne and CEPR and Max Planck Institute for Research on Collective Goods

[§]UC Santa Barbara and University of Toronto and NBER and CEPR and IZA

[¶]LMU Munich and CEPR

[‖]University of Konstanz and IZA

Starting with Taylor (1919), a tradition of research on workplace productivity emphasizes the importance of *workplace control*, broadly defined as firms' actions to structure and control employee behavior at work. Workplace control includes monitoring, communication, and other management practices. The degree and form of workplace control varies widely, e.g., many firms have extensive rules about how to handle situations whereas others do not.[1] Randomized controlled trials (RCTs) in field settings, detailed below, emphasize the value of employee monitoring and other forms of control in improving performance. However, lab experiments suggest that excessive employer control has the potential to signal mistrust and contribute to a negative environment (Falk & Kosfeld, 2006).

Firms commonly use checklists as a control tool. These are structured lists that workers fill out or work through. Checklists are celebrated as a powerful tool to help workers remember things and coordinate production (Gawande, 2010). Famously used in surgery and aviation, checklists have been applied in many industries, including retail. Studies show large benefits of checklists, but little is known about whether and when checklists can be harmful, and this is for several reasons. First, as is true for many management practices (Bloom *et al.* , 2014), firms' use of checklists is highly non-random, making it difficult to estimate causal returns. Second, modern firms often use numerous checklists, so even if one believes a workplace is "overmonitored," as often alleged by popular observers,[2] it is hard to know which checklists to modify. Third, it seems likely that returns to monitoring through checklists are heterogeneous—some workers may benefit from extra structure, while others may find checklists useless and insulting. Capturing this heterogeneity seems critical for a full understanding of checklists.

We survey workers and managers to identify two potentially harmful forms of monitoring, namely, two checklists called the operational checklist and daily protocol, leading the firm to randomly eliminate the two checklists in half of stores. We believe this to be the first large-scale RCT on removing checklists or monitoring at work.

Our partner is a major German bakery chain with 145 stores and over €100m of annual revenue. Prior to our intervention, the firm was using checklists in many aspects of production. Workers needed to record extensive information, not only about their products (e.g., when they took bread out of the oven), but also on customer interactions, such as whether they smiled. Drawing on a deep collaboration with the firm and top management, we conduct intensive pre-RCT interviews and surveys, and discover that several checklists

[1] E.g., in clothing retail, Nordstrom's employee handbook consists of the single rule "Use good judgment in all situations" while other firms have detailed rules.

[2] That US workplaces may be overmonitored has been alleged in many contexts, from call-centers to Amazon warehouses to tech (Guendelsberger, 2019). See also the 2022 articles in the Economist ("Welcome to the era of the hyper-surveilled office") and NY Times ("The Rise of the Worker Productivity Score").

are perceived as especially low-value (i.e., high time costs and limited benefits).

Our RCT is grounded in a simple conceptual framework of checklists, as laid out in Section 1. Monitoring through checklists helps firms address moral hazard problems, coordinate production, and remind workers of tasks. However, checklists also entail costs, both directly in terms of time and indirectly in terms of other factors, such as by reducing worker happiness or signaling mistrust. The framework grounds which checklists are best to remove and which stores may benefit most from checklist removal.

As detailed in Section 2, our bakery chain represents an ideal setting for our RCT. First, the sample is large. Second, we access granular administrative data, coupled with the ability to conduct high-quality surveys. The administrative data cover detailed aspects of sales, customers, and orders hour by hour, which is critical for examining how workers and managers use time and how they substitute time on checklists to other tasks. Because of our deep collaboration also with the firm works council, the surveys have high response rates, plus in-depth open-ended questions, which are critical for understanding mechanisms. Unusually, we survey not only store employees and managers, but also regional managers (the bosses of store managers) and have them predict in which stores the RCT will be effective.

In Section 3, we estimate that removing checklists increases sales by 2.7%. The impact on sales is similar during busy and slower times. While removing monitoring could lead to wasted food, employee misbehavior, or coordination failures, we find no negative impact on shrinkage (a joint measure of food waste and worker stealing) or mystery shopping scores, i.e., the scores given by undercover shoppers. Surveys indicate that checklist removal increases worker trust and commitment. Google reviews reveal that, in treated stores, consumers are more likely to perceive fast line speed and high product quality, illustrating how positive sales effects may manifest. Turning from sales, there is no overall impact of the treatment on employee attrition. Still, there is a strong, negative effect on attrition of store managers, who do a lot of checklist completion and who may naturally appreciate doing less. In contrast, the treatment has a positive, though statistically insignificant, effect on attrition for unskilled workers without vocational training who may benefit from structure and checklists.

Our initial discussions with regional managers highlighted their expectation that treatment effects would likely be highly heterogeneous across stores. In our survey of regional managers, conducted prior to randomization, managers predict that in about half of their stores the treatment would be effective, and in the other half it would not. Thus, in our RCT pre-registration, we focused strongly on this aspect of heterogeneity. Splitting the sample based on whether regional managers predicted the store would work, we observe vast differences in results (Section 4). Among stores where the RCT was predicted to be successful, removing checklists increases sales by 5%. There are broad-based improvements in store

operations, with round-the-clock improvements in sales, statistically increases in customers, and a decrease in shrinkage. In contrast, in stores where the treatment was not predicted to work, the impact on both store-level outcomes and employee attrition is zero. If anything, mystery shopping scores are slightly worse, though the impact is not statistically significant.

To better understand these heterogeneous effects, we dig into the free text of regional managers' responses on why the treatment would work in particular stores. Among stores where regional managers predicted the treatment will work, in about one-third of cases, regional managers explicitly mention something about workers enjoying the removal of checklists, consistent with a utility cost to excessive monitoring. In about two-thirds of cases, regional managers mention something about the absence of problems, consistent with traditional views of monitoring to help detect and avoid problems.

The firm was quite satisfied with the results of the RCT, as discussed in Section 5. Unlike past interventions in the literature, our RCT subtracted instead of adding something, so the direct cost to implement the RCT was minimal. With benefits of checklist removal exceeding costs by a ratio over 50, the firm decided to implement a firmwide checklist removal program. Interestingly, while the RCT removed two checklists, the firm restored the daily protocol in the firmwide rollout even though the operational checklist was eliminated.

Our paper contributes to several literatures. First, it contributes to work in personnel and organizational economics, as well as social science more generally, on the returns to checklists and monitoring.[3] Most influentially, the physician Atul Gawande (2010) summarizes studies and in-person observations from a number of domains, including those of surgeons (see Ko *et al.* (2011) for a review), airline pilots (Boorman, 2001), and investors, to argue that checklists can have profound positive organizational consequences. Our findings show that the returns to monitoring need not be positive, as we estimate sizable positive benefits of removing checklists. The central reason, we believe, is the presence of indirect costs of monitoring. Using lab experiments with assigned roles, Falk & Kosfeld (2006) show that workers react negatively and often choose low effort when being controlled by the manager. Our paper suggests that such insights extend into the field as well, and we offer a framework that rationalizes why monitoring may be good for some tasks, but bad for others.

In economics RCTs on monitoring, closest to ours is Nagin *et al.* (2002), who consider a field experiment where a call-center company exogenously varies its monitoring rate in some

[3]Economics RCTs showing benefits of monitoring include Nagin *et al.* (2002), Duflo *et al.* (2012), Jackson & Schneider (2015), Gosnell *et al.* (2020), and Kelley *et al.* (2021). In most of these studies, monitoring is added instead of taken away. There are also many observational studies which document benefits to monitoring, especially in the trucking industry (Hubbard, 2000, 2003). In contrast, lab experiments are more likely to illustrate potential overmonitoring (Dickinson & Villeval, 2008; Falk & Kosfeld, 2006). As surveyed by Ravid *et al.* (2023), there is also a large psychology literature on monitoring which focuses on correlations of monitoring with survey outcomes.

call-centers. They show that increasing the declared monitoring rate leads to a decrease in suspected bad calls, but that a certain share of workers do not appear to respond to additional monitoring, due to a belief that workers should behave in an appropriate manner. Despite key differences in the nature of the RCTs,[4] we believe both papers are highly complementary and point to broader conceptions of how monitoring affects workplace behavior beyond the classic contract theory perspective (Holmstrom, 1979), both why some workers behave well despite limited monitoring (Nagin *et al.* , 2002) and why some workers and teams perform poorly while monitored (our paper).[5] Our results indicate that some forms of monitoring can harm firm performance and be a disamenity to employees, and that one can identify such forms of monitoring by surveying workers and managers.

Also closely related to ours is Bandiera *et al.* (2021), who conduct an RCT in Pakistan where authority is transferred to tax collectors from their monitors. They show that delegating to tax collectors increases their performance. Instead of changing authority, our study changes monitoring holding authority fixed.

Second, our paper contributes to work in personnel economics on the heterogeneous returns to management practices and on the impact of managers. Amid substantial work on the importance of management practices in general (Bloom *et al.* , 2012, 2019), growing research emphasizes that management practices are complementary to one another (Milgrom & Roberts, 1990; Ichniowski *et al.* , 1997), and that their impact may be contingent on other factors within an organization (Blader *et al.* , 2020). We show that there is substantial heterogeneity in the return to a management practice, namely, checklists, based on regional manager beliefs. Manager beliefs are somewhat correlated with some observable traits of stores, e.g., managers correctly predict that the treatment will be larger in smaller stores, but there is substantial predictiveness of manager beliefs beyond observable characteristics. As surveyed in Roberts & Shaw (2022), a rich and growing literature examines what do non-CEO managers do and their impact, often emphasizing the role of managers in motivating and teaching employees. Our results suggest that an important way managers can add value is by having private information about their teams, consistent with theories of dispersed information in organizations (Dessein, 2002; Dessein & Santos, 2006).

[4]First, Nagin *et al.* (2002) examine audit rates, a non-checklist form of monitoring. Second, Nagin *et al.* (2002) study intensive margin changes in monitoring, whereas we study extensive margin changes (i.e., eliminating monitoring). Third, in Nagin *et al.* (2002), production is individual, whereas our workers work in teams, and this matters for coordination benefits of monitoring. Fourth, our study is about workers reacting negatively to excessive monitoring, whereas Nagin *et al.* (2002) is about some workers behaving well despite a lack of monitoring. Fifth, the metrics studied in Nagin *et al.* (2002) suggest that less monitoring is bad in their context, whereas our results suggest that less monitoring is good on average.

[5]de Rochambeau (2020) shows that randomly monitoring Liberian truckers increases their effort, though there are some workers who reduce their output after being monitored. Hiring students to identify coins, Belot & Schröder (2016) show that randomly added monitoring can backfire on some dimensions of performance.

Third, it provides a clear example of a successful large-scale intervention that is "subtractive," i.e., involves taking something away. Psychologists argue that individuals and firms systematically overlook subtractive interventions (Adams *et al.* , 2021), perhaps because humans are hard-wired to look for new things (Klotz, 2021) or because it is easier to take credit for an addition. We are not familiar with prior economics RCTs that improve outcomes via subtraction. That our RCT is subtractive is substantively important, as it makes our RCT's cost quite low. Several additive management practices also yield performance improvements broadly similar to ours, but our RCT's benefit to cost ratio over 50 is the largest that we are aware of in the management practice literature.

Fourth, our paper makes a methodological contribution to RCTs. Beginning with DellaVigna & Pope (2018), work uses expert predictions to examine how the results of an RCT compare to priors of experts, i.e., to see to what extent a result is surprising (DellaVigna *et al.* , 2019). Rather than having experts predict the average results of the RCT (e.g., that the treatment will affect sales by a certain amount), our RCT has experts predict store by store whether the treatment will be effective in that particular store. We are aware of very limited prior work that uses expert predictions in RCTs in this manner, but we believe this methodology may be useful in other contexts.[6] Experts in our context have substantial knowledge about which units will be most affected.

Fifth, our paper relates to discussion in behavioral economics on the relevance of lab findings to the field. In experimental economics, there is substantial discussion on trust vs. control (Ellingsen & Johannesson, 2008; Falk & Kosfeld, 2006; Herz & Zihlmann, 2022). We show that concerns about employer overcontrol are relevant also in a large firm.

1 Conceptual Framework

What is the average impact of checklists on performance, and how would this impact vary across stores within a firm? Besides shedding light on these two questions, our framework helps motivate which checklists are best to remove and also models the implications of treatment effect heterogeneity according to regional manager expectations. Checklists are randomized in our RCT, so we focus on the impact of checklists instead of the decision to adopt them. For simplicity, we model the binary comparison of having checklists versus not.

As in Garicano (2000), the firm faces *problems*, though we think of problems in a very broad sense, covering issues of information and agency. First, problems can be memory

[6] E.g., doctors could predict which individual patients will respond to a drug. Broadly relatedly, Dal Bó *et al.* (2021) ask supervisors of government agricultural workers to rank which workers should get free phones, and Bryan *et al.* (2021) have loan officers predict how individual microfinance clients will fare under treatments. Our setup differs in that we focus on predictions of higher-up experts in a private-sector firm.

problems, such as where people on a surgery team forget certain steps (Gawande, 2010) or where bakery workers forget to put doughnuts at the correct angle. Second, and very importantly for us, these can be coordination problems, e.g., a bakery worker forgets to pass along to the next shift at what time the bread was made. Finally, these can also be moral hazard problems where workers behave opportunistically (Nagin *et al.* , 2002). To keep things as simple as possible, we assume that problems occur exogenously with probability p, but the logic of our model can be easily extended to having workers choosing whether to behave opportunistically. When a problem occurs, the cost to the firm is k. Thus, without checklists, firm profits are $-pk$.[7]

Checklists helps the firm and its employees to identify problems.[8] The impact of checklists on problem-solving is given by m, and represents the probability that a problem is detected and solved in full. Using checklists involves direct cost, c, which can include the technology itself, but in our setting is primarily the time cost of filling out checklists.

In addition, monitoring entails an indirect cost θ to firm performance. Many people seem to dislike being monitored and controlled, perhaps because it is intrinsically unpleasant to fill out checklists, but also because monitoring and control can be viewed as a sign of disrespect (Ellingsen & Johannesson, 2007, 2008). Ellingsen & Johannesson (2008) model workplace respect in terms of second-order beliefs, i.e., a worker's belief about the firm's belief about whether she is altruistic or competent. Being respected can be important for firm performance, both because it makes workers more likely to stay with the firm (Friebel *et al.* , 2023) but also because it motivates them to work harder (Cai & Wang, 2022). Alternatively, control could crowd out intrinsic motivation to work hard (Benabou & Tirole, 2003; Rebitzer & Taylor, 2011). It is natural that θ could depend on p and k, i.e., the checklist may feel most onerous when it serves little purpose, such as when there are few problems to solve or when the cost of problems is small.[9]

Therefore, the profits from checklists are $-(1-m)pk-c-\theta$, and the returns from our treatment of removing checklists are $c+\theta-mpk$. This expression allows us to characterize whether the treatment is likely to be positive or negative, as well as to predict what are the stores where the treatment will have the largest benefit. Specifically, our treatment is likely

[7]Stores may differ not in the frequency of problems, but rather in ability to solve them (Garicano, 2000). Thus, one can alternatively define p as the share of problems stores cannot solve without checklists.

[8]We think of checklists as highly structured forms of documentation. These include, of course, a list that worker checks off. Checklists also include forms of documentation where workers enter simple information in a structured fashion, such as how much money is in the cash register, how much was expected, and a list of IT problems. Checklists can be done using pen-and-paper or electronically. Checklists can be performed in a group (e.g., a surgical team) or individually.

[9]Pilots may be fine with checklists because the cost of problems is high. In retail, the costs of problems is lower (e.g., someone skips a doughnut), so checklists may feel more unpleasant. Having θ depend on p and k can also explain stores with fewer problems being more frustrated by a sense of excessive control.

to be positive when there are important direct and indirect costs of control by checklists, as well as when a firm faces infrequent problems, when the memory technology can less reliably identify problems, and where the cost of those problems is lower. It seems likely to us that stores would vary most in the frequency of problems, p, and in the indirect costs to checklists, θ.[10]

This framework also raises the possibility that there could be heterogeneity across stores within a firm in the returns to using checklists. Regional managers may know that some stores experience coordination problems more frequently; stores may also vary in outcomes if some workers dislike checklists more than others (e.g., if some workers find them more disrespectful or wasteful than others), and store employees may differentially complain about these costs to regional managers. Given that there are multiple factors affecting whether monitoring has positive effects and that some factors (like frequency of coordination problems) are hard to observe in data, it is natural to ask regional managers to make predictions about whether a treatment will work in a store.

Formally, the performance impact of the treatment is $z = c + \theta - mpk$. Regional managers observe a private signal $\hat{z} = z + \epsilon$ of treatment implications in a store, and state a subjective belief $B = 1(E(z|\hat{z}) > z^*)$ about whether the treatment will work in a store, where z^* is a threshold level of effectiveness.[11] The private information a manager has is represented by the signal's precision, $h_\epsilon = \frac{1}{\sigma_\epsilon^2}$. Managers believe the treatment will work when the treatment effect is above a threshold. Thus, the more private information that regional managers have about z, the greater is $E(z|B = 1)$, i.e., the average effect of the treatment among stores where the regional manager predicts the treatment will work. Likewise, the more private information that regional managers have, the greater is $E(z|B = 1) - E(z|B = 0)$, i.e., the difference in treatment effects between stores where managers think the treatment will work relative to stores where managers think the treatment will not work.

Our framework focuses on store performance, in line with our RCT pre-registration, but is easily extended to cover worker attrition. It is natural that the direct and indirect costs of control through checklists is not only reflected in performance, but also in worker utility from the job. Our treatment is likely to reduce attrition most for workers with higher personal costs of, but could increase attrition for workers who personally benefit from checklists that could provide valued structure. Dube *et al.* (2022) find that workers care deeply about

[10] Of course, there could also be heterogeneity across stores in m and k, e.g., if certain stores have greater costs when problems arise. Given the multiplicative term mpk, heterogeneity in p leads to the same effect in the model as heterogeneity in m or k.

[11] In reality, regional managers may observe multiple signals, e.g., one on the frequency of problems, and one on the indirect costs of checklists. Our framework follows the RCT, where we elicit a manager's overall belief about the treatment's effectiveness in each store, doing so in the RCT for brevity and clarity.

being respected and provide evidence that this matters for turnover. It is natural that more qualified workers would experience more positive effects on attrition of checklist removal.

In sum, theory does not make clear predictions about the overall impact of checklist removal, as there are costs and benefits to checklists.

2 Study Background

The firm. Our study firm is one of the largest bakery chains in one densely populated region of Germany.[12] Like most bakery chains, the firm is family-owned. The CEO is also the founder of the modern version of the chain. Many of the top executives helped set up the chain with the CEO over the last 40 years. The firm has roughly 2,000 employees. The firm has one plant which produces raw products (e.g., unbaked bread which is baked in store ovens) for the firm's 145 stores. About 90% of the bakery stores are located next to grocery stores, with hours fixed by the rental contract with the grocery store chain.[13] The firm has a reputation for quality products, as evident, among other places, in online reviews. Wages are slightly above the German minimum wage (Dustmann *et al.* , 2009).

Most employees work in the stores with an average of 13 employees per store (including the store manager). Most of the firm's employees hence work in the sales (or operations) division, headed by three sales directors supervising 15 regional managers, which each manage 10 stores on average. There is one store manager per store.

Store managers and their team predominantly prepare and finish products on-site (e.g. sandwiches or fresh bread pre-fabricated in central production but finished in store ovens). They also manage the in-store flow and presentation of goods they receive from headquarters several times a day; maintain and clean the machines; keep the store tidy and manage the sales process including customer advice; and operate the cash register.

The firm's culture is control-oriented. Detailed instructions, checklists, and regular top-down communication are used to ensure quality standards are met. Workers are also monitored by store managers and mystery shoppers. There is no formal communication between stores. Some employees, mainly those with longer tenure, may know some colleagues in other stores but this is not encouraged.

Why the firm did the RCT. Our collaboration with the firm started in 2020. In exploratory talks about potential projects, two signs indicated concerns about overcontrol

[12] In Germany, most bakery chains operate in particular regions.

[13] A typical position for a bakery store from our firm is located in the same building as a grocery store, but outside the layout of the grocery store. The bakery has its own separate entrance and is open on different days and times (e.g., Germany grocery stores are closed on Sunday, but bakeries are open).

and overmonitoring. First and foremost, we came across a 2018 employee survey which indicated broad dissatisfaction with checklists at the firm. Second, the head of HR was concerned about employee turnover, especially of trained workers, and separately expressed concern about overmonitoring (via feedback from the works council), and we thought that the two could be linked.[14] To jointly explore these observations in greater detail and rigor, we formed a project team consisting of two of this paper's coauthors; the heads of both HR and accounting/controlling; multiple employees from those two departments; one sales director; and the head of the works council.

In the 2018 employee survey, employees anonymously complained about what they deemed excessive control through time-consuming checklists. While some project team members believed that some checklists might be inefficient or counterproductive, there was no comprehensive list of checklists or broad understanding of their costs and benefits. Thus, we set out to gather survey data on all checklists at the firm. We did not gather survey data on non-checklist aspects of the firm's control system (e.g., compensation, mystery shopping), as employees in the survey did not express dissatisfaction with these elements.

Identifying potentially harmful forms of monitoring. The project team began by creating a comprehensive list of all 22 in-store checklists.[15] RAs conducted in-depth in-person surveys with 21 store managers and 18 workers in 22 randomly selected shops about beliefs regarding time use duties. Given the control-oriented culture, we were concerned that there would be issues of trust, so we asked for in-person surveys to be done to get more truthful and accurate information than via online surveys. To further establish trust, the RAs were driven to the stores by the head of the works council, who introduced the RAs to the survey respondents, emphasizing that they could trust the RAs. For 21 of the checklists, respondents were asked the following questions:[16]

1. To what extent does the checklist help the company achieve its goals (1-10 scale)?
2. To what extent does the checklist help avoid mistakes (1-10 scale)?
3. How often do you fill out the checklist each week?
4. How many minutes do you spend each time filling out the checklist?

[14] Trained workers are trained via apprenticeships paid in part by the employer (Dustmann & Schoenberg, 2012).

[15] Our approach of using a committee to generate comprehensive lists broadly follows idea generation and process optimization procedures used in large firms like Toyota (Womack *et al.* , 2007).

[16] One of the 22 checklists was omitted from interviews, namely, the one involving consent for working on Sundays. According to the works council, it is legally impossible to drop, so it was not asked about.

Figure 1 gives results on the survey, focusing on the value in helping the firm achieve its goals (Q1) and the weekly time cost (Q3 and Q4 combined). Five checklists stand out for having both relatively low value and high time cost. Three of these were considered "sacred cows" and impossible to remove for political reasons or because they were related to the unique selling proposition of the firm.[17] The two remaining duties were the **operational checklist** (*Operative Liste*) and the **daily protocol** (*Tagesprotokoll*). The daily protocol and especially the operational checklist also score poorly in terms of avoiding mistakes (Appendix Figure B4).

Workers and managers in the in-depth interviews have similar average beliefs about how much time the checklists take: These beliefs are correlated ($\rho = 0.77$) with the beliefs of top management members from our project team, gathered in a project team meeting prior to conducting the in-depth interviews. The data from our in-depth interviews thus are high quality. It also shows that top management was aware of the time required for checklists.

According to self-reports, workers spend an average of 319 minutes or almost 5.5 hours per week across the checklists. Store managers spend even more, spending 499 minutes per week or over 8 hours. Thus, according to self-reports, roughly 15-20% of worker and manager time is spent on checklists each week. Employees may exaggerate self-reported time spent on checklists, but this is not an important concern for our study.[18]

In a meeting in October 2020, the researchers presented analyses on these surveys and recommended removing these two checklists via an RCT. The firm decided to do so. The firm is no stranger to experimentation, and frequently runs "pilots" in selected shops (e.g., new products, marketing campaigns, shop design). Thus, the fact that there were significant changes in some shops would not have been considered unusual by employees.

Within top management, there were two broad "schools of thought" regarding the firm's checklists. One group emphasized the benefits of monitoring, pointing out the importance of *Struktur* (structure) for workers, especially given the firm has 145 stores which cannot be consistently monitored personally by top management. The other group emphasized the costs of checklists, both the time involved and the notion that monitoring signals disrespect. Thus, executives had pre-RCT debates which paralleled tradeoffs in our conceptual framework.[19]

[17] For example, one sacred cow is the "sample roll" checklist, where every time a bakery bakes a batch of rolls they need to send five rolls to headquarters for potential examination or testing. Bread rolls are considered key to the firm's unique selling proposition, so this checklist could not be removed.

[18] Whether checklists take 15-20% of time at work or some fraction of this, our surveys indicate *substantial* time spent on checklists. Our focus is not on measuring time on checklists, but rather on estimating tradeoffs involved in eliminating two of them.

[19] Executives in the pro-structure school of thought helped introduce many checklists to the firm, including the operational checklist and the daily protocol. These executives have much longer tenure than those emphasizing the costs of checklists.

Operational checklist. The operational checklist is a detailed form with things to be done.[20] As seen in Figure 2, which provides the operational checklist from right before the RCT, it is a constant reminder for workers about how they are supposed to do their jobs. In our initial focus groups, many workers view the list as somewhat insulting. Employees are required to sign each item of the operational checklist every day. Workers do the checklist daily at different times.

Most items on the operational checklist are updated each month. Thus, employees spend some time reading it each day, so they know what they are signing. Some executives initially thought that without the operational checklist, stores would experience operational problems and that workers would not follow company guidelines (e.g., employees would forget to keep shelves clean and to smile at customers). Appendix C.3 provides two examples of older versions of the operational checklist, one from Aug. 2019 and one from Jan. 2017.[21]

Workers spend an average of 32 minutes per week on the operational checklist (25th percentile = 14 minutes, p50 = 24m, p75 = 35m). Store managers spend less time, devoting an average of 15 minutes per week to the checklist (p25 = 6m, p50 = 7m, p75 = 20m).

Stores receive the same information that is in the operational checklist in the form of a weekly newsletter. For example, the newsletter already tells the stores about the correct placement of Berliner doughnuts. In short, workers are constantly being reminded how to do their daily job, including in the newsletter, and then the operational checklist reminds them and requests signatures regarding what they have already been reminded of.

Daily protocol. The second checklist we study is the daily protocol, where you write down all the things that happened during the day. As seen in Figure 3, this includes how much money is in each cash register, items sold out, IT problems, and information to pass along to the next shift (this last point seemed especially appealing in bigger stores). In contrast to the operational checklist, some workers find more value in the daily protocol.

Workers spend a mean of 38 minutes per week on the daily protocol (p25 = 14m, p50 = 18m, p75 = 70m). Reflecting that the task is often done by managers, managers spend an average of 52 minutes per week on the daily protocol (p25 = 35m, p50 = 35m, p75 = 70m). Unlike the operational checklist, the daily protocol does not change over time, but it still requires significant time to provide the required information. Employees do the daily

[20] Examples include: I put the rolls at the right place in the shelves, I put the sugar on the Berliner doughnut in the right shape, I know about the Covid restrictions.

[21] Comparing the old checklists with Figure 2 reveal several things. First, the format varies over time, suggesting that workers can't necessarily just "breeze through" without reading. Second, the checklist asks about broadly similar things over time. Third, the Dec. 2020 checklist is longer than the older ones. The expanding nature of the checklist could be one reason why workers dislike it, and is also consistent with firm leadership using "management by exception" where new content is added to the checklist upon finding new problems.

protocol at the end of shifts.

The completed operational checklist and daily protocol forms are rarely examined by corporate headquarters. Indeed, many workers believe that they are never looked at by headquarters, which may heighten workers' aversion toward the checklists.

RCT setup with regional managers. Our RCT treatment is removing two checklists in treatment stores. Regional managers and sales managers were invited to a meeting on Feb. 16, 2021 with top executives and the research team. Regional managers were informed there would be a 6-month RCT and were given detailed guidelines about it. They were also given the chance to ask questions and were informed that surveys would be administered.

In the meeting, several regional managers spontaneously expressed strong views on the stores in which the treatment would be effective. This suggested possible heterogeneous treatment effects and that regional managers may have strong local knowledge on this heterogeneity. Thus, in March 2021, before knowing which stores were in control or treatment, regional managers made predictions by phone about in which stores the treatment would be effective (Appendix C.1). One coauthor interviewed all 15 regional managers (100% response rate).[22] We motivated the phone call to regional managers by stating that there was significant heterogeneity in managers' informal predictions (and rationales) for whether the treatment would work during the February 2021 meeting. To make the predictions as natural as possible, we asked regional managers for verbal responses, which we later convert into a binary response of whether it will work. For almost all responses, the conversion to binary responses is clear and unambiguous, as detailed in the Appendix. No incentives are used for this prediction because it is a subjective one.[23]

Single-treatment RCT. The RCT uses a single treatment of removing two checklists for three reasons. First, our 2020 pre-RCT survey revealed two checklists that were low-value and politically feasible to remove, so it was natural and managerially relevant for the firm to remove two at once. Second, pre-RCT power calculations indicated that we'd be well-powered to detect a treatment effect of 3% for one treatment, but possibly under-powered for multiple treatments. Finally, we expected and pre-registered substantial treatment heterogeneity, and we would be under-powered to detect heterogeneity with multiple treatments.

[22]The interviews were conducted by a coauthor (a chaired German professor) rather than an RA. This was done to show respect to senior managers, as well as to elicit serious and complete responses.

[23]Even if it were possible to incentivize predictions, there are four advantages of not using incentives. First, not using incentives avoids "incentive effects" for regional managers to influence or manipulate outcomes in stores to match predictions. Second, avoiding incentives reduces prediction salience, e.g., where predictions would "stick out" mentally for regional managers. Third, not using incentives seemed natural for higher-ranking managers. Fourth, reviewing the literature, Haaland *et al.* (2023) argue that incentives are not needed to accurately elicit beliefs and discuss how incentives can sometimes worsen elicitation.

RCT setup with store managers and workers. Store managers and workers were informed about the RCT via the firm's weekly newsletter. The information came in a message on the store intranet on Tuesday April 6, 2021 (after the Easter holiday) and also in paper form in the bundle of papers for the weekly documentation duties. In contrast to regional managers, workers and managers were not informed that there was an RCT or that the change would last for a certain period of time. Workers and managers in the treatment group indicated full awareness of the treatment. This is natural given that the RCT removed checklists that were an important part of the normal job.

The firm's message informing treatment stores about the change came from the firm's COO, the son of the CEO, which gave credibility and importance to the change:

> "At [FIRM] we constantly ask ourselves how and where we can improve to make your daily work easier. Together with the workers' council, we started discussions on day-to-day business checklists (daily protocol, expiry date checklist, weekly report, etc.) at [FIRM] last year.
>
> Starting April 6th, 2021 we will no longer process the operational checklist and the daily protocol in your store and will drop them without any replacement.
>
> This gives you more freedom to organize yourselves and we trust you that the essential processes (such as the arrangement of the products in the sales counter, covid measures, customer communication) will continue to be done in a company-compliant manner.
>
> We believe that time saved on checklists is an opportunity, which we can use for training new colleagues and communicating with customers."

The message emphasizes two factors, paralleling our predictions on direct and indirect effects. First, it emphasizes how the firm trusts workers (indirect effect). Second, it emphasizes the extra time (direct effect), and that workers should use the extra time for customers and colleagues. While one could worry that workers are being "led" to think a certain way, it would be highly artificial for a firm to make a large change like removing significant checklists without explanation. Moreover, even if workers were led somehow, it would be unlikely to explain the persistence of the main effects, or that effects vary substantially by regional manager expectations.

The framing of the letter is positive (not completely neutral), in keeping with the language used by for the firm in discussing policy changes. For example, in 2022, the firm increased hourly pay by €1 and used comparable language.

We ensured that the RCT was carried out as planned. Checklists are delivered every week to stores in a bundle. We sent an RA to monitor that the checklist bundles delivered to

treatment stores did not contain the operational checklist or daily protocol, but that control stores did.[24] We also confirmed with regional managers, the head of HR, and one sales director in May 2021 that the treatment was being carried out as planned and there were no issues with implementation.

RCT timing. The experiment began on April 6, 2021. Checklists were removed in treatment stores. The authors presented the results to the firm in Dec. 2021. Given the success of the RCT, the firm decided to roll out the treatment to control stores starting at the end of Jan. 2022. Specifically, the operational checklist remained removed, but the daily protocol was (re-)introduced, given that some workers found it useful and less onerous.

The RCT was registered on the AEA RCT Registry on April 14, 2021. Our analyses closely follow the registration. Based on theory and our interactions with the firm, we pre-registered that there would be treatment effect heterogeneity according to team size, team tenure, and regional manager predictions. We pre-registered that the RCT would last for 6 months. However, for logistical reasons, the firm left the RCT in place for 10 months.[25]

Data. We use administrative data from the firm to create two main panel datasets. First, we create a store-level panel with detailed hourly data on sales by store. This dataset also includes information on mystery shoppers. Second, we create a worker-month panel of regular employees covering worker attrition and worker absence.

The pre-treatment store manager survey was conducted in March 2021 as a phone survey conducted by RAs and a response rate of roughly 95%, with N=135. The pre-treatment regional manager survey was conducted after regional managers knew of the existence of the RCT, but before they knew which stores were in the treatment group. The main purpose of this survey was to assess regional manager beliefs about which stores would respond positively to the treatment. The during-RCT store manager survey was conducted in Nov. 2021, also by phone.

Finally, there was a during-RCT worker survey, conducted in-store with pen and paper in October 2021. This survey was conducted using a large number of RAs who personally visited the stores and collected the questionnaires.

Randomization. We conducted a stratified randomization using 4 dimensions of stratification: pre-RCT head count (above or below mean), pre-RCT sales (above or below mean), pre-RCT store ranking in the firm's performance league (above or below mean, with

[24] After 6 weeks, the firm asked if the RA could come only every couple of weeks, which we agreed to.

[25] The firm promised us that an endline survey would be conducted toward the end of the RCT. However, one of the authors had a baby and our main contact went on holiday at the same time, causing the endline survey (and end of the RCT) to be postponed for 4 months. This fortuitously gives us more data and was obviously not driven by any statistical power considerations.

this variable described more in Appendix A.3), and region (9 regions). This gives us 46 strata. Randomization was conducted using "randtreat" in Stata. As seen in Table 1 below, we observe strong balance across various characteristics.

3 Overall Results

To estimate the impact of the treatment on store-level outcomes, we use ANCOVA specifications following McKenzie (2012). Using data from the RCT period, we estimate OLS models where we control for the mean of the dependent variable in the pre-RCT period ($y_{s,pre}$), as well as year-month fixed effects (γ_t) and pre-RCT store characteristics used in the stratified randomization (X_s):

$$y_{st} = \alpha_0 + \alpha T_s + \beta y_{s,pre} + \gamma_t + X_s + \epsilon_{st}$$

where y_{st} is the outcome of store s in year-month t.[26] Throughout the paper, standard errors are clustered by store, reflecting the level of randomization. To estimate impacts on employee attrition, we consider linear probability models where the decision of whether to attrite is regressed on the treatment dummy, as well as person- and store-level controls.

Store-level outcomes. Panel A of Table 2 shows that the treatment boosts sales. Overall sales go up by 2.7%, statistically significant at the 10% level. Sales increase during the busier part of the day for bakeries (7am to 2pm) and in the less-busy time segment (after 2pm). The number of customers increases by 2.3%, narrowly missing statistical significance, suggesting that more customers are coming through the door instead of solely upselling more.

One concern with removing checklists is that it could lead to a decrease in product quality, a decrease in employee effort, and an increase in employee misbehavior. However, we see little evidence for that. Shrinkage and the mystery shopping score are both unchanged. With 95% confidence, we reject that shrinkage increases by more than 0.033 log points and we reject that mystery shopping score decreases by more than 0.13 standard deviations.

Besides estimating overall effects for the entire RCT period, it is useful to show effects over time. Panel (a) of Figure 4 shows the sales results from Equation (1) estimated separately by 5-month period, which natural to use given the RCT lasts 10 months. Effects are similar across both periods of the RCT. Even in the second period, coming 6-10 months after the treatment is introduced, checklist removal increases sales by 3%, which is statis-

[26] All our findings are unchanged to doing simple ANCOVA where we don't control for variables used in stratification. Our conclusions are also robust to controlling for strata dummies, though results are more imprecise, which we believe occurs because we have a high ratio of strata to stores (Bruhn & McKenzie, 2009), including 14 singleton strata. Including dummies for singleton strata is akin to excluding them from analysis.

tically significant at the 5% level. The relatively constant effects over time are also seen breaking up the RCT into other time divisions, such as quarters. Figure 4 also estimates differences between treatment and control stores during the post-RCT rollout, which we postpone discussing until Section 6.

While sales increases, a natural question is whether there are important aspects of operations that suffer from our treatment. Besides analyzing the overall mystery shopping score, we also analyze individual components of mystery shopping. As seen in Panel (a) of Appendix Table B1, we see no evidence of harm on any component scored by mystery shoppers. This is true across simple checks, like whether employees show their name badge, present free samples in the correct way, and upsell in the correct way, but also in terms of following guidelines on store appearance, customer interactions, and quality of the rolls.

Attrition. Panel B of Table 2 examines effects of the treatment on employee attrition. As is typical in German retail, attrition is relatively low—at least compared to US retail firms—at about 2% per month or about 25% annually, meaning about 1/4 of workers exit each year. The relatively low attrition rate places limits on statistical power.

As seen in column 1, there is no overall effect of the treatment on attrition. However, this masks heterogeneity by skill. A critical distinction is between trained and untrained employees. Trained workers already did a 3-year apprenticeship, often from the bakery firm, and the firm is keen to retain these workers receiving expensive training. In contrast, untrained workers have fewer skills and are less important to retain.[27] Trained worker attrition decreases by 0.45 percentage points (hereafter, "pp") per month, which is a 35% decrease relative to control. However, untrained worker attrition increases by a statistically insignificant 0.64pp per month, an increase of 20%. The attrition difference by worker training is statistically significant ($p = 0.02$). Untrained workers know less and may benefit from added structure; for trained workers, this structure may be unnecessary and unwanted.

Within trained workers, some are store managers and some are not, and effects are driven by managers. Manager attrition decreases by over 1pp per month, a reduction of roughly half, and statistically significant at the 10% level. The decrease is seen in raw counts: there are 10 store manager quits in control stores, but only 5 in treatment stores. Why could there be especially large effects on manager attrition? One reason is that costs of checklists are especially strong for managers, particularly for the daily protocol. Managers spend almost an hour per week on the daily protocol, while workers spend half an hour.

[27] The firm prides itself on high-quality products, and trained workers are a key part of their high-quality strategy. Appendix Table B2 shows that trained and untrained workers vary massively in characteristics. In the month before the RCT, trained workers have a median tenure of 10.7 years, while it is only 2.9 years for untrained workers. Trained workers have base pay that is one-third higher than that of untrained workers.

In pre-RCT focus groups and discussion with the firm, there was a feeling that checklists took managers away from high-value activities like mentoring and teaching workers. It is also possible that utility costs of checklists are especially bothersome for managers. Store managers are supposed to monitor and lead—by using extensive checklists, the firm may communicate that it doesn't trust managers to monitor and lead by themselves.

As seen in Appendix Figure B5, there is no evidence that the impact on manager attrition fades over time. As seen in panel (a), if anything, the treatment effect appears to grow over the RCT, but we cannot reject that effects are constant over time.

Magnitudes. How should we think about the magnitudes of the estimates? In a study in a different German bakery chain, Friebel *et al.* (2017) find that providing a team performance bonus led to an increase in sales and customers by 3%. Thus, the overall effect of the impact of removing checklists is similar to the impact of providing a team performance bonus. However, our treatment is much more cost-effective and profitable: the treatment in Friebel *et al.* (2017) involves a pay increase of 2.2%, while compensation is kept constant in our RCT. The seminal monitoring RCT by Nagin *et al.* (2002) look at effects on suspicious calls, but do not have data on sales.

Another study of a particular management practice is an RCT on work from home by Bloom *et al.* (2014). This study finds that working from home led to a 4% increase in calls per minute, which is also similar to our effect on sales. Bloom *et al.* (2014) also find that work from home reduces attrition by half, which is broadly similar to the attrition effect we observe on managers. However, the effect we observe in stores in which the treatment is predicted to work is larger (though with a large standard error, meaning we cannot reject that the effect would be at half, though we can reject that the effect is zero).

An RCT by Alan *et al.* (2023) also observes reductions in attrition concentrated among managers. Working with Turkish firms, the authors examine the impact of a module by a consulting company designed to improve the relational atmosphere at work. This module reduces manager attrition by roughly 80% while having much smaller impacts on worker attrition. The impact of checklist removal on manager attrition is thus broadly similar to the effect of a workplace relational module.

In sum, removing two perceivedly low-value checklists in our setting yields treatment effects on the order of some of the most promising and highly regarded management interventions. At the same time, we believe that our effect sizes are plausible. While it may be surprising that removing checklists has quantitatively substantial effects, we emphasize that the ones removed were perceived as particularly onerous and ineffective.

Using worker surveys and customer reviews to understand mechanisms.

Given the sizable effects, we ask why they occur. Why does sales increase? Why does retention increase for many workers? To shed light on these questions, we survey workers during the RCT, and we also scrape data from Google reviews of the stores. The worker surveys illustrate how the treatment affects worker attitudes, and the Google reviews show how effects manifest in terms of store operations.

Survey of worker attitudes. Panel A of Table 3 shows that the treatment increased workers' commitment to their store by 0.21 standard deviations (σ), statistically significant at the 5% level. The treatment also increased workers' sense of trust between headquarters and workers by 0.27σ. These estimates are consistent with the notion in the conceptual framework that removing checklists can convey trust and build commitment. Another possible theory is that freeing up time on checklists allows managers and workers to invest more time in training. However, there was no effect of the treatment on workers' perception of whether their latest hire was well-trained. The final column shows that there is no effect of the treatment on basic quality control. Workers were asked about whether the firm continued to do basic quality control over several aspects of work, and we observe that checklist removal did not significantly limit whether stores engaged in basic quality control.

Customer reviews. To further understand the impact on sales, Panel B examines the impact of the treatment on star ratings on Google reviews. As seen in column 1, the treatment led to a 0.20 point increase in Google star reviews. Online reviews are known generally to be heavily skewed to the left in average score (Tadelis, 2016), and ours are no exception, with an average rating in control stores of 4.1. As seen in columns 2-6, the share of 1s, 2s, 3s, and 4s is lower in treatment stores, while the share of 5s significantly increases. As seen in column 7, the treatment does not significantly affect the number of reviews a store gets, suggesting that the treatment affects customer experiences instead of selection into reviews.

To shed light on what happens in treatment stores to warrant higher customer ratings, Panel C of Table 3 performs a text analysis of the Google reviews. Using the text of scraped reviews, an RA measured whether there was anything positive said about the product, service, shop appearance, speed of service, value for money, and product availability. We describe the classification procedure in greater detail in Appendix A.8, including the issue that many reviews do not contain text.

The text analysis yields several findings. First, the share of reviews mentioning speed of service increases by roughly 160%, from 0.5% in control store to 1.3% in treatment stores, an increase that is statistically significant at 5%. Second, the share of reviews mentioning something positive about the product increases by 4.4pp, an increase of 17% relative to Control stores. Employees who feel more committed and trusted and who have more time may put greater effort into displaying and producing high-quality baked goods. The effect on

quality could also reflect the speed channel, where stores that sell faster have fresher products and fresher products are regarded as higher-quality. There is also a 42% increase in positive comments about shop appearance, though this increase is statistically insignificant, as well as a statistically insignificant increase in comments about value for money (consistent with the perceived increase in quality).[28]

Overall, customers in treatment stores report being served faster and receiving a higher-quality product, and this manifests itself in higher overall ratings.[29]

4 Treatment effect heterogeneity

Regional managers had strong beliefs about in which stores the treatment would be successful. Thus, we focus analysis of treatment heterogeneity on regional manager predictions. In stores where regional managers predict the treatment to work, the treatment increases sales and decreases trained worker and manager attrition. Moreover, effects on sales and attrition are differentially stronger, in economic and statistical significance, in stores where the treatment is predicted to work. After presenting these results, we consider other pre-registered dimensions of heterogeneity: team size and tenure.

Heterogeneity by regional manager predictions. Table 4 separates treatment effects on store outcomes by regional manager predictions, showing that effects are much stronger in stores where regional managers predict the treatment to be beneficial. In such stores (Panel A), sales increase by 5.2%, statistically significant at the 5% level, with similar increases among busy and slow sales. The number of customers increases by 4.8%, and shrinkage—a combination of wasted product and theft—goes down 2.4%, though this decrease is not statistically significant. In contrast, for stores where the treatment is not predicted to work (Panel B), the effects on sales are zero and shrinkage *increases* by 2.4%, though this decrease is also insignificant.

Panel C of Table 4 shows p-values regarding equality of the treatment effect in stores where regional managers predict the treatment to work and not to work. In addition to showing two-sided p-values, which are common for analyzing interaction terms, we also present one-sided p-values, which we believe are more appropriate given the explicitly one-

[28]The qualitative characteristics are modestly correlated with one another, but appear distinct. In a principal component analysis, the first component, receiving positive weight from all the characteristics, only explains 35% of the variance. Thus, it is not the case that all the characteristics listed here appear to represent the same underlying trait.

[29]In their RCT on group bonuses, Friebel *et al.* (2017) argue that their observed increase in sales is driven by an increase in speed of service, suggesting that this is a plausible mechanism for treatment effects in bakeries. Friebel *et al.* (2017) do not have data on customer reviews, so our evidence is more direct.

sided prediction of store managers (i.e., dividing stores in the ones where the treatment will work and ones where it will not work). Under the one-sided p-values, all interactions with respect to sales, customers, shrinkage are statistically significant at $p = 0.06$ or less.

Returning to Figure 4, panels (c) and (d) show results over time by regional manager expectations. Restricting to stores where regional managers predict the treatment will work, the treatment is pronounced in the first quarter, consistent with the large distaste that many workers and managers at the firm expressed toward checklists. However, we cannot reject that the treatment effect is constant over the RCT.[30]

Table 5 shows the treatment effect on attrition separating by regional manager predictions. Overall attrition decreases by 0.5pp in stores where it is predicted to work and increases by 0.5pp in stores where it is predicted not to work. Both effects are not statistically significant, but the difference is statistically significant ($p = 0.03$ under a one-sided test). Turning to trained workers, who are the firm's main focus for attrition, attrition decreases by 1pp or roughly 2/3 in stores where the treatment is predicted to work, and this effect is significant both for trained non-managers and for managers. The drop in trained worker attrition is entirely driven by stores where treatment is predicted to work.

The drop in manager attrition is also fully driven by stores where the treatment is predicted to work. In those stores, attrition decreases by 2.2pp per month, essentially a complete reduction relative to control. In the raw data, in stores where the treatment is predicted to work, there are 8 store manager quits in control stores, but only 1 in treatment stores. In contrast, in stores where the treatment is not predicted to work, the effect is zero. This difference is statistically significant at the 5% level.[31]

Robustness of regional manager predictions as a source of heterogeneity, and other heterogeneity dimensions. To assess the robustness of regional manager predictions as a source of heterogeneity, we apply alternative approaches to estimating heterogeneous treatment effects, all of which strongly support robustness. We focus on the effect on sales and begin with linear regressions containing interactions terms of the treatment variable with many additional pre-RCT store characteristics (one at a time). The interaction

[30] Appendix Figure B7 shows the impact of the treatment over time in stores where the treatment is predicted to work using an event study framework. In contrast to our baseline ANCOVA results, we use store fixed effects and focus on the interaction of treatment status with dummies for quarter since the start of the RCT. Here, too, one cannot reject that the treatment effect is constant throughout the RCT.

[31] Interestingly, store manager attrition is 3x higher in stores where the treatment is predicted to work compared to not to work. There are two intuitive reasons for this, both grounded in our conceptual framework. First, regional managers may have private information about which managers are most at risk at quitting, perhaps in part due to excessive monitoring and an overly bureaucratic culture, and they predict that the treatment will be most effective for such managers. Second, stores where the treatment is predicted to work may have fewer problems, and store managers such stores exhibit positive selection in their quits.

on treatment X manager expectation remains highly robust, as seen in Appendix Table B4. However, a concern with this approach is that the correlation between pre-RCT characteristics may lead to efficiency loss and, especially on small samples like ours, biased estimates. Thus, we rely on machine learning methods designed to estimate statistical relationships with many, possibly correlated, regressors. First, we perform an elastic net-regularized linear regression which prunes variables that do not have sufficient predictive power (Zou & Hastie, 2005). The interaction on treatment X regional manager expectation is one of only a small number of the interaction terms to escape pruning, and is much more predictive than the rest.

We next try two other approaches that work with more general relationships than linear regression: sorted effects (Chernozhukov *et al.* , 2018) and causal random forests (Wager & Athey, 2018). The sorted effects method evaluates the conditional treatment effect for each observation (treated or not) given its characteristics and the user-specified function that links the outcome of interest to those characteristics and whose parameters are estimated from data. The sorting of observations by the magnitude of the estimated treatment effect produces the most and least affected groups, thereby enabling comparisons between these groups in terms of their characteristics. The causal random forests method works broadly similarly, the main difference being that it uses a nonparametric procedure (random forest), rather than a user-specified equation as in sorted effects, to evaluate conditional treatment effects. We adapt both methods to allow for clustering at the store level and account for multiple hypotheses testing in multidimensional comparisons across the most and least affected groups.

The application of the above two methods reveals substantial treatment effect heterogeneity across stores. To summarize this heterogeneity, for each method, we group stores into quartiles of the estimated store-specific treatment effect and re-estimate the average treatment effect on log sales for each quartile using equation (1), reporting the findings in Table B9. The two methods produce broadly similar treatment effect estimates, and there is also a high degree of agreement in the classification of stores into treatment effect quartiles: the two methods produce the same classification results for more than three-quarters of the stores.

Table B10 reports the averages and standard deviations of pre-RCT headcount and tenure and regional manager predictions by quartile and by method. Stores where regional managers predicted the treatment to work are much more likely to be highly positively affected (quartiles 3 or 4), which supports our earlier results showing the importance or regional manager predictions as a source of heterogeneity. The highly affected stores also tend to be smaller in headcount and employ more experienced workers, even those this latter

difference is less significant statistically.

Mechanisms for the regional manager predictions. Why are regional manager expectations predictive of the treatment effect? What is the rationale for their predictions, both positive and negative? To address this question, we use the raw text from regional managers' pre-RCT predictions. The text of regional manager predictions is provided in Appendix Tables B7 and B8.[32]

Looking through the responses, there are two salient features of text responses for stores where regional managers predicted that the treatment would work. First, in many cases, regional managers mention that workers will enjoy having less checklists. For one store the regional manager said that workers "Would be very happy about less bureaucracy, less work as a result, do not like to work with notes and strict rules." This explanation would fall under the utility cost of monitoring described in Section 1. Second, in many cases, regional managers talked about how teams would be unlikely to face problems, especially because the team already had good communication. An example prediction is that one store "Could live without bureaucracy, very communicative branch management." Some predictions mention both that reducing monitoring will be good for worker utility and that there are no anticipated problems. For example, one manager predicted that the "Team will be glad when operational list is gone. No problems expected. Will work out!"

Table 6 summarizes key facts about regional manager predictions. In stores where regional managers believe the treatment will be successful, in 37% of predictions, regional managers mention something about checklist removal benefiting worker utility. Likewise, in 71% of predictions, regional managers mention something related to ability to overcome problems. Thus, regional manager predictions strongly support both (1) the traditional economic view of monitoring as a way of addressing problems (Holmstrom, 1979; Halac & Prat, 2016) and (2) theories emphasizing utility costs of monitoring (Falk & Kosfeld, 2006).

Table **??** examines correlates of regional manager predictions, showing that observable characteristics explain only a modest share of regional manager predictions ($R^2 = 0.17$). The largest predictor of regional manager predictions if a store's pre-RCT mystery shopping score, with regional managers believing that removing checklists will be more effective in stores with higher pre-RCT mystery shopping scores. Pre-RCT Log Sales and pre-RCT mean worker tenure are not significant predictors of regional manager expectations.

A natural concern in interpreting the results on regional manager predictions is whether results could be due to managers behaving differently in treatment vs. control stores. However, in the predictions, no regional manager said anything about an intent to behave dif-

[32]The text recorded is the notes taken during the phone call with regional managers. Written notes were taken instead of recording calls due to privacy considerations in Germany.

ferently in treatment vs. control stores, such as by visiting treatment stores more often. A different concern is that regional managers might have private information not about the efficacy of treatment, but rather about the coming of external shocks to stores (e.g., there will be a large festival next to a store in the coming months). However, no regional managers said anything in their prediction about external shocks.

5 Profits, Further Analyses, and Validity Threats

5.1 Profit Implications of Results

The RCT is highly profitable for the firm, with a benefits to cost ratio of over 50:1. To see this, note that the treatment effect on sales of 2.7% implies that stores receive an extra roughly €2,050 per month, given average monthly sales per stores of €75k (i.e., (exp(.027)-1))*75k = €2050. Aggregated over 145 stores and the 10 months of the RCT, the total revenue benefit is almost €3m. Using a share of value added in bakery chains of 0.56 (Friebel *et al.* , 2022), the gains from the RCT are .56*3m = €1.7m.

In contrast, checklists are inexpensive to remove. Counting up hours spent by the project team, as well as time spent by executives, regional managers, and others in implementing the RCT, the total cost of time seems unlikely to exceed €31k (details in Appendix A.9). This yields a benefit to cost ratio on the order of 50:1 or more. We view such a ratio as relatively conservative and would be even larger if it accounted for the reduction in turnover among trained workers.[33] This is one of the largest benefit to cost ratios that we are aware for any management practice.[34]

Another to assess profitability is to look at profit per store-month. Using a profit margin of 1% (as indicated by top management), we estimate that our RCT more than doubles the profit margin.

5.2 Direct and Indirect Effects of Checklist Removal

Separate from regional manager predictions, what drives the improvements in store performance that we observe, as well as the reductions in manager attrition? Are people using the extra time that they have to perform other tasks, which we can think of as the direct effect of checklist removal? Or is there some other mechanism such as increased happiness, trust,

[33] It is fairly cheap for the firm to find untrained workers. Trained workers are much more expensive to hire and train, and trained workers are much harder to find. Adding the direct hiring costs of trained workers to the analysis, which are significant (Blatter *et al.* , 2012), would further increase the benefit to cost ratio.

[34] Friebel *et al.* (2022) have a benefit to cost ratio of roughly 2:1.

or respect? The regional manager predictions indicate that at least for some stores, regional managers believe that indirect effects will be present, believing that removing checklists will make workers happier.

Appendix Table B5 examines heterogeneity in the overall treatment effect on sales based on the amount of time that stores spend on the daily protocol in the pre-RCT period. As seen by the key interaction term, there is no evidence that the treatment effect on sales varies with pre-RCT time spent on the daily protocol. Rather than looking at the quantity of time, one can instead focus on when stores tend to do the daily protocol in the pre-RCT period. We find no evidence that the treatment effect is larger during the time periods when stores generally do the daily protocol. Recall that the daily protocol takes more time compared to the operational checklist.[35]

These two pieces of evidence fail to support direct effects of the treatment, i.e., that checklist removal increases sales by allocating extra time to other activities. One additional piece of evidence in favor of indirect effects comes via the firmwide rollout, which we discuss shortly below.

5.3 Are the Sales Effects Due to Turnover?

A natural question is whether the improvements in sales we observe are due to the lower turnover of trained workers and managers (Cai & Wang, 2022). A formal mediation analysis provides no evidence that the sales effect is mediated by lower turnover (Appendix A.10). In addition, the time paths of sales and turnover effects are different. If we focus on quarters instead of 5-month periods, sales improves immediately in the first quarter, whereas turnover effects are negligible in the first quarter of the RCT and become larger each quarter. Together, this suggests that our sales effects are unlikely to be driven by our turnover effects.

5.4 Threats to Validity

Control store frustration. Could it be the case that our treatment effects are driven not by positive change in the treatment stores, but rather by something negative in control stores? Perhaps employees in control stores were frustrated they were not selected for treatment. We were very mindful of this point, and thus, in all stores, workers and store managers were not

[35] Rather than looking at heterogeneity in pre-RCT time use, we can also exploit a direct question we asked to store managers about whether checklist removal helped freed up time in the stores (yes or no). We see no evidence that the treatment effect on store outcomes varies based on whether store managers believed that the treatment help free up time, though of course such an analysis requires the caveat that it is heterogeneity based on an endogenous variable (i.e., whether time is freed up).

informed that they were part of an RCT, and employees in control stores were not informed about any possible removal of checklists. Still, people may talk to one another, and indeed, in designing the RCT, the head of HR thought that it's likely that some store managers would talk to one another.

To address and anticipate any contamination, regional managers were provided with written guidelines (see Appendix C.4) on what to say if workers or store managers asked about checklist removal. Specifically, people were told that there was a pilot project with researchers from the University of Cologne in some stores, randomly selected for fairness reasons so that everyone has the same chance, and with the lottery done jointly with the research team and works council. Workers were told they could contact the works council with any questions.[36]

To measure the effect of any contamination, workers and managers were surveyed in November 2021, 8 months into the RCT, on whether they knew about a pilot project where checklists were removed in some stores. About 3/4 of store managers and 1/2 of worker employees in control stores knew about the pilot project (i.e., the RCT). However, they expressed essentially no annoyance about the existence of the RCT. For people who knew about the RCT, the average level of annoyance was only a very low 2 on a scale from 1 to 7. All our results are robust to dropping the small number of stores where store managers or workers expressed any level of annoyance.

That annoyance is so low is quite expected. Neither the researchers nor the works council head received any compladints. Furthermore, people at the firm are used to pilots where some things are done in some stores, but not others.[37] That people also do not care about the existence of RCT squares with other studies like Bloom *et al.* (2014) where workers are explicitly told that they are randomized into work from home or not.

Regional manager effort. Could the effects we observe be driven by regional managers reallocating effort between control and treatment stores (e.g., regional managers stop spending time on control stores to focus on improving performance in treatment stores)? Anecdotally, the firm believes this is very unlikely because regional managers had other key concerns during the RCT, namely, the issue of covid.[38] Finally, using the during-RCT survey of store managers, we see no impact of the treatment on how much time store managers report interacting with regional managers.

[36]Providing this helps establish trust, as Germans have strong trust toward works councils, which are chosen democratically. When German employees have issues at work, they contact their council. At our firm, we know that the council would be willing to contact us if there were any problems because the counsel contacted us once when one store manager didn't receive their voucher for participating in a pre-RCT survey.

[37]Past pilots include introducing high-quality coffee or reducing prices, all in some stores but not others.

[38]For example, both the head of HR and a sales manager believed that the RCT was no longer especially salient to regional managers.

Separate from the overall treatment effect, could regional manager effort drive the fact that the treatment effect is entirely concentrated in stores where regional managers predicted that the treatment would work? As mentioned above, we avoided giving incentives for predictions precisely with this concern in mind. In addition, there was no career benefit for regional managers of predicting correctly. Finally, in the during-RCT survey of store managers, there is no impact of the treatment on time with regional managers even when restricting to stores where regional managers expected the treatment to work.

Hawthorne effects. A separate concern in any RCT is whether subjects could alter their behavior in order to please the researchers (Levitt & List, 2011). As stated above, workers and store managers were not informed that they were part of an RCT, though there was some information leakage. We have two responses to this concern. First, our treatment effects persist 10 months into the future. It seems unlikely to us that Hawthorne Effects would stay for so long. Second, Hawthorne Effects cannot easily explain our key heterogeneity results by regional manager expectations.[39]

Contemporaneous policy changes. Another concern in any RCT is the presence of contemporaneous policy changes. However, this was not the case in our firm. In the data, if one regresses log base wage on the treatment dummy and controls, as in Panel B of Table 2, there is no significant impact.

Multiple hypothesis testing. In a study addressing heterogeneous treatment effects and multiple outcomes, one worries that treatment effects could be spurious due to multiple hypothesis testing. The main way that we address this point is through the rigorous pre-registration of our RCT. Our main outcomes were listed in the pre-registration before the RCT began, and we explicitly say that our primary outcome is store sales. In addition, we explicitly say that our heterogeneity analysis will focus on heterogeneity according to regional manager expectations.

Estimates combining all stores are somewhat noisy. The estimates on the overall effect of checklist removal in Section 3 are sometimes somewhat noisy and we cannot always reject null effects with high confidence. A key reason for this is that we are studying an intervention, checklist removal, where it is very likely theoretically that effects will be highly heterogenous. Our overall results pool stores where the treatment is predicted to work and where the treatment is predicted not to work. We hope that our methodology of asking managers to predict whether the treatment will be effective can be useful in studying other management interventions.

[39] Hawthorne effects could only drive heterogeneity by regional manager expectations if regional managers had private information about the extent of Hawthorne effects across stores. No regional manager mentioned Hawthorne Effects in their explanations about why the treatment would work in particular stores.

Covid. The RCT took place in April 2021 - January 2022. Is there any external validity concern from covid? The covid lockdown in Germany was almost over in March 2021 and was over by May 2021, and food retail establishments including bakery stores like ours were excerpt from the lockdown. All stores were fully open during the RCT, including the coffee area of stores.[40] Both the operational checklist and daily protocol were used before, during, and after the pandemic. The operational checklist often had an item or two related to covid (see Figure 2 for an example), but these were otherwise unaffected.

Autonomy and local information. Separate from utility benefits of removing checklists, one alternative explanation for our effects could conceivably be that removing checklists gives workers autonomy to make better decisions. That is, they are no happier or more committed to the firm, but not having rules could allow workers to exercise better judgment, whether in terms of how to speak to customers (e.g., "Good morning" vs. "Hi") or how to present or place the products.

There are several pieces of evidence against this interpretation. First, the RCT did not actually change workers' autonomy. Everyone was still required to give a certain number of cookies and interact with customers in a certain way—they simply were no longer required to sign forms guaranteeing that they had behaved in a certain way. Workers were still reminded of the contents of the operational checklist in the newsletter delivered on the firm intranet. Second, aspects of the mystery shopping score are still monitored via mystery shopping. Finally, on the worker survey, we measure whether workers feel more autonomous as a result of the RCT, and we see no difference between treatment and control stores, despite observing that they are more committed and feel greater trust.

Could the checklists have been initially useful? While the RCT finds that removing the checklists is beneficial, this does not necessarily imply that the checklists were always harmful. Perhaps the checklists were initially valuable when first introduced, e.g., as a way of instilling good routines. We cannot rule this out, but there is evidence against this possibility. First, the treatment effects on sales do not vary with a store's average worker tenure, either tenure at the start of the RCT or average tenure over time. Second, the impact on sales does not vary with the age of a store. If the operational checklist and daily protocol are useful for establishing routines, one would expect that removing them might be less beneficial in younger stores, but we see no evidence of that.[41]

[40] The coffee areas were closed during the middle of the lockdown, but were open by the start of the RCT.

[41] We observe that 28 of the 145 stores begin operations after January 2014. Even if checklists were initially valuable, and became less valuable over time, this does not mean that the firm was wrong to remove them. It just means that benefits may fade once many new workers join the firm.

5.5 External Validity

Like all RCTs, our results are specific to our organizational context, namely, a leading firm in the German bakery chain industry. Do other firms use checklists and in what form? To address this question, we conducted informal surveys with a set of comparable firms, namely, German bakery chains with more than 50 branches, regarding checklists. We surveyed five chains in a large metropolitan area in Germany (different from that of our study firm). We conducted 21 in-person interviews in total (3-6 per firm) using a research assistant. All five bakery chains report using checklists, and this simple fact is consistent across the employees interviewed. While checklists vary by firm, the checklists tend to focus on operational issues, broadly similar to the issues in the operational checklist. The interviews suggested also that workers spent significant time on checklists.

The heterogeneity of the effects we observe suggests that returns to treatments like ours may be heterogeneous across firms. In contexts where problems come up frequently and/or are expensive to deal with, checklists may play a crucial role and their elimination could be harmful. On the other hand, in contexts where having a checklist is time-consuming or is interpreted as a sign of mistrust, eliminating checklists may be more beneficial.

Turning from external validity regarding firms, how should we think about external validity in regard to tasks? Our RCT removed two low-value checklists. Should we be concerned that these were not "typical" monitoring tasks or that we did not randomly select which monitoring tasks to remove? We believe the answer is decidedly not. Our contribution is **not** to estimate the return to removing typical or randomly chosen checklists. If we had done so, effects would likely have been far more negative, as economists generally believe that firms try to optimize and that most duties serve some purpose. Instead, we view our key contribution as providing a methodology that researchers can use in other contexts to study removing checklists and other tasks.[42]

6 Firmwide Rollout

The firm was quite satisfied with the outcomes of the RCT. The research team presented preliminary results from the RCT to the study firm in December 2021. Given the success of the RCT, the firm immediately rolled out checklist removal to the whole firm, implemented at the end of January 2022. Beyond the quantitative results of the RCT, the firm regularly receives informal feedback from workers and managers at the stores.

[42] We believe that our methodology could be useful in other organizational contexts, as long as workers and managers are comfortable providing honest opinions about the value of tasks.

However, in the firmwide rollout, only the operational checklist was removed. The daily protocol was reinstated. A key reason was that feedback from workers and managers supported some value to having the daily protocol. Some workers and managers thought that having the protocol was useful for coordinating production (Alonso *et al.* , 2008). In treated stores, managers and workers in the during-RCT surveys were asked about whether they agreed that it was good to remove each of the checklists. Both workers and managers strongly supported the removal of the operational checklist, giving mean agreement levels of 5.7 and 6.0, respectively, on a scale between 1 and 7, where 1 means strongly disagree and 7 means strongly agree. However, workers and managers were less supportive of removing the daily protocol, with mean agreement levels of 4.9 and 3.1, respectively. As of September 2022, i.e., 8 months after the rollout, the firm has continued not having the operational checklist.

Sales and trained worker attrition. Figure 4 shows the difference between treatment and control stores in sales during the post-RCT period. Results also appear in Appendix Table B6, which repeats Table 2 while restricting to the post-RCT period. Under the firmwide rollout, the difference in sales between treatment and control stores drops from 2.7% to 1.6%. This is a drop in the coefficient of almost half, and the difference is no longer statistically significant. The difference in trained worker attrition is starker, going from a coefficient of -0.44pp during the RCT to a coefficient of 0.46pp during the rollout.

In sum, when checklists are standardized across treatment and control stores, the difference in sales almost halves, and the difference in trained worker attrition completely disappears. Results broadly support the stability of the treatment effects of checklist removal.

Why wasn't the rollout differentiated by store? Given the observed heterogeneity results, why didn't the firm implement checklist removal only in stores where regional managers expected the treatment to work? There are several reasons. First, while there are sizable positive effects of checklist removal in stores where regional managers expected the treatment to work, there are not sizable negative effects of checklist removal in stores where managers expected the treatment not to work. Thus, in stores where regional managers predicted not to work, checklist removal was roughly neutral instead of harmful. Second, while the firm thought it was logistically feasible to differentiate store procedures for the period of an RCT, the firm did not think that this would be feasible from a longer-run perspective. The firm often adds new stores, and would need to be surveying regional managers about whether the treatment would be effective in a new store, and regional managers would need to do this with limited information about the characteristics of the new store.

Conclusion from the rollout. The message from the RCT and reinforced by the

rollout is not that all checklists are bad. Rather, the firm discovered that certain checklists were not a good fit for the organization. The firm eliminated the checklist that many workers regarded as annoying or demeaning. However, it kept the daily protocol, which helps coordinate production across shifts and days of the week.

7 Conclusion

Scholarship often focuses on benefits of monitoring, both in general and for checklists. In a large German bakery chain, we use a novel methodology of surveying workers and managers about the relative value of different checklists, documenting wide variation across checklists in time cost and perceived value. Removing two of the perceivedly lowest-value checklists improves average store performance as measured by sales and store manager attrition. The size of performance benefits are comparable to those from starting major management practices like team bonuses, but the costs are much smaller. There is no evidence of more unexpected problems in treated stores, as evidenced by mystery shopping, though with the caveat that some problems are rare and hard to observe. In online reviews, customers of treated stores give higher ratings, and the text of reviews indicate improvements in speed of service and product quality. In surveys, treated workers feel more committed to their stores and perceive greater trust between workers and managers at the firm, relative to control workers.

Pre-RCT conversations with managers suggested that treatment effects could be highly heterogeneous across stores, broadly consistent with work in economic theory on the signaling effects of monitoring (Benabou & Tirole, 2003). Thus, we asked regional managers to predict treatment efficacy for all their stores before treatment assignment, and we find that positive treatment effects on performance are entirely concentrated in stores where regional managers predicted the treatment to be effective. This result cannot be explained by regional managers spending more time with, or being otherwise partial to, those stores. Rather, it suggests that managers have private knowledge about which stores are most likely to benefit from checklist removal. Text analysis of regional manager predictions indicate that managers weighed heavily which stores would benefit from the added structure of checklists, and which stores' workers would experience utility benefits from checklist removal. The treatment also works better in smaller stores, presumably because their teams can better coordinate without formal procedures involving checklists.

Our findings suggest an expansive view of monitoring beyond the classical conception as a costly tool for detecting low effort (Holmstrom, 1979). We find that some types of monitoring can harm firm performance and be a disamenity for skilled workers.

Our evidence suggests that the intervention of removing these checklist is a "win-win,"

improving firm profits and worker welfare, at least for trained workers. The conclusion on worker welfare is supported by the fact that trained turnover decreases and workers expressed satisfaction with checklist removal. Thus, our removal of workplace control and monitoring exemplifies how management interventions can be beneficial for both sides, and improve overall welfare. The Google reviews (and sales effects) suggest that customers may have benefitted from the RCT too.

Our work also connects to work in personnel economics on the value of managers. Besides motivating and teaching employees, middle managers may be valuable because of knowledge they have about their teams. This knowledge seems hard to codify, as regional manager predictions are only weakly correlated with observed store characteristics.

The RCT lasted for 10 months before checklist removal was rolled out firmwide. This is a long period of time compared to most management practice RCTs (Bloom *et al.* , 2020), and the impact on sales is strongly present in months 7-10 of the RCT. The rapid firmwide rollout of checklist removal is a testament to the durability of the treatment effects.

Given our RCT reveals that the firm was not fully optimizing before the RCT, a natural question is why. One conjecture is that top management didn't know the share of people who are bothered by the onerous checklists and didn't know how costly it was for workers. Top management had a sense of how much time the operational checklist and daily protocol took, but they may have underestimated what share of people would find it highly distasteful.

We look forward to future RCTs examining the benefits and costs (both direct and indirect) of monitoring and workplace control. We believe that eliciting expert opinions regarding the likely effect of an RCT in particular units is a methodologically novel and useful tool to help detect treatment effect heterogeneity.

References

Adams, Gabrielle S., Converse, Benjamin A., Hales, Andrew H., & Klotz, Leidy E. 2021. People Systematically Overlook Subtractive Changes. *Nature*, **592**(7853), 258–261.

Alan, Sule, Corekcioglu, Gozde, & Sutter, Matthias. 2023. Improving Workplace Climate in Large Corporations: A Clustered Randomized Intervention. *Quarterly Journal of Economics*, **138**(1), 151–203.

Alonso, Ricardo, Dessein, Wouter, & Matouschek, Niko. 2008. When Does Coordination Require Centralization? *American Economic Review*, **98**(1), 145–79.

Bandiera, Oriana, Best, Michael Carlos, Khan, Adnan Qadir, & Prat, Andrea. 2021. The Allocation of Authority in Organizations: A Field Experiment with Bureaucrats. *Quarterly Journal of Economics*, **136**(4), 2195–2242.

Belot, Michele, & Schröder, Marina. 2016. The Spillover Effects of Monitoring: A Field Experiment. *Management Science*, **62**(1), 37–45.

Benabou, Roland, & Tirole, Jean. 2003. Intrinsic and Extrinsic Motivation. *Review of Economic Studies*, **70**(3), 489–520.

Blader, Steven, Gartenberg, Claudine, & Prat, Andrea. 2020. The Contingent Effect of Management Practices. *Review of Economic Studies*, **87**(2), 721–749.

Blatter, Marc, Muehlemann, Samuel, & Schenker, Samuel. 2012. The Costs of Hiring Skilled Workers. *European Economic Review*, **56**(1), 20–35.

Bloom, Nicholas, Sadun, Raffaella, & Van Reenen, John. 2012. The Organization of Firms Across Countries. *Quarterly Journal of Economics*, **127**(4), 1663–1705.

Bloom, Nicholas, Liang, James, Roberts, John, & Ying, Zhichun Jenny. 2014. Does Working from Home Work? Evidence from a Chinese Experiment. *QJE*, **130**(1), 165–218.

Bloom, Nicholas, Brynjolfsson, Erik, Foster, Lucia, Jarmin, Ron, Patnaik, Megha, Saporta-Eksten, Itay, & Van Reenen, John. 2019. What Drives Differences in Management Practices? *American Economic Review*, **109**(5), 1648–83.

Bloom, Nicholas, Mahajan, Aprajit, McKenzie, David, & Roberts, John. 2020. Do Management Interventions Last? Evidence from India. *AEJ Applied*, **12**(2), 198–219.

Boorman, Daniel. 2001. Today's Electronic Checklists Reduce Likelihood of Crew Errors and Help Prevent Mishaps. *ICAO Journal*, **56**(1), 17–21.

Bruhn, Miriam, & McKenzie, David. 2009. In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *AEJ: Applied*, **1**(4), 200–232.

Bryan, Gharad T, Karlan, Dean, & Osman, Adam. 2021. *Big loans to small businesses: Predicting winners and losers in an entrepreneurial lending experiment.* Tech. rept. National Bureau of Economic Research, WP 29311.

Cai, Jing, & Wang, Shing-Yi. 2022. Improving Management through Worker Evaluations: Evidence from Auto Manufacturing. *Quarterly Journal of Economics*, **137**(4), 2459–2497.

Chernozhukov, Victor, Fernandez-Val, Ivan, & Luo, Ye. 2018. The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages. *Econometrica*, **86**(6), 1911–1938.

Dal Bó, Ernesto, Finan, Frederico, Li, Nicholas Y, & Schechter, Laura. 2021. Information Technology and Government Decentralization: Experimental Evidence from Paraguay. *Econometrica*, **89**(2), 677–701.

de Rochambeau, Golvine. 2020. *Monitoring and Intrinsic Motivation: Evidence from Liberia's Trucking Firms.* Mimeo, Science Po.

DellaVigna, Stefano, & Pope, Devin. 2018. Predicting Experimental Results: Who Knows What? *Journal of Political Economy*, **126**(6), 2410–2456.

DellaVigna, Stefano, Pope, Devin, & Vivalt, Eva. 2019. Predict science to improve science. *Science*, **366**(6464), 428–429.

Dessein, Wouter. 2002. Authority and Communication in Organizations. *Review of Economic Studies*, **69**(4), 811–838.

Dessein, Wouter, & Santos, Tano. 2006. Adaptive Organizations. *Journal of Political Economy*, **114**(5), 956–995.

Dickinson, David, & Villeval, Marie-Claire. 2008. Does Monitoring Decrease Work Effort?: The Complementarity Between Agency and Crowding-out Theories. *Games and Economic be-*

havior, **63**(1), 56–76.

Dube, Arindrajit, Naidu, Suresh, & Reich, Adam D. 2022. *Power and Dignity in the Low-Wage Labor Market: Theory and Evidence from Wal-Mart Workers*. Working Paper 30441. National Bureau of Economic Research.

Duflo, Esther, Hanna, Rema, & Ryan, Stephen. 2012. Incentives Work: Getting Teachers to Come to School. *American Economic Review*, **102**(4), 1241–78.

Dustmann, Christian, & Schoenberg, Uta. 2012. What Makes Firm-Based Vocational Training Schemes Successful? The Role of Commitment. *American Economic Journal: Applied Economics*, **4**(2), 36–61.

Dustmann, Christian, Ludsteck, Johannes, & Schönberg, Uta. 2009. Revisiting the German Wage Structure. *Quarterly Journal of Economics*, **124**(2), 843–881.

Ellingsen, Tore, & Johannesson, Magnus. 2007. Paying Respect. *Journal of Economic Perspectives*, **21**(4), 135–150.

Ellingsen, Tore, & Johannesson, Magnus. 2008. Pride and Prejudice: The Human Side of Incentive Theory. *American Economic Review*, **98**(3), 990–1008.

Falk, Armin, & Kosfeld, Michael. 2006. The Hidden Costs of Control. *American Economic Review*, **96**(5), 1611–1630.

Friebel, Guido, Heinz, Matthias, Krueger, Miriam, & Zubanov, Nikolay. 2017. Team Incentives and Performance: Evidence from a Retail Chain. *American Economic Review*, **107**(8), 2168–2203.

Friebel, Guido, Heinz, Matthias, & Zubanov, Nikolay. 2022. Middle Managers, Personnel Turnover, and Performance: A Long-Term Field Experiment in a Retail Chain. *Management Science*, **68**(1), 211–229.

Friebel, Guido, Heinz, Matthias, Hoffman, Mitchell, & Zubanov, Nick. 2023. What Do Employee Referral Programs Do? Measuring the Direct and Overall Effects of a Management Practice. *Journal of Political Economy*, **131**(3), 633–686.

Garicano, Luis. 2000. Hierarchies and the Organization of Knowledge in Production. *Journal of Political Economy*, **108**(5), 874–904.

Gawande, Atul. 2010. *The Checklist Manifesto*. Picador.

Gosnell, Greer K., List, John A., & Metcalfe, Robert D. 2020. The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains. *Journal of Political Economy*, **128**(4), 1195–1233.

Guendelsberger, Emily. 2019. *On the Clock: What Low-wage Work Did to Me and How it Drives America Insane*. Hachette UK.

Haaland, Ingar, Roth, Christopher, & Wohlfart, Johannes. 2023. Designing Information Provision Experiments. *Journal of Economic Literature*, **61**(1), 3–40.

Halac, Marina, & Prat, Andrea. 2016. Managerial Attention and Worker Performance. *American Economic Review*, **106**(10), 3104–32.

Herz, Holger, & Zihlmann, Christian. 2022. *Adverse Effects of Control: Evidence from a Field Experiment*. CESifo Working Paper No. 8890.

Holmstrom, Bengt. 1979. Moral Hazard and Observability. *The Bell Journal of Economics*, 74–91.

Hubbard, Thomas N. 2000. The Demand for Monitoring Technologies: The Case of Trucking. *Quarterly Journal of Economics*, **115**(2), 533–560.

Hubbard, Thomas N. 2003. Information, Decisions, and Productivity: On-Board Computers and Capacity Utilization in Trucking. *American Economic Review*, **93**(4), 1328–1353.

Ichniowski, Casey, Shaw, Kathryn, & Prennushi, Giovanna. 1997. The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines. *AER*, **87**(3), 291–313.

Jackson, C. Kirabo, & Schneider, Henry S. 2015. Checklists and Worker Behavior: A Field Experiment. *American Economic Journal: Applied Economics*, **7**(4), 136–68.

Kelley, Erin M., Lane, Gregory, & Schönholzer, David. 2021. *Monitoring in Small Firms: Experimental Evidence from Kenyan Public Transit.* Mimeo, IIES.

Klotz, Leidy. 2021. *Subtract: The Untapped Science of Less.* Flatiron Books.

Ko, Henry CH, Turner, Tari J, & Finnigan, Monica A. 2011. Systematic Review of Safety Checklists for use by Medical Care Teams in Acute Hospital Settings—Limited Evidence of Effectiveness. *BMC Health Services Research*, **11**(1), 1–9.

Levitt, Steven D., & List, John A. 2011. Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments. *American Economic Journal: Applied Economics*, **3**(1), 224–238.

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

Milgrom, Paul, & Roberts, John. 1990. The Economics of Modern Manufacturing: Technology, Strategy, and Organization. *American Economic Review*, **80**(3), 511–528.

Nagin, Daniel, Rebitzer, James B., Sanders, Seth, & Taylor, Lowell J. 2002. Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment. *American Economic Review*, **92**(4), 850–873.

Ravid, Daniel M., White, Jerod C., Tomczak, David L., Miles, Ahleah F., & Behrend, Tara S. 2023. A meta-analysis of the effects of electronic performance monitoring on work outcomes. *Personnel Psychology*, **76**(1), 5–40.

Rebitzer, James B., & Taylor, Lowell J. 2011. Extrinsic Rewards and Intrinsic Motives: Standard and Behavioral Approaches to Agency and Labor Markets. *Handbook of Labor Economics.*

Roberts, John, & Shaw, Kathryn. 2022. *Managers and the Management of Organizations.* Working Paper 30730. National Bureau of Economic Research.

Tadelis, Steven. 2016. Reputation and Feedback Systems in Online Platform Markets. *Annual Review of Economics*, **8**, 321–340.

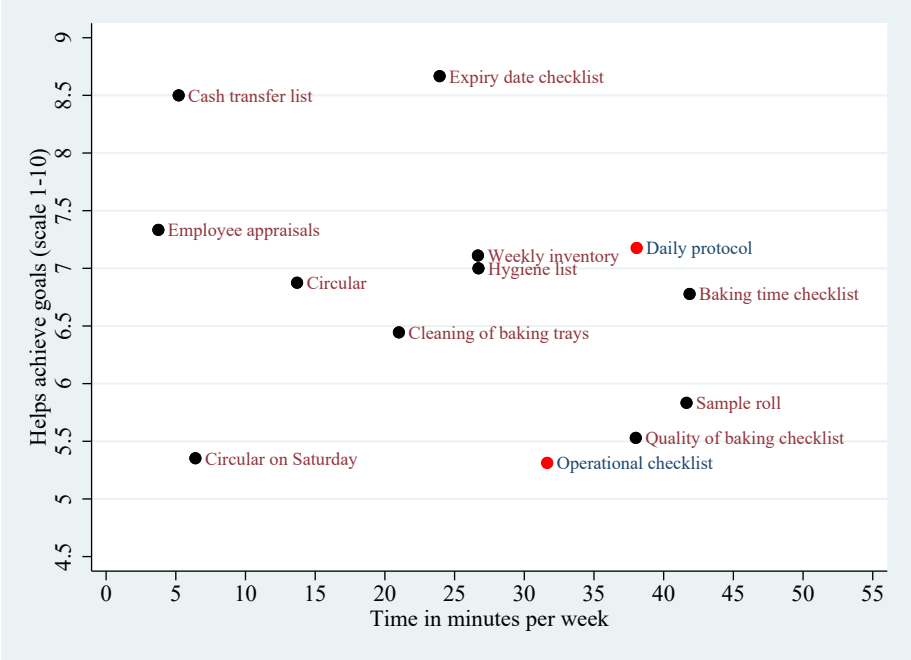
Taylor, Frederick Winslow. 1919. *The Principles of Scientific Management.* Harper & Brothers.

Wager, Stefan, & Athey, Susan. 2018. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, **113**(523), 1228–1242.

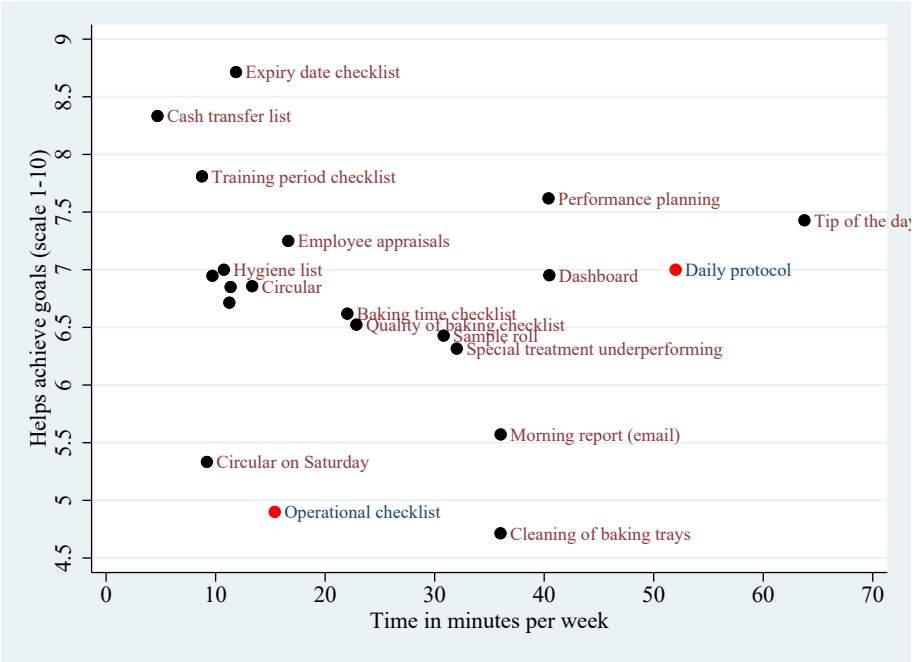
Womack, James P, Jones, Daniel T, & Roos, Daniel. 2007. *The Machine That Changed the World: The Story of Lean Production.* Simon and Schuster.

Zou, Hui, & Hastie, Trevor. 2005. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **67**(2), 301–320.

Figure 1: Variation Across Checklists in Time per Week and Help in Obtaining Goals



(a) Workers



(b) Managers

Notes: This figure uses data from the in-depth, pre-RCT surveys of workers and managers described in Section 2. For each checklist, the figure plots the mean amount of time across workers spent on the checklist, as well as the mean level of agreement with the statement: "The checklist helps (FIRM) to get better and reach company goals." Our pre-RCT surveys ask separately about reaching company goals and avoding mistakes. Results on avoiding mistakes are similar, and are shown in Appendix Figure B4.

Figure 2: Operational Checklist from December 2020 (i.e., the month when the top management decided to conduct the RCT with the research team). Bolding and highlights from the original.

	Mo	Tue	Wed	Thu	Fr	Sat	Sun
1. Covid							
a) Current **covid guidelines** followed! Collecting customer contacts, serving customers: gloves, wearing face mask, keeping distance, airing out the shop	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.
b) Covid hotline: PHONE NUMBERS All questions concerning covid, quarantine, sickness pay	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.
2. Opportunities to increase sales							
a) Spelt products initiative phase 2 Hand over all new spelt flyers to all customers, but do **NOT** put them in the bread bag! **Please destroy old flyers** Recall: Spelt products are: LIST OF 12 DIFFERENT PRODUCTS	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.
b) Bring your own cup initiative correctly implemented? For additional cups contact your regional manager	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.
c) Snack of the month December Cheese-ham-cabbage → Be aware of combined offers	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.
d) Please be mindful of the **appearance of the Berliner doughnut**. In a recent store visited, the sugar was partly scraped off on the side of a Berliner. Carefully touch the Berliners with a cake tong on the side; never touch a Berliner with the cake tong on the top, as sugar might be scraped off; monitor other reasons why sugar is scraped off on Berliners	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.
e) **Roasted almonds correctly placed** Loosely placed on a baking tray in the cake counter, on top of 2-4 packed, not yet closed bags of almonds	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.
f) **Christmas cookies** **Sufficient amount of the mini spelt almond cookies?** **→ If you do a free sample, put 4 mini spelt almond cookies in a 1 kg bag and hand it to the customers!** Sufficient amount of Christmas bags 4 kg Sufficient amount of all Christmas cookies? Follow order processes! Product assortment: - Cookie basket on top pf the counter: All types of almond cookies, coconut cookies, shortbread cookies (5 types) - Edge of the cake counter: Tree cake, gingerbread, Christmas cake - In the counter: alternating between puff cookies and shortbread cookies	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.
g) Product trial Blueberry-pudding snack in LIST OF SHOPS	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.
3. Organizational implementation tasks							
a) **New bonus system for wasted & returned goods** since Dec 1st Make sure to check every day If you have questions, contact your regional manager	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.
b) **Coffee bags** When making and selling coffee, please first empty **old** coffee bags before opening new ones	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.	Sign.

Notes: This figure provides an example of the operational checklist. It is translated from German and has firm-identifying information redacted.

Figure 3: Daily Protocol from December 2020. Bolding and highlights from the original.

Date: _____ Store: _____

	Cash register number	Cash ACTUAL	Cash TARGET	Difference	Sign.	Safe bag	
						Banknote	Coins
1.		€	€	€			
2.		€	€	€			
3.		€	€	€			
4.		€	€	€			
5.		€	€	€			
6.		€	€	€			
7.		€	€	€			
8.		€	€	€			
9.		€	€	€			
10.		€	€	€			
11.		€	€	€			
12.		€	€	€			

Sales (€)		Working hours		Performance	

Special orders "sold out" → should we order more?

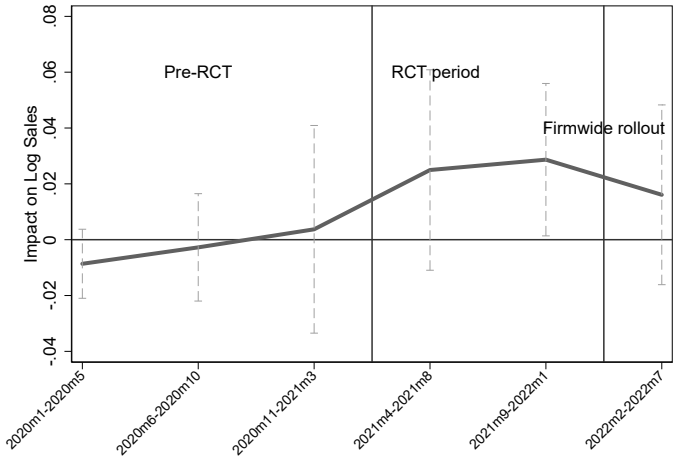
Facility or IT problems, etc.

Shift changes

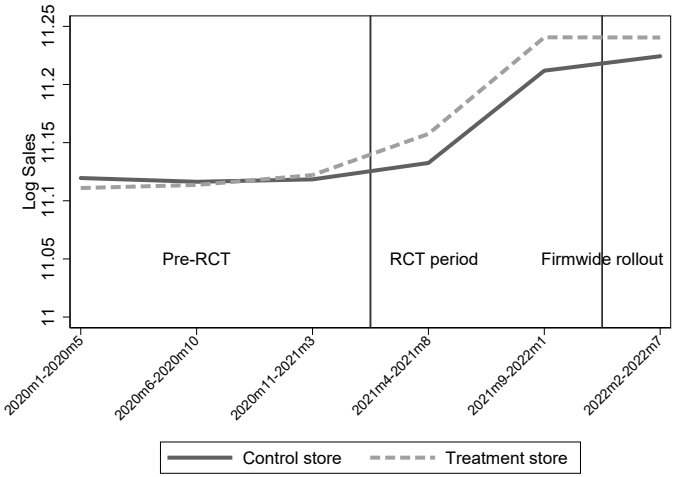
Additional information for the next shift

Notes: This figure provides an example of the operational checklist. It is translated from German and has firm-identifying information redacted.

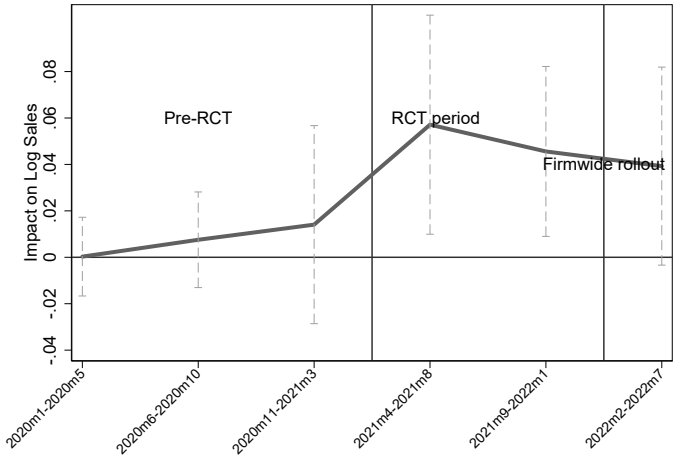
Figure 4: Differences Between Treatment and Control Stores Over Time in Sales



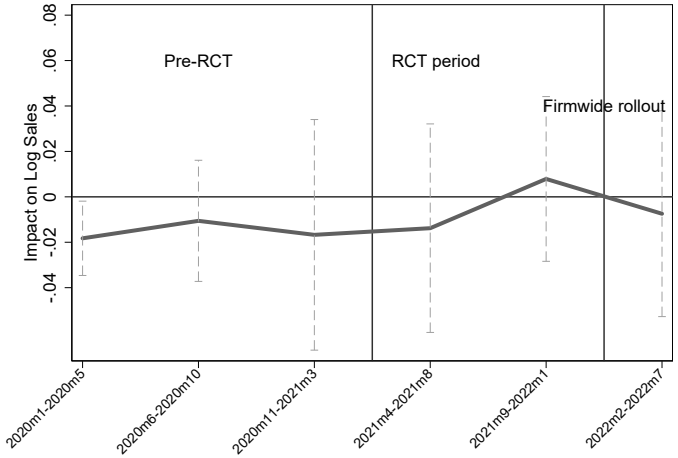
(a) All Stores



(b) All Stores, Two Lines



(c) Stores Where RCT Predicted to Work by Reg. Mgrs.



(d) Stores Where RCT Not Predicted to Work by Reg. Mgrs.

Notes: Panel (a) is similar to that in column 1 of Panel A of Table 2, but we split separately by period of the RCT. The first period of the RCT is April-August 2021 and the second period is September 2021-January 2022. We also show three periods before the RCT, as well as the post-RCT rollout, where we have data for six months. Likewise, panels (c) and (d) here are similar to column 1 in panels (b) and (c) of Table 4. Panel (b) compares treatment versus control stores with two separate lines. The control line plots the control store means, whereas the treatment store plots control means plus the treatment effect in each period.

Table 1: Comparing Pre-Treatment Store Means across the Treatment Groups ($N = 145$ stores): Randomization Check

	Log Sales	Log Busy Sales	Log Slow Sales	Log Customers	Shrinkage	Mystery Shopping Score	Head count	Store League Ranking
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-0.02	-0.01	-0.02	-0.00	-0.00	-0.05	0.53	-4.45
	(0.05)	(0.05)	(0.07)	(0.05)	(0.03)	(0.08)	(0.77)	(7.22)
Constant	11.16***	10.83***	9.86***	9.85***	-2.06***	18.98***	13.27***	77.93***
	(0.03)	(0.03)	(0.05)	(0.03)	(0.02)	(0.06)	(0.50)	(4.82)
p-val	0.73	0.76	0.72	0.97	0.90	0.52	0.49	0.54

Notes: This table compares pre-RCT store-level characteristics across treatment and control stores. Robust standard errors in parentheses * significant at 10%; ** significant at 5%; *** significant at 1%

Table 2: Impacts of the Treatment on Store Outcomes and Employee Attrition (x100)

Panel A: Store Outcomes	(1)	(2)	(3)	(4)	(5)	(6)
Dep. var.:	Log Sales	Log Busy Sales	Log Slow Sales	Log Customers	Log Shrink -age	Mystery Shopping Score (normed)
Treatment	0.027*	0.026*	0.034*	0.023	0.002	0.003
	(0.015)	(0.014)	(0.019)	(0.015)	(0.016)	(0.070)
Observations	1,431	1,431	1,431	1,431	1,431	1,161
Mean dep. var. if Treat=0	11.17	10.86	9.838	9.762	-2.099	-0.0284
Stores	145	145	145	145	145	144

Panel B: Worker Turnover	(1)	(2)	(3)	(4)	(5)
Sample of workers:	All	Untrained Workers	Trained workers	Trained Non-Mgrs	Trained Managers
Treatment	0.07	0.66	-0.44*	-0.23	-1.09*
	(0.24)	(0.40)	(0.25)	(0.27)	(0.61)
Observations	13,271	6,489	6,782	5,403	1,379
Mean dep. var. if Treat=0	2.038	2.806	1.254	1.159	1.647
Workers	1637	863	774	624	150
2-sided p-val: trained v. untrained			0.02		
2-sided p-val: manager v. non-mgrs					0.10

Notes: Standard errors clustered by store are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Panel A Notes: An observation is a store-month during the RCT. Busy sales are sales between 7am and 2pm, whereas slow sales are sales after 2pm. Shrinkage is the share of product lost as a share of total sales revenue. Each regression controls for the mean of the dependent variable in the pre-RCT period, year-month dummies, and several pre-RCT store characteristics (above/below median sales, above/below median head count, above/below median store league performance ranking, and region).

Panel B Notes: An observation is a worker-month during the RCT. Coefficients are multiplied by 100 for ease of exposition. All regressions control for the same controls as in Panel A, as well as a quadratic in worker tenure and worker gender. Since an observation is a worker instead of a store, we control for the store-level mean attrition rate in the pre-RCT period. The "trained v. untrained" p-value is a two-sided p-value from a test of whether the treatment effect is larger for trained workers compared to untrained workers. Likewise, the "manager v. non-mgrs" p-value is a two-sided p-value from a test of whether the treatment effect is larger for managers relative to non-managers. These tests are from a regression where worker skill (i.e., being a trained worker or a manager) is fully interacted with all regressors.

Table 3: Impacts of the Treatment on Worker Survey Outcomes and Customer Reviews

Panel A: Worker Survey	(1)	(2)	(3)	(4)
Dep. var.: (all normed)	Commitment to one's store	Trust bwn. HQ & workers	Last new hire was well-trained	Basic quality control
	(1)	(2)	(3)	(4)
Treatment	0.214**	0.268*	0.005	-0.076
	(0.098)	(0.138)	(0.123)	(0.105)
Observations	390	394	368	394

Panel B: Google Review Scores	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dep. var.:	Average rating	Share 1s	Share 2s	Share 3s	Share 4s	Share 5s	Number of ratings
Treatment	0.204**	-0.033	-0.011	-0.004	-0.032	0.078**	4.029
	(0.091)	(0.022)	(0.012)	(0.012)	(0.026)	(0.034)	(4.151)
Stores	140	140	140	140	140	140	140
Mean DV if Treat=0	4.059	0.107	0.0380	0.0730	0.252	0.530	28.55

Panel C: Google Revs, Text Analysis	(1)	(2)	(3)	(4)	(5)	(6)
Dep. var.: Whether there is a positive comment regarding:	The product	Service	Shop appearance	Speed of service	Value for money	Product availability (placebo)
Treatment	0.054**	-0.002	0.007	0.008**	0.006	-0.006
	(0.026)	(0.024)	(0.007)	(0.004)	(0.006)	(0.012)
Stores	140	140	140	140	140	140
Mean DV if Treat=0	0.265	0.204	0.0175	0.00531	0.0157	0.0575

Notes: Standard errors clustered by store are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Panel A Notes: An observation is a worker. Surveyed workers are anonymous so we cannot control for pre-RCT store characteristics.

Panels B and C Notes: An observation is a store during the RCT. We control for the mean of the dependent variable in the pre-RCT period, plus the pre-RCT store characteristics listed in Table 2. There are a few stores for which Google reviews are not available both during and before the RCT.

Table 4: Treatment Effects on Store Outcomes are Sizable in Stores where Regional Managers Predict the Treatment Will Work, but Negligible in Stores where the Treatment is Not Predicted to Work

Dep. var.:	Log Sales	Log Busy Sales	Log Slow Sales	Log Customers	Log Shrink -age	Mystery Shopping Score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Stores Where RCT Predicted to Work by Regional Managers						
Treatment	0.052**	0.050**	0.058***	0.048**	-0.024	0.080
	(0.020)	(0.019)	(0.022)	(0.019)	(0.021)	(0.089)
Observations	744	744	744	744	744	597
Mean dep. var. if Treat=0	11.09	10.77	9.761	9.684	-2.063	0.0410
Stores	76	76	76	76	76	75
Panel B: Stores Where RCT Predicted Not to Work by Regional Managers						
Treatment	-0.003	-0.003	0.004	-0.006	0.024	-0.068
	(0.020)	(0.019)	(0.027)	(0.020)	(0.020)	(0.109)
Observations	687	687	687	687	687	564
Mean dep. var. if Treat=0	11.27	10.96	9.929	9.852	-2.142	-0.108
Stores	69	69	69	69	69	69
Panel C: Comparing Treatment Effects by Regional Manager Predictions						
2-sided p-val: Panels A v. B	0.05	0.05	0.12	0.05	0.10	0.29
1-sided p-val: Panels A v. B	0.02	0.03	0.06	0.03	0.05	0.15

Notes: Each panel here is similar to Panel A of Table 2. The difference is that we split the sample based on whether or not a regional manager predicted the treatment would work in each store. The p-values in Panel C are from tests of how treatment effects vary based on regional manager prediction. These p-values are based on regressions where regional manager predictions are fully interacted with all regressors. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 5: Heterogeneity by Regional Manager Predictions: Impacts of Treatment on Employee Attrition (x100)

Workers:	All	Untrained Workers	Trained workers	Trained Non-Mgrs	Trained Managers
Panel A: Stores Where RCT Predicted to Work					
Treatment	-0.46	0.22	-1.08***	-0.70*	-2.23**
	(0.30)	(0.51)	(0.37)	(0.41)	(0.85)
Mean dep. var. if Treat=0	2.056	2.667	1.505	1.306	2.228
Observations	6,595	3,126	3,469	2,691	778
Workers	829	422	407	320	87
2-sided p-val: trained v. untrained			0.05		
2-sided p-val: manager v. non-mgrs					0.06
Panel B: Stores Where RCT Not Predicted to Work					
Treatment	0.47	1.03	0.09	0.12	0.55
	(0.40)	(0.63)	(0.37)	(0.37)	(0.87)
Mean dep. var. if Treat=0	2.019	2.931	0.967	1	0.806
Observations	6,676	3,363	3,313	2,712	601
Workers	878	483	395	328	67
Panel C: Comparing Treatment Effects by Regional Manager Predictions					
2-sided p-val: Panels A v. B	0.07	0.32	0.03	0.14	0.02
1-sided p-val: Panels A v. B	0.03	0.16	0.01	0.07	0.01

Notes: Each panel here is similar to Panel B of Table 2. The difference is that we split the sample based on whether or not a regional manager predicted the treatment would work in each store. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 6: Responses from the Regional Manager Survey: Explanations for Why the Treatment Will Work

Explanation	Share
A Utility Explanation, Such as People Like Not Having Checklists or Feeling Less Stressed About Checklists	37%
A Problem Explanation Such as Not Experiencing Problems or Team Having Good Communication or No Bureaucracy Needed Because People Know Procedures	71%
Regional Managers Will Invest More Time in a Store if it is Treated (e.g., Visiting or Calling More Frequently)	0%
Treatment Stores are Likely to Experience Outside Shocks to Performance During the RCT	0%

Notes: These data are from the pre-RCT regional manager prediction survey. The numbers are based on examining the free text responses of regional managers. We restrict to the 78 stores where regional managers predict that the treatment will work. For 21 of the stores, the regional manager made a prediction, but did not provide a clear explanation (e.g., the regional manager just said "Yes, will work") and the percentages are based on the 57 stores where regional managers provided explanations. Of the 21 stores with no explanations, 14 of those cases come from 2 regional managers, one of whom was picking up their kids during the survey and the other one had just arrived at an appointment. These two regional managers gave no explanation for all of their predictions, though still made yes/no predictions for all stores, and appeared to take these predictions very seriously. Given the short time window between informing regional managers about the RCT and performing the randomization, there was only of couple weeks to conduct the regional manager surveys, so it was not possible to re-schedule. The text of the explanations, translated into English, appear in Appendix Tables B7 and B8.

Table 7: Regional Manager Predictions are Correlated with Some Pre-RCT Store Characteristics, but the Predictive Power is Relatively Low

	(1) Base	(2) Lasso-selected regressors
Treatment store	-0.025	
	(0.080)	
Pre-RCT Log Sales	0.015	
	(0.237)	
Pre-RCT mystery shopping score	0.286***	0.282***
	(0.094)	(0.072)
Pre-RCT mean head count	-0.026*	-0.025***
	(0.015)	(0.007)
Pre-RCT store league performance ranking	0.000	
	(0.001)	
Pre-RCT mean tenure of workers	-0.000	
	(0.001)	
Observations	144	144
R-squared	0.166	0.165

Notes: An obs. is a store. Robust SEs in parentheses. Column 2 uses the regressors selected by lasso, where λ is selected by cross-validation. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 8: Impact of the Treatment on Log Sales: Heterogeneity by Team Size

	(1)	(2)	(3)	(4)
Sample	Small teams	Big teams	All	All
Treatment	0.079**	0.010	0.079**	0.051
	(0.034)	(0.016)	(0.037)	(0.047)
Big team at firm			0.044	0.041
			(0.030)	(0.030)
Treatment X Big team			-0.069*	-0.063
			(0.041)	(0.043)
Treatment X Predict success				0.049
				(0.031)
Predict that treatment will work				0.005
				(0.018)
Observations	305	984	1,289	1,289
Mean DV if Treat=0	10.82	11.27	11.17	11.17
Stores	35	110	145	145

Notes: An observation is a store-month during the RCT. Standard errors clustered at the store level are in parentheses. A big team is defined as having a store head count above 10. Controls are the same as in Table 2. * significant at 10%; ** significant at 5%; *** significant at 1%

Web Appendix, "Is This Really Kneaded? Identifying and Eliminating Potentially Harmful Forms of Workplace Control", by Friebel, Heinz, Hoffman, Kretschmer, and Zubanov

Appendix A provides additional discussion on various topics. For each subsection, we give the relevant section of the main paper that it accompanies. Appendix B contains additional figures and tables. Appendix C provides materials used by the firm in the RCT.

Appendix A Additional Discussion and Information

A.1 Procedure for Identifying All Checklists (Section 2)

We asked the top management to present all checklists. In the meeting, the sales director, who was part of the project team, presented step by step all checklists from the stores. He forgot one checklists—the head of the workers' council informed him about it at the end. No one in the meeting from the project team was aware of any documentation duties that were missing. In a second step, we asked the store managers and workers at the end of the in-dept interviews whether any checklists were missing on our lit. It turned out that no checklists were missing.

A.2 Did Treatment Stores Continue to Use Informal Versions of the Removed Checklists? (Section 2)

Did store managers continue to use any sort of informal version of the checklists during the RCT despite being in the treatment group? We investigated this using a few questions in our survey of store managers at the end of the RCT. For the operational checklist, there was no effort by store managers to replace it. In a survey, not a single store manager reported continuing to use an informal version of the checklists. For the daily protocol, some managers reported making abbreviated informal notes about information that was in the daily protocol to communicate across shifts. For example, some managers sent notes saying they forgot to clean the oven or that someone was sick.

We draw two main conclusions from these surveys on implementation. First, with very few exceptions, treatment store managers did not replace the two checklists with any informal checklist instead. Second, the limited replacement that was done was for the information in the daily protocol concerning information between shifts, consistent with there being value for the daily protocol for some store managers, but not for the operational checklist.

A.3 Store League Performance Ranking (Section 2)

The store league performance ranking is broadly inspired by the Bundesliga (Germany's professional soccer league) and is decided upon by the firm's top management. Top management reviews a large number of performance indicators for each store during the last few weeks (for example, sales, shopping, waste, targets) and creates a subjective assessment for

each store. There is no formula that is used to calculate the ranking. The rankings are not used for assignments of bonuses, but are instead used to mark stores needing improvement.

A.4 Randomization Procedure and Controlling for Stratification Variables in the Empirical Analysis (Sections 2-3)

As described in Section 2, we perform a stratified randomization using region, pre-RCT sales, pre-RCT head count, and pre-RCT store league performance ranking. This was for several reasons. First, Bruhn & McKenzie (2009) advocate for stratifying based on geography and baseline outcomes, leading us to include region and pre-RCT sales. Second, analysis of variance suggested that region and pre-RCT head count were strong predictors of pre-RCT sales. Third, our institutional knowledge that it would be useful to also consider store league performance ranking in the stratified randomization, as it is a variable of interest to some firm managers.

In our empirical analysis, we control for the variables used in stratification in above/below median form. We found that this slightly improves power relative to above/below mean, but results are very similar in both cases. We also obtain similar results controlling for the stratification variables in continuous form.

A.5 Construction of the Employee-Month Panel (Section 2)

Our worker-month panel is based on regular workers at the firms. Additionally, in Germany, there are "minijobbers" who are allowed to work up to 12 hours per week, and for whom the firm doesn't pay employment taxes (Tazhitdinova, 2022). We exclude minijobbers from our sample when analyzing workers. We do this because minijobbers are very different from the other workers in our firm, who work around 30 hours per week, whereas minijobbers only work 7-8 hours per week on average. Minijobbers are supposed to be there on a temporary basis and are expected to attrite. All our results are similar when including minijobbers. This is unsurprising given that minijobbers comprise only about 8% of hours worked during the RCT. If minijobbers are instead included in the sample, all our conclusions are unchanged, with no effect of the treatment on overall employee attrition, but with a negative effect on skilled worker attrition. The treatment has no significant impact on minijobbers' employee attrition, consistent with the idea that their eventual attrition is expected.

A.6 Framing to Workers Explaining Checklist Removal (Sec. 2)

In Section 2, we discuss how the framing of the treatment was not neutral. In particular, in telling treatment store workers that the checklists would be removed, it was emphasized to workers that the firm trusts it workers, and that extra time freed up can be spent on customers and colleagues.

As we discuss in Section 2, it would have been highly artificial for us to have implemented our treatment with a neutral framing, so we did not. Still, it worth reflecting on the implications of framing for the interpretation of our results.

We acknowledge that part of the effects we estimate could be due to framing. However, we believe it is highly unlikely that a pure framing effect could lead to our RCT's quite

sizable effects on sales and attrition that persist for 10 months. Prior work on framing in the field tends to estimate moderate effects that are fairly context-specific.[1] We view the framing of the RCT as complementary to the potential signaling of removing monitoring, i.e., the framing helps people understand the signaling. We also wish to point out that from a managerial standpoint, it is less policy-relevant to use neutral standpoint. To make policy changes comprehensive to workers, companies want to use positive framings, so using a positive framing is natural.

In the experimental economics literature, there is a debate about the importance of framing in relation to results on the costs of control. Schnedler & Vadovic (2011) and Hagemann (2007) provide evidence that the negative impact of control on effort (Falk & Kosfeld, 2006) depends on framing. In particular, they show that a negative framing of control induces negative responses, whereas a neutral framing has a limited effect.

A.7 Survey of Worker Attitudes

Panel A of Table 3 shows results on commitment to the store. We also asked about commitment to the firm, and there was no impact of the treatment on commitment to the firm. In lower-skill retail jobs, we believe it is natural for workers to have their greatest commitment to the store and their work-team instead of the firm overall. When interviewed, store managers emphasize that "we" pertains to the store instead of the overall company.

A.8 Use of Data on Google Reviews (Section 3)

General issues with data on online reviews. As discussed in Tadelis (2016), there are various issues with using data on online reviews. Reviews are left-skewed—we address this by showing effects on the whole distribution of point responses, and indeed, our effects occur due an increase in the number of 5's. Another concern with online reviews is that some reviews are fake. However, there is no reason why our treatment would affect whether a store reviews fake reviews. Further, our firm has a traditional management culture and would be unlikely to make fake reviews. Since many customers do not leave reviews, a concern is whether the treatment affects selection into receiving a review. As seen in column 7 of Panel B of Table 3, there is no statistically significant impact of the treatment on the number of Google reviews that a store receives. While statistically indistinguishable from zero, the coefficient is +4, meaning that treatment stores receive four more ratings. Research on online reviews suggests that not leaving a review is more likely to be a sign of a negative experience (Dellarocas & Wood, 2008; Tadelis, 2016), suggesting that the "true effect" of the treatment on positive customer experiences would be even higher if customers left reviews at the same rate across treatment and control stores.

Identifying dates. The data on Google Reviews were gathered in November 2022. To confidently identify reviews left during the RCT, we focus on reviews marked as posted one year ago.

[1] Hossain & List (2012) find that whether an incentive is gain- or loss-framed matters for team, but not individual performance. However, De Quidt *et al.* (2017) find no effect of framing on performance.

Identifying qualitative characteristics. During the RCT period, 6163 reviewers posted 7009 customer reviews, of which 3068 had text comments in addition to the star rating. Having read a selection of the comments, we identified the following specific topics that were most frequently mentioned, positively or negatively: product characteristics (taste, look, smell), service quality (were sales personnel friendly and helpful?), store appearance (looks, ambience, hygiene level), product availability, speed of service, and value for money. We then instructed a research assistant to read through all the 3068 reviews with text and indicate which of the above topics was mentioned in the text of each review. For example, the review "Gute Qualität bei den Waren. Freundliche Bedingung. Alles in allem zu Empfehlen!" ("Good quality goods. Friendly service. All in all recommended!") positively mentions product and service quality, and so the RA indicated this review as mentioning these two topics. Another review, "Personal ist leider teilweise unfreundlich, zumindest die älteren Mitarbeiter... Ware ist sehr oft leer, egal zu welcher Uhrzeit. Die Backwaren an sich sind sehr lecker" / "Unfortunately, some of the staff are unfriendly, at least the older employees... Shelves are very often empty, no matter what time of day. The baked goods themselves are very tasty", positively mentions product but negatively mentions service quality and product availability, which topics were accordingly indicated by the RA. We validated the RA's choices by reading a selection of reviews themselves, finding a near 100% correspondence.

85% of the text reviews mentioned one or several of the above topics, but the remaining 15% were too general or hard to classify. The majority of those reviews contained an overall positive assessment, for example, "Gut" / "Good", "OK", "Wie immer alles bestens"/ "As always, everything is fine", "Top Bäckerei" / "Top bakery" or simply positive emojis. A few were hard to classify (e.g., "Berliner Jungs kommen gar nicht aus Berlin." / "Berlin boys don't even come from Berlin.", "Früher war sehr gut aber durch viele ausländische Mitbürger ist ja bisschen runter gekommen" / "It used to be very good, but the many foreign citizens have brought it down a bit"). We excluded these reviews from the text-based analysis (Panel C of Table 3) but kept them in the analysis of review scores (Panel B of Table 3).

A.9 Profit Calculation Details (Section 5.1)

The project team had two days of half a day meetings in the nine-person full group. Assigning each person a day rate of €1500, the total cost of these meetings is €13.5k. There were also four meetings of half a day in the small group. We use 7 half-person days a rate of €1200 (reflecting the people at the small group meetings are less seniors), and obtain a cost of €4.2k. There were around 5 person-days for data transmission at €800 per day, yielding a cost of €4k. There were also costs of training the 15 regional managers, plus two top directors, for which the cost was about one work day, or €1000. Furthermore, we used RA time of about €9k. Summing up, we obtain a time cost of roughly €31k.

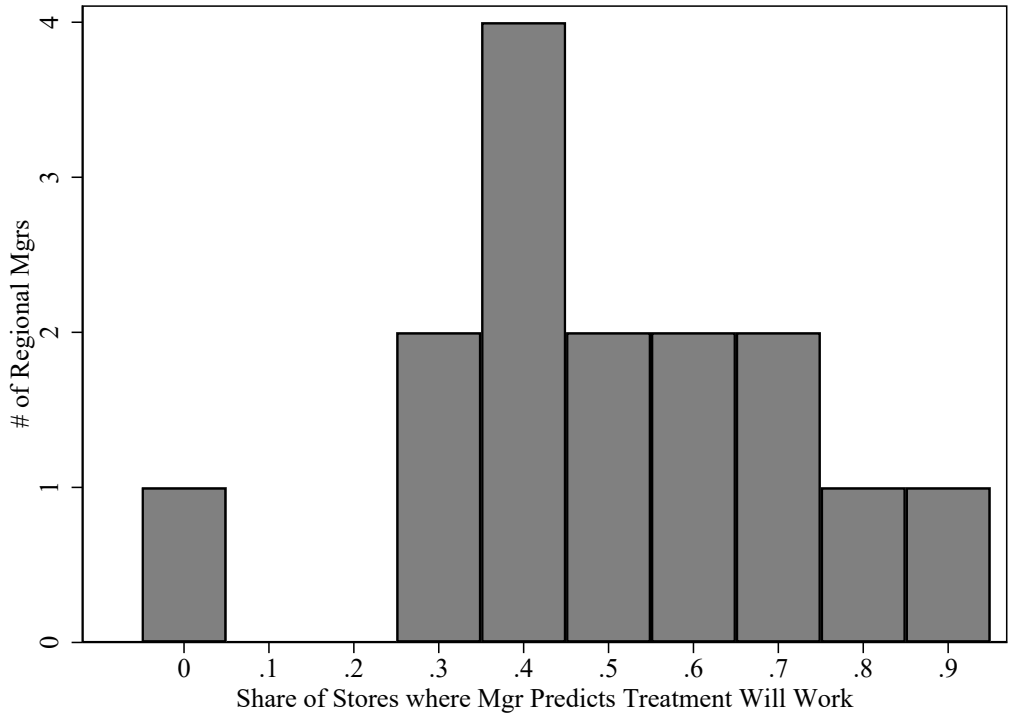
A.10 Mediation Analysis (Section 5.3)

We use a mediation analysis (Imai *et al.* , 2010a,b) to address the question of whether our estimated sales effects are due to lower turnover. We estimated the models in Panel A of Table 2 while adding a control variable for the attrition of trained workers in each store-month. We also ran the results using trained manager attrition. In both cases, the

estimated treatment effects are extremely similar when controlling for a store's monthly attrition rate. We also estimated the models in Table 4 and observe no evidence of mediation when restricting to stores where regional managers predict the treatment will work, or while restricting to stores where regional managers predict the treatment will not work.

Appendix B Appendix Figures and Tables

Figure B1: Variation in Manager-Level Rates of Predicting that the Treatment Will Work

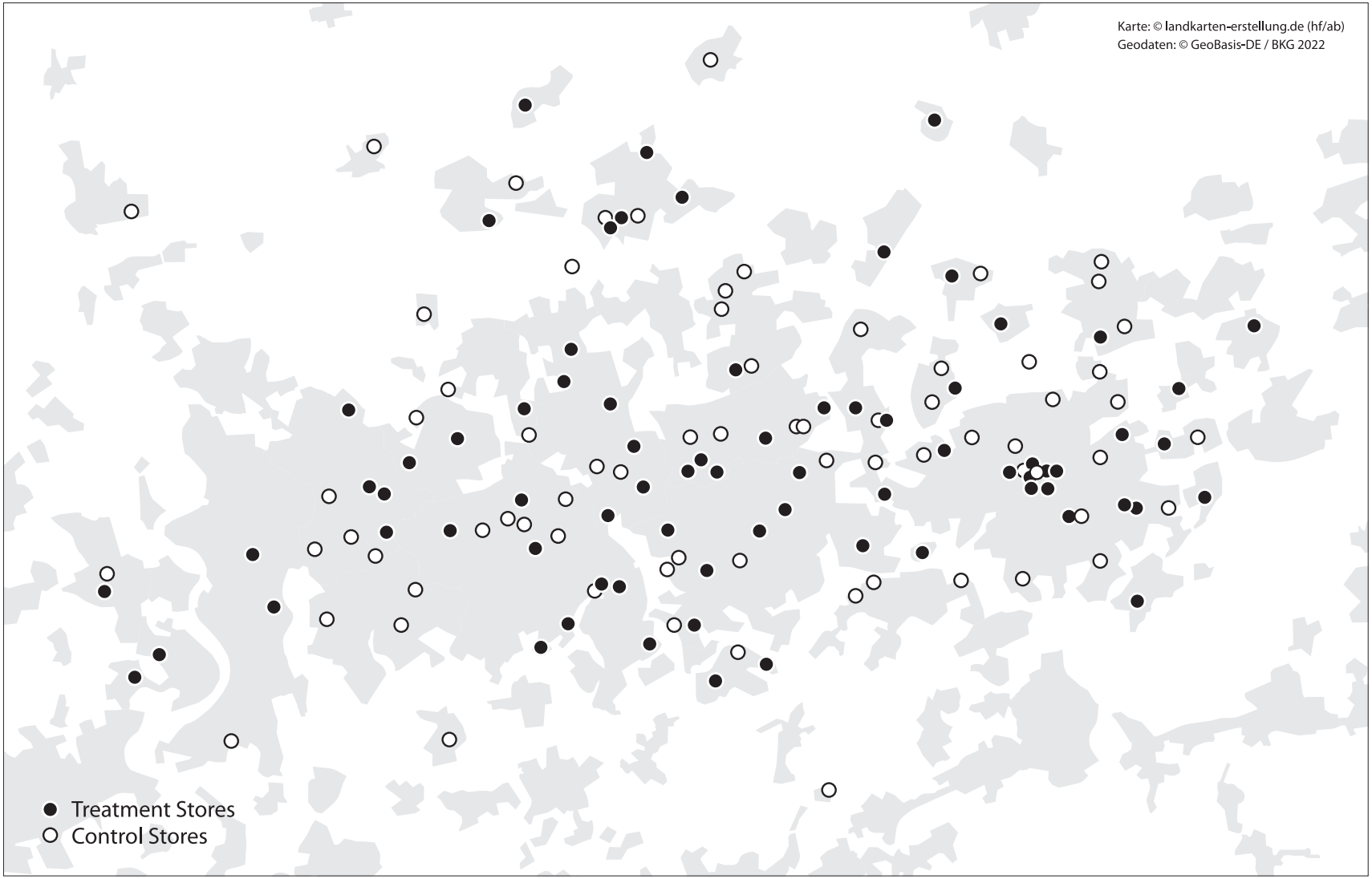


Notes: This figure shows the distribution across managers in rates of predicting that the treatment will work. There are 15 regional managers, who are responsible for roughly 10 stores each. For example, we see that there are 2 regional managers who predict that the treatment will work in between 25-35% of their stores.

Figure B2: Picture of a Sample Bakery

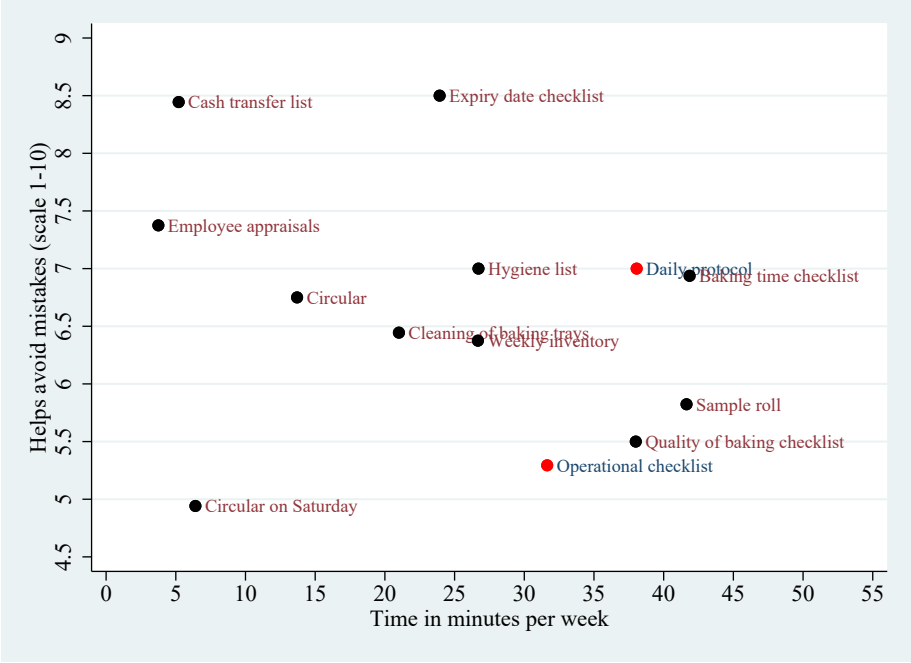
Notes: Identifying information about the firm has been redacted from this picture.

Figure B3: Location of Treatment and Control Stores

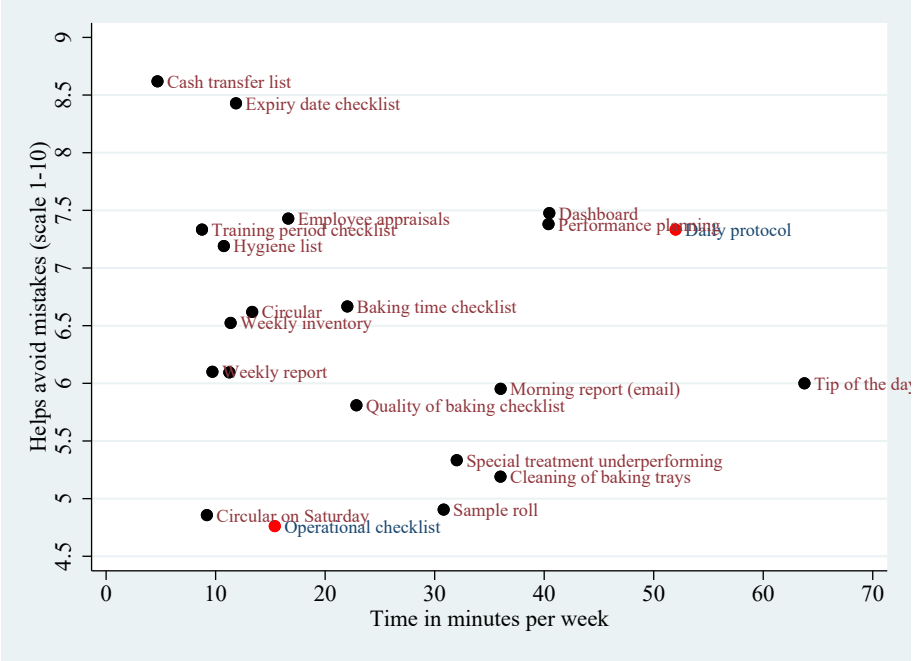


Notes: This figure shows the location of treatment and control stores on a map, with identifying information redacted.

Figure B4: Variation Across Checklists in Time per Week and Help in Avoiding Mistakes



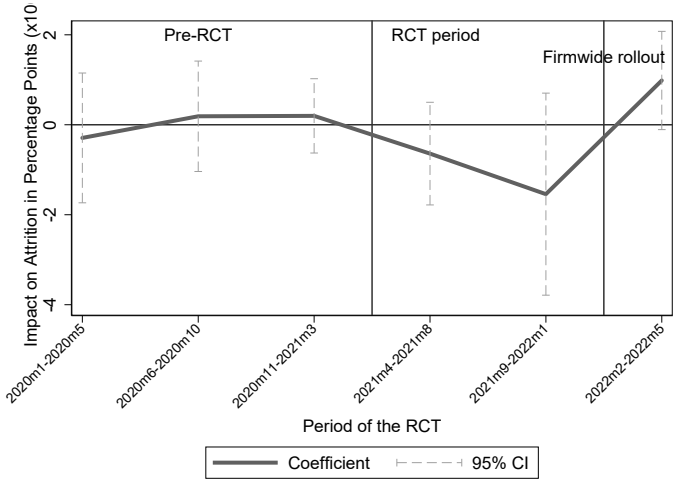
(a) Workers



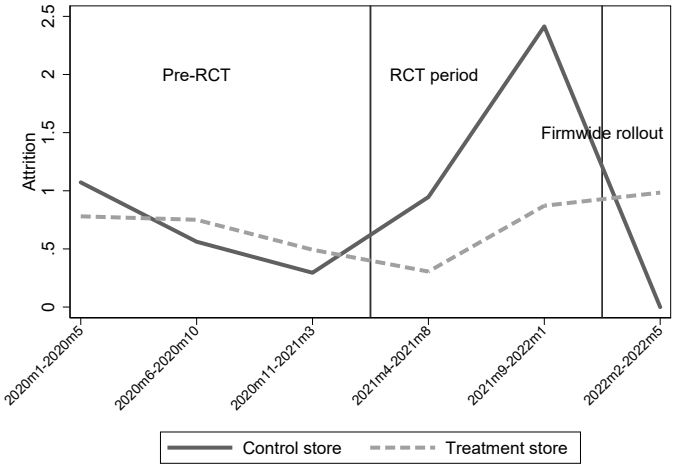
(b) Managers

Notes: Help in avoiding mistakes is measured using: "The checklist helps (FIRM) avoid mistakes."

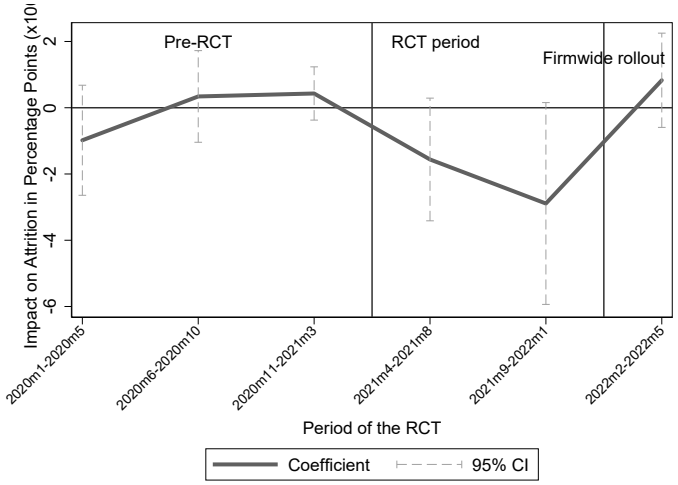
Figure B5: Differences Between Treatment and Control Stores Over Time in Store Manager Attrition



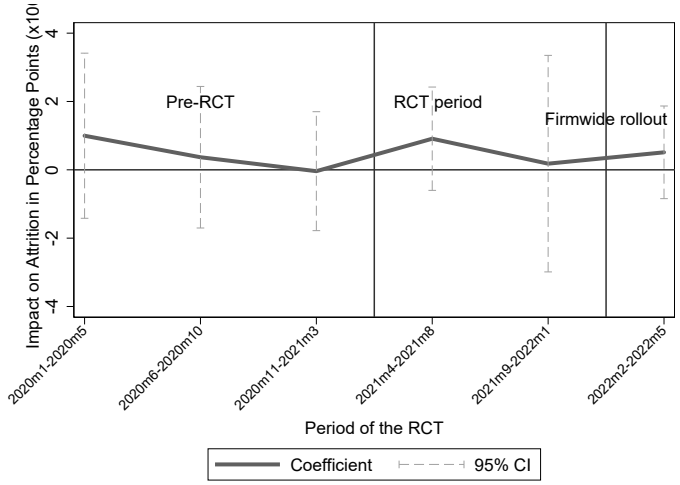
(a) All Stores



(b) All Stores, Two Lines



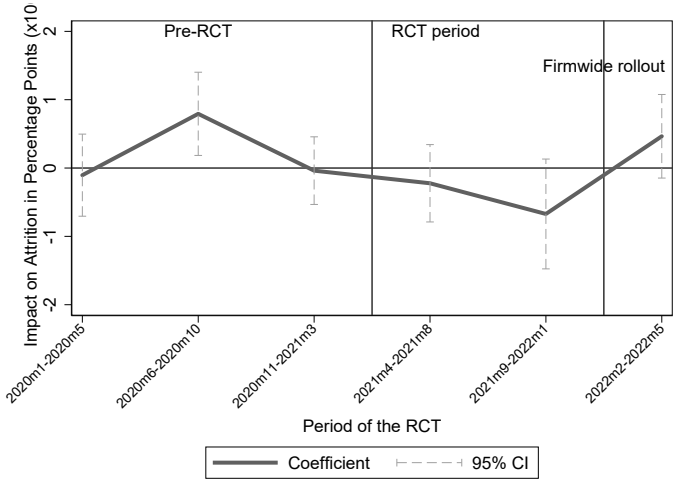
(c) Stores Where RCT Predicted to Work by Reg. Mgrs.



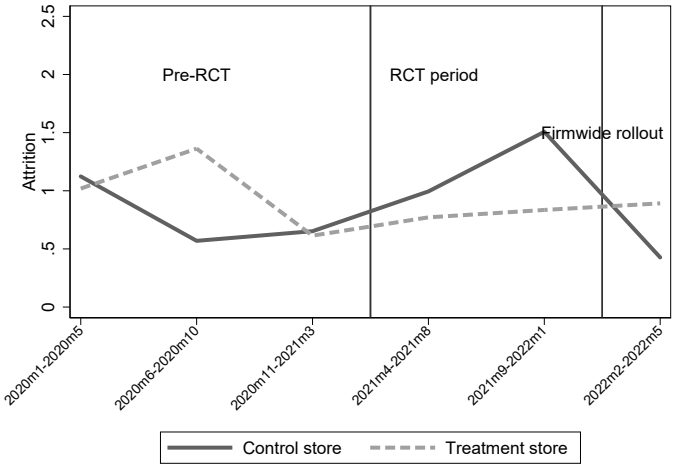
(d) Stores Where RCT Not Predicted to Work by Reg. Mgrs.

Notes: Panel (a) is similar to that in column 5 of Panel B of Table 2, but we split separately by 5-month period of the RCT. Likewise, panels (c) and (d) here are similar to column 5 of Table 5. Panel (b) compares treatment versus control stores with two separate lines. The control line plots the control store means, whereas the treatment store plots control means plus the treatment effect in each period.

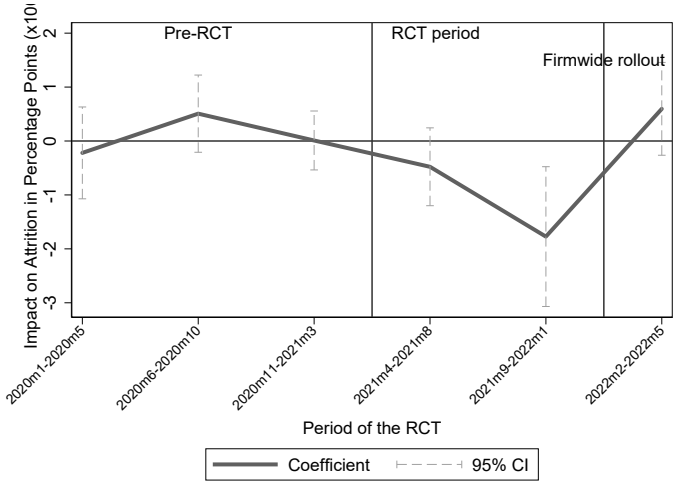
Figure B6: Differences Between Treatment and Control Stores Over Time in Trained Worker Attrition



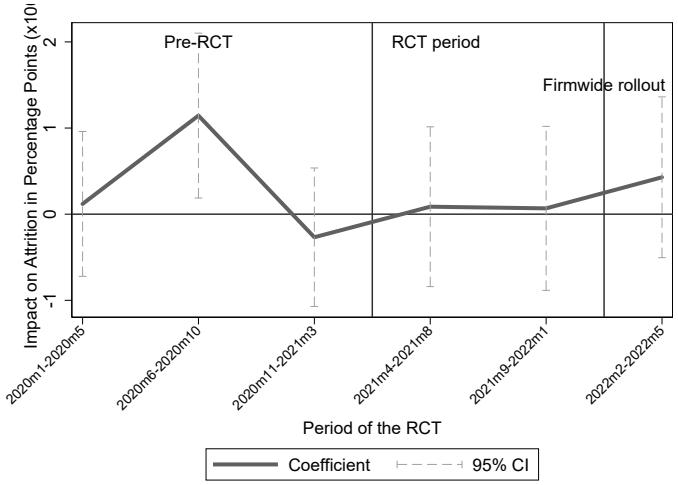
(a) All Stores



(b) All Stores, Two Lines



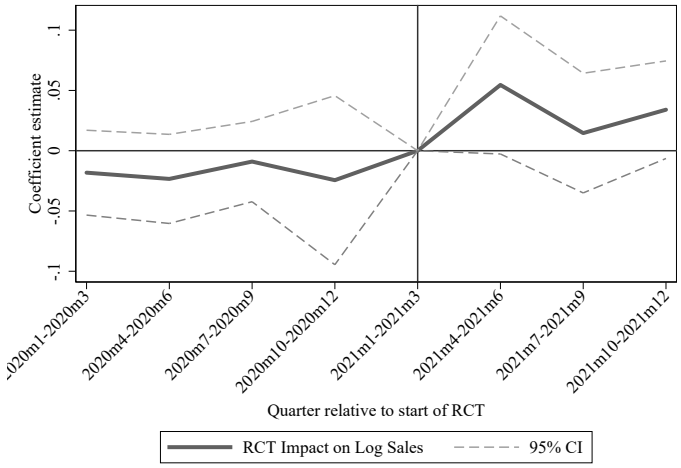
(c) Stores Where RCT Predicted to Work by Reg. Mgrs.



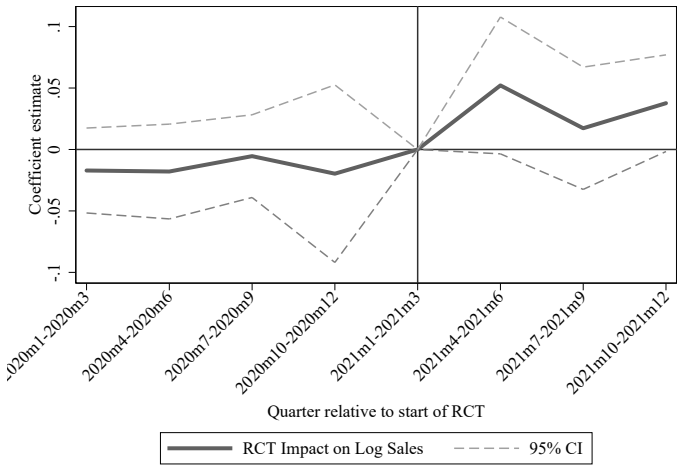
(d) Stores Where RCT Not Predicted to Work by Reg. Mgrs.

Notes: Panel (a) is similar to that in column 3 of Panel B of Table 2, but we split separately by 5-month period of the RCT. Likewise, panels (c) and (d) here are similar to column 3 of Table 5. Panel (b) compares treatment versus control stores with two separate lines. The control line plots the control store means, whereas the treatment store plots control means plus the treatment effect in each period.

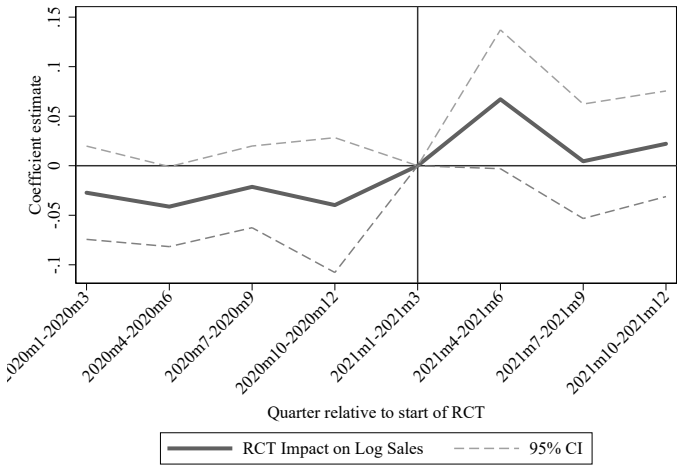
Figure B7: Event Study Impacts of the Treatment: Stores Where Treatment Expected to Have Effect



(a) Log Sales



(b) Log Busy Sales



(c) Log Slow Sales

Notes: This figure shows the event study impacts of checklist removal.

Table B1: Impacts of the Treatment on Individual Components of the Mystery Shopping Score

	Name badge	Sales procedure	Product present-ation	Free sample	Advert-ising	Customer interact-ion	Sales quest-ions	Upsell	Golden roll	Other roll	Store appear-ance
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Panel A: All Stores											
Treatment	0.001	-0.001	0.018	0.000	-0.021	-0.004	0.002	0.014	-0.023	0.003	0.030
	(0.020)	(0.008)	(0.020)	(0.000)	(0.042)	(0.036)	(0.004)	(0.014)	(0.036)	(0.007)	(0.029)
Observations	1,161	1,161	1,161	1,161	1,161	1,161	1,161	1,161	1,161	1,161	1,161
Mean DV if Treat=0	1.890	1.986	2.917	1	1.840	2.037	0.995	0.0254	2.581	0.984	2.679
Stores	144	144	144	144	144	144	144	144	144	144	144
Panel B: Stores Where RCT Predicted to Work by Regional Mgrs											
Treatment	0.050**	-0.007	0.024	0.000	-0.027	-0.062	0.000	0.022	0.081*	-0.002	0.026
	(0.025)	(0.012)	(0.025)	(0.000)	(0.057)	(0.048)	(0.000)	(0.016)	(0.047)	(0.011)	(0.038)
Observations	597	597	597	597	597	597	597	597	597	597	597
Mean DV if Treat=0	1.885	1.990	2.929	1	1.942	2.092	1	0.0153	2.542	0.983	2.642
Stores	75	75	75	75	75	75	75	75	75	75	75
Panel C: Stores Where RCT Predicted Not to Work by Regional Mgrs											
Treatment	-0.044	0.002	0.019	0.000	-0.025	0.065	0.008	-0.002	-0.106*	0.007	0.030
	(0.031)	(0.011)	(0.029)	(0.000)	(0.064)	(0.048)	(0.007)	(0.018)	(0.053)	(0.008)	(0.041)
Observations	564	564	564	564	564	564	564	564	564	564	564
Mean DV if Treat=0	1.896	1.982	2.902	1	1.723	1.975	0.988	0.0371	2.625	0.984	2.721
Stores	69	69	69	69	69	69	69	69	69	69	69

Notes: This table presents analyses similar to those in column 6 of Table 2. The difference is we look at the individual components of the mystery shopping scores. * significant at 10%; ** significant at 5%; *** significant at 1%

Table B2: Comparing Trained vs. Untrained Workers in Mean Characteristics

	Untrained	Trained	Trained Non-mgr	Trained Manager
Female	.95	.99	.99	1
Age	36.68	42.35	41.5	45.66
Base wages in euros	10.44	13.88	13.53	15.24
Tenure in yrs	4.84	11.62	10.82	14.7
Tenure of 1yr or less	.19	.05	.06	0
Tenure of 1-2yrs	.18	.06	.07	.01
Tenure of 2-5yrs	.29	.11	.14	.01
Tenure of 5-10yrs	.16	.26	.25	.29
Tenure more than 10yrs	.19	.52	.48	.68

Notes: This table compares workers of different types using data from March 2021, which is the month before the RCT began.

Table B3: Examining Alternative Explanations for Larger Sales Effects in Stores where Managers Predict Treatment to Work

	(1)	(2)	(3)	(4)	(5)
Treat X RCT Predicted to Work	0.055**	0.053*	0.051*	0.075***	0.047*
	(0.027)	(0.030)	(0.028)	(0.027)	(0.025)
2-sided p-val	0.04	0.08	0.07	0.01	0.07
1-sided p-val	0.02	0.04	0.04	0.00	0.03
Variable to add:					
Treat*(Pre-RCT Log Sales)	X				
Treat*(Pre-RCT mean head count)		X			
Treat*(Pre-RCT mean tenure of workers)			X		
Treat*(Pre-RCT mystery shopping score)				X	
Treat*(Pre-RCT store league performance ranking)					X

Notes: This table accompanies the discussion in Section 4. It displays how key interaction term coefficients vary as we include regressors for an additional characteristic, as well as the interaction of treatment times the characteristic. * significant at 10%; ** significant at 5%; *** significant at 1%

Table B4: Examining Alternative Explanations for Larger Sales Effects in Stores where Managers Predict Treatment to Work

	(1)	(2)	(3)	(4)	(5)	(6)
Treatment X Predict success	0.055**	0.053*	0.051*	0.075***	0.047*	0.051**
	(0.027)	(0.030)	(0.028)	(0.027)	(0.025)	(0.024)
Treat*(Pre-RCT Log Sales)	-0.051					-0.122
	(0.052)					(0.092)
Treat*(Pre-RCT mean head count)		-0.002				0.003
		(0.004)				(0.005)
Treat*(Pre-RCT mean tenure of workers)			-0.001			-0.002
			(0.005)			(0.005)
Treat*(Pre-RCT mystery shopping score)				-0.087**		-0.128***
				(0.037)		(0.046)
Treat*(Pre-RCT store league performance ranking)					-0.0001	-0.001*
					(0.0003)	(0.000)
2-sided p-val, Treat X Predict success	0.04	0.08	0.07	0.01	0.07	0.04
1-sided p-val, Treat X Predict success	0.02	0.04	0.04	0.00	0.03	0.02

Notes: This table accompanies the discussion in Section 4. It displays how key interaction term coefficients vary as we include regressors for an additional characteristic, as well as the interaction of treatment times the characteristic. * significant at 10%; ** significant at 5%; *** significant at 1%

Table B5: Heterogeneity in Sales Effects Based on Time Spent on the Daily Protocol

	(1)
Treatment	0.032
	(0.031)
Treatment X Time spent on daily protocol	-0.000
	(0.001)
Time spent by store on daily protocol, overall	0.000
	(0.001)
Observations	1,221

Notes: An observation is a store-month during the RCT. Standard errors clustered at the store level are in parentheses. Each regression controls for the mean of the dependent variable in the pre-period and year-month fixed effects. A big team means more than 10 workers at the store.* significant at 10%; ** significant at 5%; *** significant at 1%

Table B6: Differences Between Treatment and Control Stores During the Post-RCT Firmwide Rollout

Panel A: Store Outcomes	(1)	(2)	(3)	(4)	(5)	(6)
Dep. var.:	Log Sales	Log Busy Sales	Log Slow Sales	Log Customers	Shrink -age	Mystery Shopping Score
Treatment	0.016	0.021	0.008	0.015	0.015	-0.031
	(0.016)	(0.015)	(0.018)	(0.016)	(0.019)	(0.113)
Observations	852	426	852	852	852	533
Mean DV if Treat=0	11.22	10.92	10.58	9.753	-2.098	18.44
Stores	142	142	142	142	142	141

Panel B: Worker Turnover	(1)	(2)	(3)	(4)	(5)
Sample of workers:	All	Untrained Workers	Trained workers	Trained Non-Mgrs	Trained Managers
Treatment	0.07	-0.30	0.46	0.31	0.98*
	(0.33)	(0.59)	(0.31)	(0.37)	(0.56)
Observations	5,095	2,530	2,565	2,066	499
Mean DV if Treat=0	1.647	2.765	0.430	0.525	0
Workers	1365	692	673	544	129

Panel C: Google Review Scores	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dep. var.:	Average rating	Share 1s	Share 2s	Share 3s	Share 4s	Share 5s	Number of ratings
Treatment	0.052	-0.011	-0.006	0.005	0.001	0.008	1.958
	(0.054)	(0.011)	(0.007)	(0.005)	(0.008)	(0.013)	(9.207)
Stores	141	141	141	141	141	141	141
Mean DV if Treat=0	4.171	0.0737	0.0347	0.0896	0.251	0.551	133.4

Notes: Standard errors clustered by store are in parentheses. Panels A and B here are similar to Table 2, and Panel C here is similar to Panel B of Table 3. However, instead of analyzing data from the RCT, it uses data from the post-RCT rollout. * significant at 10%; ** significant at 5%; *** significant at 1%

Table B7: Regional Manager Predictions, Part 1

Yes	Prediction
1	Would be very happy about less bureaucracy, less work as a result, do not like to work with notes and strict rules, will work
0	Both: Some employees are happy about fewer guidelines, others need strict rules
0	Will have a positive impact on employee satisfaction; but: poor communication of initiative by store management expected, might have negative impact on sal
1	Great, well-coordinated team in the store, everything fits in the store, would appreciate less bureaucracy
0	Both: employees will be happy, but individual employees need more restrictions
1	Well-coordinated team, has been working together for a long time, very good communication within the team, would be glad, no negative effects; will work!
0	Negative effects, as the team is still very fresh, new management in place, processes not yet internalised; Negative sales
0	Both. Employees will be glad, mixed team with some old and many young employees.
1	Would perhaps miss the list; but: no negative consequences in the store; on the contrary: positive impact!
0	Will be glad; but: implementation of processes not secure, chaotic store; internal evaluations (e.g. strawberries on a cake) mostly negative. Might be chaotic w
1	Would implement this very well, would also get along well without paper and clear structure; employee satisfaction will increase
0	Many new staff members, store is a bit chaotic, need structure and guidance, want guidance
1	Get along without bureaucracy; would feel more comfortable if there was less pressure because of less bureaucracy. Will work
1	Get along without bureaucracy, nothing would change in the operational processes without bureaucracy, staff already understood important things
0	Mixed picture; have too high returns on bakery products, returns will get worse. Unclear how it will work
1	Get along without bureaucracy, nothing would change. Therefore, will work
0	Need structure, will not work without it, otherwise the store will sink into chaos and lose focus
0	Need structure, haven't been around long, bureaucracy is important support, returns on bakery products
0	Need structure and bureaucracy, otherwise staff will have problems
1	Yes, will work
1	Yes will work
1	Yes, will work
0	No, will not work
1	Yes, will work
0	No, does not work
0	No, does not work
1	Yes, will work. Clear yes
0	No, does not work. No way
0	No, does not work. No way
1	Will work. Good and organized store management; very conscientious and tidy. Implementation will work
0	Need assistance. Complicated without lists, young store management, young team needs guidance
0	Undecided. Maintain documentation obligations, as other structure is difficult to implement; old store management, which wants to maintain habits
1	Store team does not need lists. Committed, thoughtful and conscientious
0	Store desperately needs structure which is provided by bureaucracy; organized store management , bad team. Will not work without lists
0	Good leadership, bad team. Would work partially
0	Would be good if lists remained. Recent change of management. Large store
1	Would work. Complete Confidence in the team
1	No documentation requirements needed. Good team. Good store
1	No documentation requirements needed. Good team and store management. Well organized
0	Will not work - team is still finding itself; guidance and structure needed; possible problems if list isn't there anymore. If there's a mystery shopping visit and
1	Does not work in this store as good as in store . . . ., but will work as well; maybe some structure needed, also autonomous possible. Will work
1	Similar to store in . . . .; team will be glad; actually need list to get routine, would also work out without list
1	Will work out without any requirements, team is confident in their performance, happy if there are no lists
1	Like in store . . . . Team will manage it, but need to stay focused. Problem: When there is a Mystery Shopping visit and expectations are not met, there will b
1	Team does not need lists. Can manage without lists. Strength in implementing processes.
1	No lists needed, works out without lists. However, when the store manager is not on duty, they sometimes not meet expectations
0	List needed for orientation. Does not work without it.
1	Definitely do not need lists, will implement everything in any case
1	Do well without a list
1	In general: will work out
1	Will work out.
1	Could do well without lists and without having problems, would like to keep daily log
0	Focus store; cannot work without clear guidelines, may result in chaos
1	There won't be any problems with less bureacracy, even if daily log is important from time to time
0	Focus store; cannot work without clear guidelines, may result in chaos
0	Cannot work without it, cash differences
1	Can do without it, store runs great
1	Can do without documentation requirements, runs great, but still relatively new store management
0	Cannot do without it, big cash and store differences and problems with sales; even if employees would like to pass on restrictions
1	Will work out without restrictions
1	Will work out without restrictions
1	Will work out without restrictions
1	Will work out without restrictions
1	Will work out without restrictions
1	Will work out without restrictions
1	Will work out without restrictions
0	Structure needed. Won't work without it
1	Will work out without restrictions
0	Staff will be glad; procedures are sometimes problematic, often not implemented, therefore bureaucracy and structure needed
0	Store management wants to maintain bureaucracy; but it could work as well. Unclear if it works out
0	?
0	Store wants to keep bureaucracy, unclear if it works out
0	Store wants to maintain bureaucracy, clear structures important for training and coaching. Unclear what happens
0	Store wants to maintain bureaucracy, important for training and coaching; mixed effects
0	Store wants to maintain bureaucracy, important for training and coaching, has to deal with store differences and leadership
0	Store wants to maintain bureaucracy, important for training and coaching, has to deal with store differences and leadership
0	Sometimes help needed, large store, operationally strong, so could also work out
1	Can be left out, very strong store management, trains staff very well
1	Can be left out, small store, few employees, can also be trained in person

Notes: This table gives the first table of regional manager predictions. What is listed here are the predictions that a coauthor wrote down in pen form during the phone calls with regional managers. Due to sensitives and legal restrictions on recording phone calls in Germany, it was not feasible to record the phone calls.

Table B8: Regional Manager Predictions, Part 2

Yes	Prediction
0	New store management old established team, need guidance, but could work out in the medium term
0	No good store management, not good at training staff, clear guidance and lists are important
1	Works out without, small store, staff are well trained and guided by store management
0	Do not leave out, big team, difficult cases,information does not go down well
1	Training on important processes is also possible this way, control can be omitted, will work out
0	New store management, lists are needed
0	New store management, lists are needed, but store management is probably good, best case: keep first, leave out later
1	Independent store, will work out without lists, employee satisfaction will improve
0	Downtown store, no positive or negative developments on sales or performance, high employee satisfaction anyway
1	Similar to other well running stores, team will be glad if lists are gone, no change in sales (maybe better sales), time is saved, no change in other numbers
0	Similar to other well running stores, team will be glad if lists are gone, no change in sales, time is saved, no change in other numbers
1	If operational list is gone, it's good for the team, it will work
1	Always enjoyed making lists and bureaucracy, but will also work out well without restrictions
0	Always enjoyed bureaucracy. Old employees and therefore difficulties without it
1	Team will be glad when operational list is gone. No problems expected. Will work out!
0	Rather neutral. Mixed effects. No operational list is good, more time for employees
0	Will not be received well,. Daily protocol and operational lists are popular; employees like bureaucracy
0	Like bureaucracy, will find another way, will neither be happy nor sad; neutral effects
0	Bureaucracy needed
1	Will work out without
1	Will work out without
0	Documentation requirements are needed
1	Could live without bureaucracy, very communicative store management
0	Daily protocol needed, operational list not necessarily. Therefore mixed effects
0	Bureaucracy needed, will not work out without
1	Strong store management, high sales, employee satisfaction 50/50, store management will not take omitting lists seriously, because there are so many other list
1	Strong store management, been there for a long time, high employee satisfaction, it will work out very well without documentation requirements
0	Currently closed, strong store management, employee satisfaction high and will improve
1	Small store, on a positive trajectory, new store management, will accept bureaucracy reduction and implement successfully. It's an opportunity!
0	Very strong store management, employee satisfaction will not change. Large store. But: operational implementation will work partially , no big problems
1	Strong store management, open to everything, high employee satisfaction, omitting lists will be successful
0	Small store, will be received positively, new store management, mixed effects
0	Very strong store management, employees been there for many year. Effects unclear
0	Will meet with resistance, will not accept anything new, will only reluctantly, if at all, let themselves be dragged into it, store management communicates this
1	Strong store management, open to everything and can implement everything well, already been there a few years
1	Employee satisfaction will improve with less bureaucracy, strong store management, will work out
1	Interested store management, will be happy about it, positive emotional response, higher employee satisfaction; Omitting will work out
1	Top motivated store management, positive emotional response, store management takes on many tasks itself, less bureaucracy will be supportive
1	Focal point store, motivated store management; store management takes over a lot of bureaucracy from staff; employee satisfaction will not improve necessarily
0	Critical store, employee satisfaction will not get better, does not work out
1	Mini store, hardly any bureaucracy, will work out
1	Mini store, hardly any bureaucracy only 3 employees will be happy when there is less bureaucracy
1	Store management will be happy that lists/ bureaucracy are gone, but then say: does not help much; employee satisfaction will not increase, but it will work v
1	Highly motivated store team, very communicative, maybe no increase in sales or staff satisfaction , because store is already productive, will work without lists
0	Old store management, if it is up to them they will continue to run lists; no change in sales, independent from restrictions - store will be ok
1	Great store management, will work hard on it and implement it well, will analyze whether it is successful. Will work. Positive influence; employees very satisfi
0	Employees are dissatisfied with the situation in the store, there are grumblings, feeling relieved because of less bureaucracy could help, unclear what happens
1	Will work, good store and well organized store management
0	Problem team, a bit chaotic. Won't work without guidelines and clear guidelines
1	Could probably work, well organized store management and team
1	Will work, but: store management is very bureaucratic
1	Store manager retiring soon. Could work out- well-functioning team; unclear if open to changes, but it will work in general
1	Could work, or rather: Will work!!
0	No, will not work
1	Will work. But team needs to know why
0	At the moment, no. Will not work
1	Yes, we implement well, but want to understand why. But: If explanation makes sense (which may be the case), it will work
1	Bureaucracy costs time; more time has a positive effect on satisfaction; will work out
0	Older employees, very bureaucratic, keep handwritten lists, love bureaucracy, unclear
1	Less bureaucracy saves time; more time = positive for employee satisfaction, young team, easy-going
1	Less bureaucracy saves time; more time = positive for satisfaction, young team, more relaxed and more free time
0	Structures and control needed
0	Will improve the general mood, are often overwhelmed with bureaucracy; employee satisfaction and sales will not improve
0	Undecided
0	Store management over 20 years in, undecided
0	Less bureaucracy will improve the general mood; but: employee satisfaction and sale will not improve. Unclear what happens
0	Undecided

Notes: This table gives the second table of regional manager predictions. What is listed here are the predictions that a coauthor wrote down in pen form during the phone calls with regional managers. Due to sensitives and legal restrictions on recording phone calls in Germany, it was not feasible to record the phone calls.

Table B9: Classification Agreement and Treatment Effects on Log Sales by Quartile of Affected Stores

	Classification agreement	TE: Random forests	TE: Sorted effects
Q1 (least affected)	0.806	-0.018	-0.004
		(0.018)	(0.023)
Q2	0.739	-0.041*	-0.010
		(0.024)	(0.019)
Q3	0.750	0.046**	0.021
		(0.021)	(0.023)
Q4 (most affected)	0.859	0.139***	0.100**
		(0.041)	(0.042)
p-value equal TE		0.000	0.094

Notes: This table reports estimated treatment effect on sales by quartile of stores affected by treatment as classified by random forests and sorted effects procedures, as well as the classification agreement between the two procedures. Standard errors clustered by store are in parentheses. Controls are the same as in Table 2. * significant at 10%; ** significant at 5%; *** significant at 1%

Table B10: Pre-RCT Headcount, Tenure, and Regional Manager Predictions by Quartile of Affected Stores

	Random forests			Sorted effects		
	Headcount	Tenure	Mgr Prediction	Headcount	Tenure	Mgr Prediction
Q1 (least affected)	15.662	70.122	0.000	15.447	72.827	0.083
	(4.637)	(21.543)	(0.000)	(4.761)	(23.220)	(0.277)
Q2	15.804	80.191	0.531	14.540	79.842	0.444
	(4.042)	(25.864)	(0.500)	(4.602)	(29.568)	(0.498)
Q3	12.790	78.661	0.821	13.376	81.224	0.719
	(4.412)	(28.378)	(0.384)	(4.330)	(20.033)	(0.450)
Q4 (most affected)	10.125	89.021	0.728	10.988	84.160	0.839
	(2.692)	(27.957)	(0.445)	(3.551)	(31.881)	(0.368)
Unadjusted p-value equal	0.000	0.011	0.000	0.000	0.262	0.000
WY p-value equal	0.000	0.002	0.000	0.000	0.075	0.000

This table reports descriptive statistics on our pre-registered dimensions of heterogeneity by quartile of stores affected by treatment as classified by random forests and sorted effects procedures. Additionally, we report p-values of the mean equality tests across the quartiles, both unadjusted and Wesfall-Young (1993)-adjusted for multiple hypothesis testing. Standard errors are clustered by store.

Appendix C Materials Used in the RCT and Firmwide Rollout

C.1 Wording Used for the Regional Manager Predictions

I presented the pilot project in a regional manager meeting in Feb 2021. I received the following feedback about the pilot project from the regional managers:

"In some shops, less documentation duties will work well in the daily business operations and will probably have a positive effect on store performance indicators. In other shops the reduction will have negative effect on the daily business and will probably have a negative impact on store performance indicators."

We as researchers are interested in your predictions!

Now I will ask you to make predictions for all of your shops (independent whether the shop will indeed be a pilot shop or not).

I have now a list of your shops (in front of me)

What do you think: If the shop XYZ indeed was a pilot shop: How well would the daily business work ("function") in the shop with less documentation duties?

C.2 Information on the RCT Provided to Store Managers and Employees

Section 2 of the paper provides the message to store workers and managers in treatment stores regarding the elimination of the two checklists. This message was translated into English by two coauthors (one native German speaking, one native English speaking). The message was relatively straightforward to translate. We translate one part as "This gives you more freedom to organize yourselves", as the German word is "freiraum", which has the dictionary meaning of freedom in English. The phrase could also be translated as "empower", as in "This empowers you to organize yourselves."

C.3 Examples of Older Versions of the Operational Checklist and Daily Protocol

Below are two examples of the operational checklist in the past.

The first is from August 2019. Instead of signing, workers indicate whether they did well, poorly, or average on different tasks.

The second is from January 2017. Workers would sign this at multiple points during the day.

INCREASING AVERAGE CUSTOMER SALES

Challenge August 2019

We are NAME OF THE COMPANY

A = Authentic → The customer realizes how authentic you are based on *your* voice, *your* smile and *your* sense of humor
P = Passion → Get excited about seeing your customers and give them compliments – selling is passion

Toolbox:

Your name				
Date				
Evaluation	+/~/-	+/~/-	+/~/-	+/~/-
1. Fulfil a desire **Eye contact, smile, confirm customer desire and maybe upgrade** Big bag used? Big serving tray used?				
2. Sample plate – point it out or physically offer it **Maybe offer a second sample?** Use a generous-sized sample – surprise the customer Ask if customer wants to buy more of the product?				
3. Fun with the customer Say one sentence more than usual + e.g. point out that they can buy more				
4. Give positive feedback to the customer and offer them the opportunity to buy more				
5. Say goodbye to each customer in an individualized way				

Customer list: Goal → Increase customer satisfaction!!!

How are we perceived by the customer? Do you personally find the presentation of the products in the sales counter appealing?
What do we really offer to the customer?
In addition to you, the store manager or sales agent leading the shift checks the following checklist at the respective points in time and signs on the checklist

1) After arrival of 1st shipment, around 7:30 am
2) After arrival of 2nd shipment, around 10:00 am
3) Shift change / start of new shift, around 12:45 pm
4) At cake time, around 3 pm
5) Evening rush hour, around 6 pm

1)	Quality: a) Put all golden rolls and one other roll of each type in a red box and evaluate the quality of the rolls (fully baked, favorable appearance,…) All types of rolls available? Were there any product shortages that were relieved, and who did it? b) Review baking plan (During the baking time? Next baking process prepared)? c) Sample roll ok? d) Give brief feedback to the women who are baking (positive encouragement… and maybe something to improve?)
2)	Service: a) Service speed ok? (Run to the customer, no queues…) b) Service friendliness ok? (Smile, eye contact with customer, melodious voice, say goodbye) c) Service advice ok? (Did you offer or recommend anything? d) Presentation ok? (Bread, cake, snack, sales counter, promotion product correctly placed?) → Is the customer really aware of our "promotion initiative" or the "hint of the day" (poster ok?) Price tags correct and placed everywhere? Sample plate full of sample goods? Price tags at sample products correctly placed? → Can customers see poster "enjoy hot" near the paninis and hot sandwiches
3)	Hygiene: a) Is the glass of the sales counter clean? If not, clean immediately! b) Look around (above and below the sales counter): Remove spider webs, keep deposit vouchers! Floor / cold sales counter are (inside) clean? c) Check: Cutlery still there + clean + polished? Enough milk, sugar, stirrers… in boxes? d) Café and coffee area outside clean? Wipe tables, sweep? Are the corners and the cushions clean? Is the bin in the café clean? All tables and chairs set up, sun umbrella opened…? e) Menu available on each table? If not – set up! – if missing, did you already order a new one? f) Doors to the side room / bathrooms clean? Smell ok? e) Clean in front of the counter? Sweep!

C.4 Guidelines Given to Regional Manager Explaining the RCT: Mid-February 2021

Guideline: regional managers

What is it about?

At [FIRM NAME] we constantly ask ourselves how and where we can improve to make our employees daily work easier. Together with the workers' council and a team of researchers from the University of Cologne, we started discussions on day-to-day business documentation duties (daily protocol, expiry date checklist, weekly report, etc.) at [FIRM NAME] in 2020.

In a joint pilot-project with the research team we will forego the daily handling of the *operational checklist* as well as the *daily protocol* in 75 randomly selected [FIRM NAME] pilot stores for an initial period of six months, starting April 6th, 2021. In doing so, we give the employees more freedom to organize themselves. The *operational checklist* and the *daily protocol* are continued in all other stores.

The aim of the pilot-project is to scientifically test what are the effects of waiving the two documentation duties. Your cooperation is essential for the success of the pilot project.

Trust your managers in the pilot stores.

What must be done in pilot stores?

Please inform all store managers and employees in pilot stores that the *operational checklist* and the *daily protocol* will no longer be used. Emphasize particularly that we want to give the employees more freedom to organize themselves and that we trust the employees will continue to do the essential processes (such as the arrangement of the products in the sales counter, covid measures, customer communication) in a company-compliant manner. You should ensure that store managers and employees in pilot stores will no longer provide written confirmation that operational processes have been implemented in the right way.

Please make it clear to employees that time saved on paperwork is an opportunity that we can use especially for training new colleagues and communicating with customers.

Will the previous information in the *operational checklist* and the *daily protocol* be recorded elsewhere in the pilot stores?

The *operational checklist* and the *daily protocol* will be dropped in pilot stores without any replacement; the employees must not confirm in writing anymore that the corresponding tasks are being completed.

In the future, the "cash balances" will be recorded exclusively by the "money transfer list" in pilot stores.

In which stores will the *operational checklist* and the *daily protocol* be dropped?

The *operational checklist* and the *daily protocol* will initially be deleted only in 75 randomly selected [FIRM NAME] (pilot) stores. **In all other stores**, the *operational checklist* and the *daily protocol* will **continue to be used in the future as before**. Please ensure this and support your store managers in the implementation.

In order to ensure fairness in the selection of pilot stores, pilot stores were chosen at random. The selection was made by the research team from the University of Cologne and was supported by the workers' council. Since the stores were selected at random, it also happens within the districts that the *operational checklist* and the *daily protocol* are kept in some stores but not in others.

Please make sure that the *operational checklist* and the *daily protocol* are continued or deleted in the "correct" stores. Please do not reintroduce the *operational checklist* and the *daily protocol* in the pilot stores on your own **under any circumstances.**

This would jeopardize the success of the entire project!

How will I respond to queries from stores managers and employees?

If you receive any questions from employees or store managers that you cannot answer, please contact your sales director.

If store managers ask why the *operational checklist* and the *daily protocol* are being continued in their stores, while hearing that this is no longer the case in other stores, please answer as follows:

As a part of a pilot project, the operational checklist and the daily protocol will no longer be used in randomly selected pilot stores for several months. For reasons of fairness, the pilot stores were randomly selected so that each store had the same chance of becoming a pilot store. The stores were drawn by a team of researchers from the University of Cologne together with the workers' council. If you have any questions about this, please do not hesitate to contact [*NAME OF THE HEAD OF THE WORKERS' COUNCIL*]*, who is supporting the project on the part of the workers' council.*

Further notes: Contact to the research team

The research team from the University of Cologne will conduct a survey among all store managers in March 2021. The aim here is mainly to determine when the store managers and employees usually fill out the *operational checklist* and the *daily protocol* and how much time this takes. As a part of the survey the research team will call the store managers directly in the stores on Wednesday mornings in March. You should inform your store managers in advance about the survey.

During the pilot project, the research team will also contact the regional managers regularly to ask for their personal impressions of the impact of the removal of the *operational checklist* and the *daily protocol*.

Appendix References

Bruhn, Miriam, & McKenzie, David. 2009. In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *AEJ: Applied*, **1**(4), 200–232.

De Quidt, Jonathan, Fallucchi, Francesco, Kölle, Felix, Nosenzo, Daniele, & Quercia, Simone. 2017. Bonus Versus Penalty: How Robust are the effects of contract framing? *Journal of the Economic Science Association*, **3**(2), 174–182.

Dellarocas, Chrysanthos, & Wood, Charles A. 2008. The Sound of Silence in Online Feedback: Estimating Trading Risks in the Presence of Reporting Bias. *Management Science*, **54**(3), 460–476.

Falk, Armin, & Kosfeld, Michael. 2006. The Hidden Costs of Control. *American Economic Review*, **96**(5), 1611–1630.

Hagemann, Petra. 2007. What's in a frame? Comment on: The Hidden Costs of Control. *Unpublished manuscript, University of Cologne*.

Hossain, Tanjim, & List, John A. 2012. The Behavioralist Visits the Factory: Increasing Productivity using Simple Framing Manipulations. *Management Science*, **58**(12), 2151–2167.

Imai, Kosuke, Keele, Luke, & Tingley, Dustin. 2010a. A General Approach to Causal Mediation Analysis. *Psychological Methods*, **15**(4), 309.

Imai, Kosuke, Keele, Luke, & Yamamoto, Teppei. 2010b. Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 51–71.

Schnedler, Wendelin, & Vadovic, Radovan. 2011. Legitimacy of Control. *Journal of Economics & Management Strategy*, **20**(4), 985–1009.

Tadelis, Steven. 2016. Reputation and Feedback Systems in Online Platform Markets. *Annual Review of Economics*, **8**, 321–340.

Tazhitdinova, Alisa. 2022. Increasing Hours Worked: Moonlighting Responses to a Large Tax Reform. *American Economic Journal: Economic Policy*, **14**(1), 473–500.