Is This Really Kneaded? Identifying and Eliminating Potentially Harmful Monitoring Practices*

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

Tobias Kretschmer[¶] Nick Zubanov[‖]

April 2023

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 by 2-3% and store manager attrition substantially decreases. Mystery shopping indicates this occurs without a rise in workplace problems. Before random assignment, regional managers predict whether the treatment would be effective for each of the stores that they oversee. Ex post, beneficial effects of checklist removal are fully concentrated in stores where regional managers predict that 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 that employees regard as demeaning, but keeping a checklist that helps coordinate production.

Keywords: Monitoring; checklists; respect; time use

PRELIMINARY. PLEASE DO NOT CIRCULATE WITHOUT PERMISSION.

*We thank the study firm and its management for their enthusiastic participation in this collaboration. We thank Alessandra Fenizia, Michael Kosfeld, Axel Ockenfels, Paul Oyer, Andrea Prat, Kathryn Shaw, Lowell Taylor, John Van Reenen, Melanie Wasserman, and numerous seminar participants for helpful comments. 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 firm's workers' council, which represents worker interests at the company level, approved the project and was involved in all steps.

[†]Goethe University of Frankfurt and CEPR and IZA

[‡]University of Cologne and CEPR

[§]U. Toronto Rotman School of Management and NBER and CEPR

[¶]LMU Munich and CEPR

[‖]University of Konstanz and IZA

Monitoring, broadly defined as keeping a close watch on production, is widely viewed as an important practice at work. Randomized controlled trials (RCTs), detailed below, show sizable performance benefits from organizations adopting monitoring, and aspects of monitoring are scored in the World Management Survey (Bloom & Van Reenen, 2007). However, monitoring need not always be beneficial for worker well-being or firm performance. Beyond the financial cost, monitoring takes time, both for the monitor and the worker being monitored. In addition to time costs, workers may dislike being monitored, not only because it prevents them from goofing off, but also because monitoring may signal mistrust (Falk & Kosfeld, 2006) and contribute to a negative work environment. Indeed, popular observers express concern that many workplaces, from call-centers to Amazon warehouses to tech firms, use unpleasant and potentially counterproductive monitoring (Guendelsberger, 2019).[1]

One of the most common forms of monitoring is checklists. Checklists are celebrated as a powerful tool to help workers remember things and coordinate production (Gawande, 2010). While most famously used in surgery and aviation, checklists have been applied in numerous industries, including retail. Studies document large benefits of checklists, but little is known about whether and when checklists can be harmful, and this is for several reasons. First, firms' use of checklists is highly non-random, making it difficult to estimate causal returns. Second, it seems likely that the returns to monitoring via checklists are highly heterogeneous, reflecting that some workers benefit from monitoring, whereas others may find it useless and insulting. Capturing this heterogeneity seems critical for a full understanding of monitoring. Third, modern firms often use numerous checklists, so even if one believes a workplace is "overmonitored," it is hard to know which checklists to modify.

In this paper, 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. As far as we know, ours is the first large-scale RCT on removing monitoring at work. Our research partner is a major German bakery chain with 145 stores and over €100m of annual revenue. The firm is family-run and prior to our intervention 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 interactions with customers, such as whether they smiled. Drawing on a deep collaboration with the firm and top management, we conducted extensive pre-RCT interviews and surveys, and discovered that several checklists were perceived as especially low-value (i.e., high time costs and limited benefits).

The RCT is grounded in a simple conceptual framework of monitoring, as laid out

[1] See also the 2022 articles in the Economist ("Welcome to the era of the hyper-surveilled office") and New York Times ("The Rise of the Worker Productivity Score").

in Section 1. Monitoring through checklists helps firms address moral hazard problems, coordinate production, and remind workers of tasks. However, checklists also entails 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 what checklists are best to remove and what stores may benefit most from checklist removal.

As detailed in Section 2, the bakery chain we study represents an ideal setting for our RCT. First, the sample is large. Second, we access highly granular administrative data, coupled with the ability to conduct high-quality, detailed surveys. The administrative data cover detailed aspects of sales, customers, and orders hour by hour, which is critical for examining how workers and managers are using their time and how they substitute time on checklists to other tasks. Because of our deep collaboration, the surveys we conduct have very high response rates, as well as 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) in detail and have them make predictions about in what stores the RCT will be most successful. Using a project team consisting of firm executives, the head of the worker council, and two of the researchers, and using surveys of workers and store managers, we document wide variation across monitoring tasks in their perceived value and how much time they take per week.

In Section 3, focusing first on the overall effects of the RCT, we estimate that removing checklists increases sales by 2.7%. The impact on sales is similar during busy and less busy times. While one may be concerned that removing monitoring would lead to wasted food, increases in employee misbehavior, or coordination failures, we observe no negative impact on shrinkage (a joint measure of food waste and worker stealing). We also observe no negative impact on mystery shopping scores, i.e., the scores prepared by undercover shoppers. Our bakery firm has relatively low attrition, and there is no overall impact of the treatment on attrition. Still, there is a strong negative effect on the attrition of store managers, who do a lot of the checklist completion and who are naturally likely to appreciate having less of it. In contrast, the treatment has a positive, though statistically insignificant, effect on the attrition of unskilled workers without vocational training who may benefit from structure and checklists.

Our initial discussions with regional managers highlighted that the impacts of the RCT on outcomes would likely be highly heterogeneous across stores. In our pre-RCT survey of regional managers, managers predicted that in about half of their stores the treatment would be effective, and in the other half they 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 the 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 down, though the impact is not statistically significant.

To better understand these 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. Unlike past interventions in the literature, our treatment was taking something away instead of adding something, so the direct cost to implement the RCT was very low. For minimal cost, the firm received a sustained increase in sales, as well as a reduction in manager turnover. Therefore, the firm decided to implement checklist removal firmwide. However, while the RCT involved eliminating two checklists, the firm decided to restore 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.[2] 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.

[2]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).

In economics RCTs on monitoring, most related to ours is a seminal paper by Nagin *et al.* (2002), who consider a field experiment where a call-center company exogenously varies its monitoring rate in some 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,[3] 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).[4] 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. In the midst of substantial work on the general importance of management practices (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. A rich and growing literature examines what do non-CEO managers do and their impact (Lazear *et al.*, 2015; Friebel *et al.*, 2022; Hoffman & Tadelis, 2021), often emphasizing the role of managers in motivating and teaching employees. Our results suggest that an important

[3] 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.

[4] 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.

way managers can add value is by having private information about their teams.

Third, our paper makes a methodological contribution to RCTs. Beginning with DellaVigna & Pope (2018), growing work uses expert predictions for the purpose of examining how the results of an RCT compared to priors of experts, that is, to see to what extent a result is surprising or not (DellaVigna *et al.*, 2019). Rather than having experts predict the average results of the RCT (e.g., that the treatment will increase or decrease 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 other work that uses expert predictions in RCTs in this manner, but we believe that this methodology may be useful in other contexts.[5] Experts in our context have substantial knowledge about which units will be most affected.

Fourth, 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 monitoring such as checklists on performance, and how would the impact of monitoring vary across stores within a firm? To address these questions, we model the impact of implementing a binary monitoring technology (monitor or not) in a store. In addition to shedding light on these two key 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. Monitoring is randomized in our RCT, so we focus on the impact of monitoring instead of the decision to monitor.

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 problems, such as where people on a surgery team forget to take 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

[5] E.g., one could ask doctors to predict which individual patients will respond to a new drug. The only other RCTs we are aware that do something similar are Bryan *et al.* (2021), who ask loan officers to predict how individual microfinance clients will fare under various treatments, and Dal Bó *et al.* (2021), who ask supervisors of government agricultural workers to rank which workers should get free cellphones. Our prediction setup differs in that we focus on the predictions of higher-up experts in a private-sector firm.

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 monitoring, firm profits are $-pk$.[6]

Monitoring such as checklists helps the firm identify problems.[7] The quality of monitoring is given by m, and represents the probability that a problem is detected and solved in full. Equally, one can assume that monitoring detects problems with 100% probability, but that only a share m of costs are recuperated. Using monitoring also 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, perhaps because it is intrinsically unpleasant to fill out checklists but also because monitoring can be viewed as a sign of disrespect (Ellingsen & Johannesson, 2007, 2008). Ellingsen & Johannesson (2008) argue that workplace respect can be thought of 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, being monitored 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., monitoring 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.[8]

Therefore, the profits from monitoring 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 to be positive when there are important direct and indirect costs of monitoring, as well as

[6] In line with Garicano (2000), one can imagine that stores differ not in the frequency of problems face, but rather in their ability to solve them. Thus, one can alternatively define p as the share of problems that a store cannot solve on its own without firm monitoring.

[7] 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.

[8] Pilots may be fine with checklists because the cost of problems is high. In retail, the costs of problems is smaller (e.g., someone doesn't buy an extra doughnut), so it may feel more unpleasant to be monitored. This could help explain why research on aviation checklists finds high returns. Having θ depend on p and k can also explain why high-functioning stores with fewer problems may be more frustrated by monitoring.

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.[9]

This framework also raises the possibility that there could be substantial heterogeneity across stores within a firm in the returns to monitoring. Regional managers may know that some stores experience frequent coordination problems and thus likely benefit from monitoring. Stores may also vary in the production costs of monitoring, such as if some workers dislike checklists more than others (e.g., if some workers find monitoring more disrespectful or wasteful than others), and stores may differentially complain about these costs to regional managers. Given that there are multiple factors affecting whether monitoring has positive effects, as well as that some factors (like frequency of coordination problems) are very difficult to observe in data, it is natural to ask regional managers to make predictions about whether a treatment will work in a store.

Formally, let the performance impact of the treatment be $z = c + \theta - mpk$. 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.[10] The private information a manager has is represented by the precision of the signal, $h_\epsilon = \frac{1}{\sigma_\epsilon^2}$. Managers believe the treatment will work when the treatment effect is above a threshold. Thus, the more private information that regional managers have about the components of 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 monitoring are not only costs to performance, but also to worker utility from the job. Our treatment is likely to reduce attrition most for workers with higher personal costs of monitoring, and could increase attrition for workers who personally benefit from monitoring (e.g., if checklists provide valued structure). Dube *et al.* (2022) find that workers care deeply

[9] 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.

[10] One would imagine that they have two signals, one regarding the frequency of problems, and one regarding the indirect costs of checklists. However, we only elicit a manager's overall belief about the treatment's effectiveness in each store.

about being respected and provide evidence that this matters for turnover. It is natural that higher qualified workers, measured both in experience and training, would have stronger effects on attrition.

2 Study Background

The firm. Our study firm is one of the largest bakery chains in one densely populated region of Germany.[11] 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 company 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.[12] The firm has a reputation for quality products, as evident, among other places, in online reviews.

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 and overmonitoring. First and foremost, we came across a 2018 employee survey which documented widespread general dissatisfaction with checklists at the firm. Second, the head

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

[12] 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).

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.[13] 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.[14] 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 documentations duties, respondents were asked the following questions:[15]

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?

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 duties stand out for

[13] Trained workers are employees who are trained through an apprenticeship, paid in part by the employer (Acemoglu & Pischke, 1998).

[14] Our approach of using a committee to generate comprehensive lists very broadly follows idea generation and process optimization procedures used in large firms like Toyota. Further details are in Appendix B.

[15] One of the 22 checklists was omitted from interviews. The "missing" documentation duty is the declaration of consent for working on Sundays. According to the works council, it is legally and politically impossible to drop, so it was not asked about. See Appendix B for details.

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.[16] The two remaining duties were the **operational checklist** (*Operative Liste*) and the **daily protocol** (*Tagesprotokoll*). The daily protocol and especially the operational checklist also perform poorly in terms of avoiding mistakes (Appendix Figure A4).

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.

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, the firm's pre-RCT debates on checklists paralleled tradeoffs emphasized in the academic literature.[17]

Operational checklist. The operational checklist is a detailed form with things to be done[18] 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

[16] For example, one of the sacred cows is the "sample roll" duty, 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 essential to the firm's unique selling proposition, so the sample roll duty could not be removed.

[17] The executives in the pro-structure school of thought had 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.

[18] Examples include: I smiled at the customers, 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.

spend some time reading it each day, so they are aware of what they are signing. Some executives initially thought that without the operational checklist, stores would experience significant operational problems, and that workers would not follow company guidelines (e.g., employees would forget to keep the shelves clean and to smile at customers).

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

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 duty we study is the daily protocol, where you write down all the things that happened during the day (see Figure 3 for the form). This includes how much money is in the cash register, how much sales taken in, and whether workers would like to pass along this information to the daily shift (this last point seemed especially appealing in bigger stores). In contrast to the operational checklist, some workers tend to find more value in the daily protocol.

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

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

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

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 had some strong local knowledge on this heterogeneity. Thus, in March 2021, before knowing which stores were in control or treatment, regional managers also 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).[19] 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 then translated into a numerical response of whether it will work. For almost all of the responses, there is little ambiguity about opinions, as we detail in the Appendix. No incentives are used for this prediction because it is a subjective one.[20]

Single-treatment RCT. The RCT uses a single treatment for several 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 would 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 effect heterogeneity, and we would be under-powered to detect such heterogeneity with multiple treatments.

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 message in the firm's weekly newsletter informing treatment stores (Appendix C.2) about the change came from the firm's COO, the son of the CEO/owner, which gave credibility and importance to the change. The message emphasizes two factors, paralleling our predictions on direct and indirect effects. First, it emphasizes how the company is

[19] The interviews were conducted by a coauthor (a chaired German professor) rather than an RA as a sign of respect to the senior managers and to elicit more serious and complete responses.

[20] 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.* (2022) argue that incentives are not needed to accurately elicit beliefs and discuss how incentives can sometimes worsen elicitation.

trusting workers (indirect effect). Second, it emphasizes the extra time (direct effect), and that workers should use the extra time for customers and colleagues. One reaction to this is that it might seem that workers are being "led" to think a certain way. However, 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. Store 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.[21] We also confirmed with regional managers, the head of HR, and one sales director in May 2021 that the treatments were 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 December 2021. Given the success of the RCT, the firm decided to roll out the treatment to control stores starting end of January 2022. Specifically, the operational checklist remained removed from the company, 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.[22]

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 covering worker attrition and worker absence.

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

[22] Specifically, we had been promised 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, leading to the endline survey (and end of the RCT) being postponed for 4 months. This fortuitously gives us more data and was obviously not driven by any statistical power considerations.

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 November 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 this variable described more in Appendix B), 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 Bruhn & McKenzie (2009). 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.[23] 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

[23] 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 to occur 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.

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. 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, and we can reject that there are significant negative effects.

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 quarter. Effects are relatively constant over the three quarters of the RCT. Even in the last quarter of the RCT, coming 6-10 months after the treatment is introduced, checklist removal increases sales by 3%, which is statistically significant at the 5% level.

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 the mystery shopping score. As seen in Appendix Table A2, we see no consistent evidence that the treatment harmed individual components of the mystery shopping. 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, interactions with customers, 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 firms, employee 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 will exit in a given year), and the relatively low incidence of attrition places limits statistical power.

As seen in column 1, there is no overall effect of the treatment on attrition. However, this masks substantial heterogeneity by employee skills. An important distinction at the firm is between trained and untrained employees. Trained workers are individuals who already did a 3-year apprenticeship, often from the bakery firm, and the firm is keen to retain these workers who have received expensive training.[24] In contrast, untrained workers have fewer skills and are less important for the firm to retain. Among untrained workers, attrition increases by a statistically insignificant 0.64 percentage points (hereafter, "pp") per month, which is an increase of roughly 20% relative to the control. However, among trained workers, attrition decreases by 0.45pp per month, which is a 35% decrease in attrition. An explanation for this is that untrained workers could benefit from added structure due

[24] As discussed above, the firm prides itself on offering high-quality products, and having trained workers is an important part of their high-quality strategy.

to their lack of knowledge at the firm. Within trained workers, some are store managers, whereas others work in the stores. The overall effect on skilled workers is driven by managers. Among managers, attrition decreases by over 1pp per month, a reduction of roughly half, and statistically significant at the 10% level. This decrease is clear 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 likely reason is that the costs of checklists are especially strong for managers, particularly in the case of the daily protocol. Managers spend almost an hour per week completing the daily protocol, whereas for workers the required duration is closer to half an hour. In pre-RCT focus groups and discussion with the firm, there was a feeling that checklists were preventing store managers from doing some high-value activities, such as mentoring and teaching workers. It is also possible that utility costs of monitoring are especially bothersome for managers. Managers are supposed to act as leaders and monitors in the store. When the firm uses extensive checklists on top of this, the firm communicates that it does not trust the store manager to perform these functions by themselves.

Appendix Figure A5 shows that there is no evidence that the treatment effect on trained worker attrition fades over time. As seen in panel (a), if anything, the treatment is slightly stronger in months 6-10 of the RCT, but we cannot reject that effects are constant over the RCT.

Magnitudes. How should we think about the magnitudes of the estimates? In a study in another 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, as the treatment in Friebel *et al.* (2017) involves an increase of wages of 2.2%, whereas 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 a seminal 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 in the workplace. They find that 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, the effect of removing two perceivedly low-value checklists in our setting leads to treatment effects on the order of some of the most promising and highly regarded past management interventions. At the same time, we believe that our effect sizes are very plausible. Some readers may be surprised that removing checklists has such quantitatively substantial effects. It is critical to remember that workers and managers regarded the checklists removed as onerous ones, with relatively low benefit and high time cost.

Using worker surveys and customer reviews to understand mechanisms. Given the prominence of the effects, an important question is why do they occur. Critically, why is sales going up? Why is retention increasing for many workers? To shed light on these questions, we surveyed workers during the RCT, and we also scraped data from Google reviews of the stores.

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.[25] 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.

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 right 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, whereas the share of 5s significantly increases.

To shed light on what is happening in treatment stores to warrant higher customer

[25]There was no impact 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.

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 the Appendix, including the issue that many reviews do not contain text. First, the share of reviews mentioning speed of service increases by roughly 160%, from 0.5% of reviews in control store to 1.3% of reviews 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 move 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).[26] Overall, customers in treatment stores report being served faster and receiving a higher-quality product, and this manifests itself in higher overall ratings.[27]

4 Heterogeneity Results

Regional managers had very strong beliefs about in which stores the treatment will be successful. Therefore, we focus our analysis of heterogeneity based on regional manager expectations. After presenting these results, we consider heterogeneity in general.

Heterogeneity in results by regional manager expectations. Table 4 separates the treatment effect on store outcomes by regional manager expectations, showing that the treatment effect is much stronger in stores where regional managers expected the treatment to be beneficial. In stores where managers expected the treatment to be beneficial (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 latter difference is

[26]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.

[27]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 German bakeries. Friebel *et al.* (2017) do not have data on customer reviews, so our evidence is more direct.

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 latter difference is also insignificant.

Returning to Figure 4, panels (b) and (c) 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.[28]

The final row in Table 4 shows p-values testing the differences between stores where regional managers predict the treatment to work. We see that the difference in effects on sales (regular, busy sales, and non-busy sales) and customers are statistically significant. We show two-sided p-values so as to be very conservative, though one can easily argue that one-sided p-values are more appropriate given the explicitly one-sided prediction of store managers (i.e., dividing stores in the ones where the treatment will work and ones where it will not work).

Likewise, Table 5 shows attrition results separating by regional manager expectations. Focusing first on overall attrition, the treatment decreases attrition by 0.5pp in stores where the treatment is predicted to work and increase attrition by 0.5pp in stores where the treatment is predicted not to work, though both effects are not significant. However, among trained workers, the treatment reduces attrition by about 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 reduction in managerial attrition is entirely driven by stores where the treatment is predicted to be successful. In stores where the treatment is expected to work, we estimate that the treatment reduces attrition by 2.2pp per month, essentially a complete reduction relative to the control group mean. In the raw data, in stores where the treatment is expected to work, there are 8 store manager quits in controls stores, but only 1 in treatment stores. In contrast, in stores where the treatment is not expected to work, the effect is zero. This difference is statistically significant at the 5% level.[29]

[28] Appendix Figure A6 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.

[29] One interesting pattern is that store manager attrition is 3 times higher in stores where the treatment is predicted to work compared not to work. There are several intuitive reasons for this, all 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.

Robustness of regional manager expectations as a source of heterogeneity. A natural question in considering our heterogeneity results by regional manager expectations is whether they are driven by some variable correlated with regional manager expectations instead of the expectations themselves. To assess the robustness of regional manager expectations as a source of heterogeneity, we consider several approaches, including machine learning, all of which strongly support robustness.

First, we interact the treatment variable with many pre-RCT store characteristics, and find that the interaction on treatment X manager prediction remains highly robust. However, a concern with this is that many pre-RCT characteristics are highly correlated. Thus, second, we perform an elastic net-regularized regression which prunes variables which do not have sufficient predictive power. Treatment X regional manager prediction is one of only a small number of heterogeneity variables to escape pruning and is much more predictive.

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 A3 and A4.

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. Among the 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 the traditional economic view of monitoring as a way of addressing problems (Holmstrom, 1979; Halac & Prat, 2016), as well as theories emphasizing the utility costs of monitoring (Falk & Kosfeld,

2006).

Table 7 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 differently 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.

Other heterogeneity. Beyond regional manager predictions, we also pre-registered that we would examine heterogeneity according to team size and team tenure. Table 8 shows that the treatment effect on sales is significantly larger in smaller teams, defined as having a head count that is 10 workers or below. In contrast, there is no significant heterogeneity according to team tenure.

5 Additional Analyses and Threats to Validity

5.1 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, 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 A1 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 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.

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.

5.2 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). An immediate piece of evidence against this is the time path of sales and turnover effects. Sales improves immediately in the first quarter (Figure 4), whereas turnover effects are negligible in the first quarter of the RCT and become larger each quarter (Figure A5). A formal mediation analysis also indicates no evidence that the sales effect is mediated by lower turnover (Appendix B). Together, this suggests that our sales effects are unlikely to be driven by our turnover effects.

5.3 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 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.3) 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.[30]

[30]Providing this helps establish trust, as Germans have strong trust toward works councils, which are

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.[31] 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.[32] 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.

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

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.

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

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

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.[33]

Contemporaneous policy changes. Another concern in any RCT is the presence of contemporaneous policy changes. However, this was not the case in our firm.

Multiple hypothesis testing. In a study addressing multiple outcomes and heterogeneous treatment effects, 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 are 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.

Covid. The RCT took place in April - December 2021. 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 (including bakery stores like ours) were excerpt from the lockdown. All stores were fully open during the RCT, including the coffee area of the store.[34] 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 firma intranet. Second, aspects of the mystery shopping score are still monitored via mystery shopping.

[33]The only way that Hawthorne effects could drive heterogeneity by regional manager expectations would be if regional managers had private information about the extent of Hawthorne effects across stores. None of the regional managers said anything about Hawthorne Effects in their explanations about why the treatment would work in particular stores.

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

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.

5.4 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.

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). As of September 2022, i.e., after 9 months, the firm has continued not having the operational checklist.

Given the heterogeneity by regional managers, one interesting question is why didn't the firm implement checklist removal only in stores where regional managers expected the treatment to work. There are two reasons against this. First, while there are sizable positive effects of checklist removal in stores where regional managers expected the treatment to work, it is not the case that there are sizable negative effects of checklist removal. Thus, while checklist removal did not yield extra returns in stores where regional managers predicted it not to work, it is not the case that such reduction proved to be 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.

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.

6 Conclusion

Scholarship often focuses on benefits of monitoring, both in general and for checklists. Working in a large German bakery chain, we use a novel methodology of surveying workers and managers about the relative value of different checklists, and we document that there is wide variation across tasks in the perceived value and time costs of monitoring. Removing two of the perceivedly lowest-value checklists improves average store performance as measured by sales and store manager attrition. The magnitudes of performance improvement are comparable to those from introducing major management practices (e.g., 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 critical caveat that some problems are rare and hard to observe. In online reviews, customers of treated stores give higher ratings, and their qualitative comments indicate improvements in speed of service and product quality. In surveys, treated workers report feeling more committed to their stores and perceive a higher level of 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 broaden economic understanding of what monitoring does in practice, suggesting 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 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.

How might our results generalize to other firms? Like all RCTs, our results are specific to our organizational context, namely, a leading firm in the German bakery chain industry. The heterogeneity of the effects we observe suggests that returns to treatments like ours may be highly 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 tasks. 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 we generally think that firms try to optimizing 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.[35]

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 direct and indirect costs of monitoring. 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.

[35] 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.

References

Acemoglu, Daron, & Pischke, Jorn-Steffen. 1998. Why Do Firms Train? Theory And Evidence. *Quarterly Journal of Economics*, **113**(1), 78–118.

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.

Bloom, Nicholas, & Van Reenen, John. 2007. Measuring and Explaining Management Practices Across Firms and Countries. *Quarterly Journal of Economics*, **122**(4), 1351–1408.

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.

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.

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

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*, **Forthcoming**.

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. 2022. Designing Information Provision Experiments. *Journal of Economic Literature*, **Forthcoming**.

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.

Hoffman, Mitchell, & Tadelis, Steven. 2021. People Management Skills, Employee Attrition,

and Manager Rewards: An Empirical Analysis. *Journal of Political Economy*, **129**(1), 243–285.

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.

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.

Lazear, Edward P., Shaw, Kathryn, & Stanton, Christopher. 2015. The Value of Bosses. *Journal of Labor Economics*, **33**(4), 823–861.

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.

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

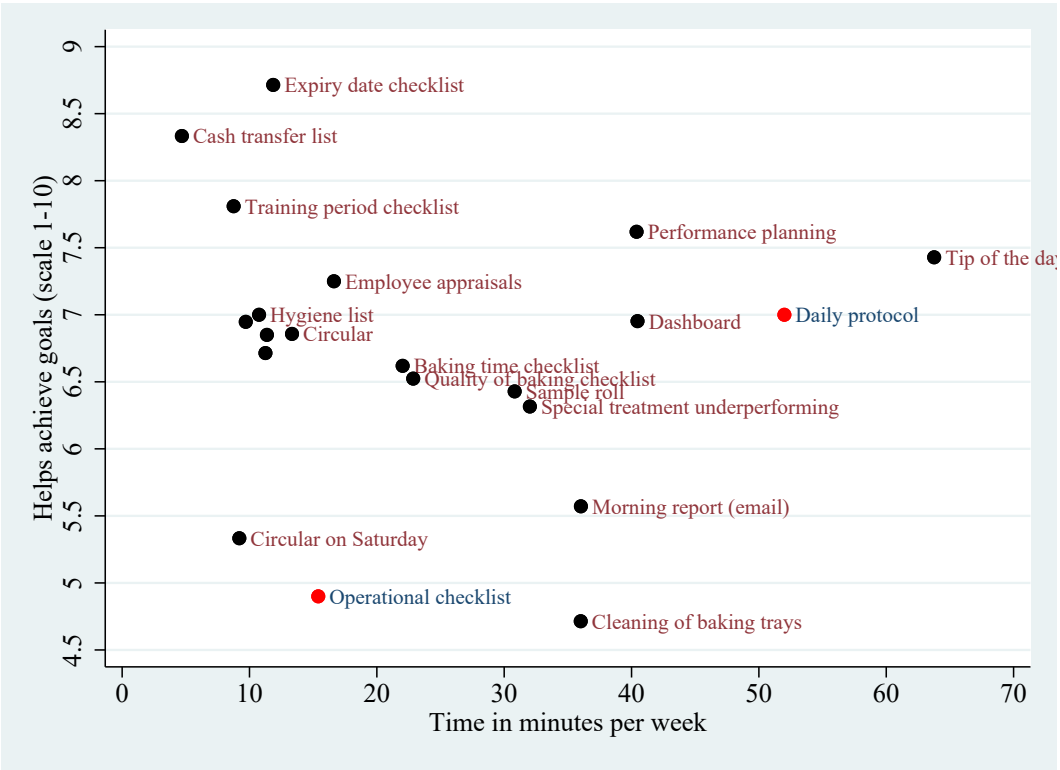
Nagin, Daniel, Rebitzer, James B., Sanders, Seth, & Taylor, Lowell J. 2002. Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment. *American Economic Review*, **92**(4), 850–873.

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*.

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

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

(a) Workers



(b) Managers

Notes: Help in obtaining goals is measured using: "The documentation duty helps (FIRM) to get better and reach company goals." Results on avoiding mistakes are shown in Figure A4.

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.

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	

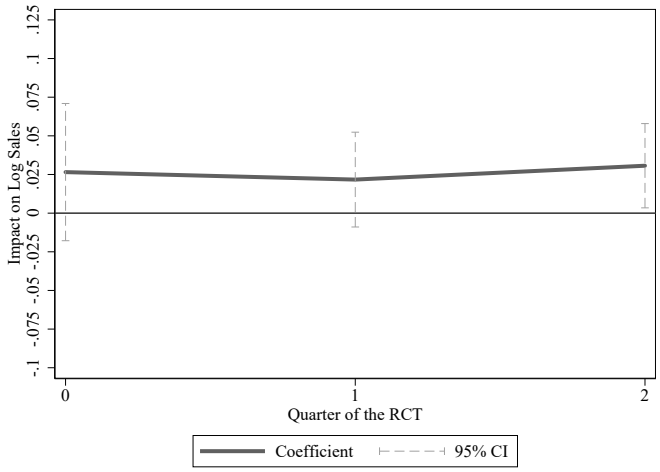
Special orders "sold out" → should we order more?

Facility or IT problems, etc.

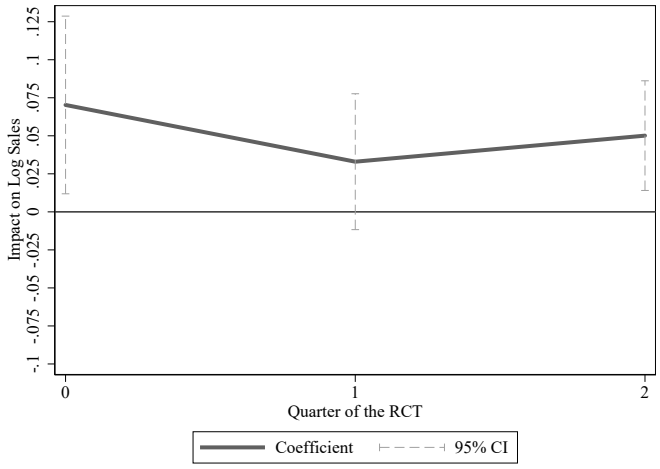
Shift changes

Additional information for the next shift

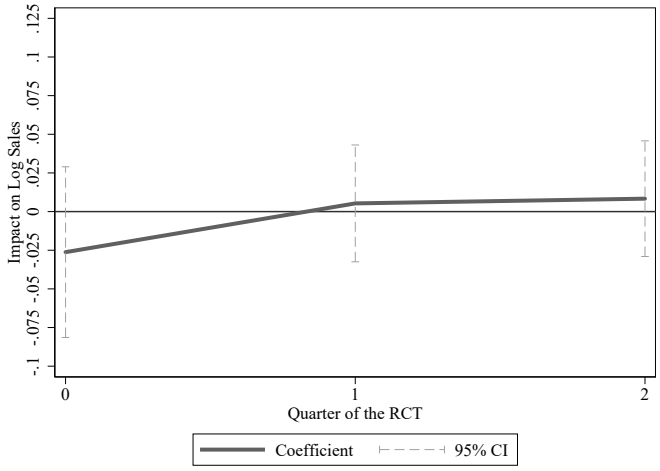
Figure 4: Treatment Effects on Sales Estimated Separately by Quarter, All Stores and Split By Regional Manager Prediction



(a) All Stores



(b) Stores Where RCT Predicted to Work by Regional Managers



(c) Stores Where RCT Not Predicted to Work by Regional Managers

Notes: This figure shows that impacts on sales do not vary significantly by quarter. Panel (a) is similar to that in column 1 of Panel A of Table 2, but we split separately by quarter of the RCT. Likewise, panels (b) and (c) here are similar to column 1 of Table 4. Quarter 0 of the RCT is April-June 2021, Quarter 1 is July-September 2021, and Quarter 2 is October 2021-January 2022.

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	-0.02	-0.03	-0.00	-0.00	-0.05	0.50	-4.98
	(0.05)	(0.05)	(0.07)	(0.05)	(0.03)	(0.08)	(0.78)	(7.24)
Constant	11.16***	10.83***	9.87***	9.85***	-2.06***	18.98***	13.30***	78.46***
	(0.04)	(0.03)	(0.05)	(0.03)	(0.02)	(0.06)	(0.50)	(4.86)
p-val	0.72	0.76	0.69	0.95	0.91	0.53	0.52	0.49

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

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	0.027*	0.026*	0.034*	0.023	0.002	0.004
	(0.015)	(0.014)	(0.019)	(0.015)	(0.016)	(0.087)
Observations	1,431	1,431	1,431	1,431	1,431	1,161
Mean DV if Treat=0	11.17	10.86	9.838	9.762	-2.099	18.93
Stores	145	145	145	145	145	144

Panel B: Worker Turnover	(1)	(2)	(3)	(4)	(5)
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 DV if Treat=0	2.038	2.806	1.254	1.159	1.647
Workers	1637	863	774	624	150
p-val trained v. untrained			0.05		
p-val manager v. non-mgrs					0.13

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. 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. 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 "p-val trained v. untrained" is a p-value from a test of whether the treatment effect is larger for trained workers compared to untrained workers. Likewise, the "p-val manager v. non-mgrs" is a p-value from a test of whether the treatment effect is larger for managers relative to non-managers.

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

Panel A: Worker Survey	(1)	(2)	(3)	(4)	(5)
Dep. var.: (all normed)	Commitment to one's store	Commitment to firm	Trust bwn. HQ & workers	Last new hire was well-trained	Basic quality control
Treatment	0.214**	-0.016	0.268*	0.005	-0.076
	(0.098)	(0.094)	(0.138)	(0.123)	(0.105)
Observations	390	390	394	368	394

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

Panel C: Google Reviews, 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	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	0.100
	(0.020)	(0.019)	(0.022)	(0.019)	(0.021)	(0.111)
Mean DV if Treat=0	11.09	10.77	9.761	9.684	-2.063	19.02
Observations	744	744	744	744	744	597
Stores	76	76	76	76	76	75
Panel B: Stores Where RCT Predicted Not to Work by Regional Mgrs						
Treatment	-0.003	-0.003	0.004	-0.006	0.024	-0.085
	(0.020)	(0.019)	(0.027)	(0.020)	(0.020)	(0.136)
Mean DV if Treat=0	11.27	10.96	9.929	9.852	-2.142	18.83
Observations	687	687	687	687	687	564
Stores	69	69	69	69	69	69
p-Val on difference between Panels A and B	0.07	0.07	0.09	0.07	0.11	0.46

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. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 5: Heterogeneity by Regional Manager Predictions: Impacts of the Treatment on Employee Attrition

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 DV 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
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 DV 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
p-Val on difference between Panels A and B	0.15	0.38	0.04	0.21	0.02

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 A3 and A4.

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

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

Notes: An observation is a store. Robust standard errors in parentheses. * 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

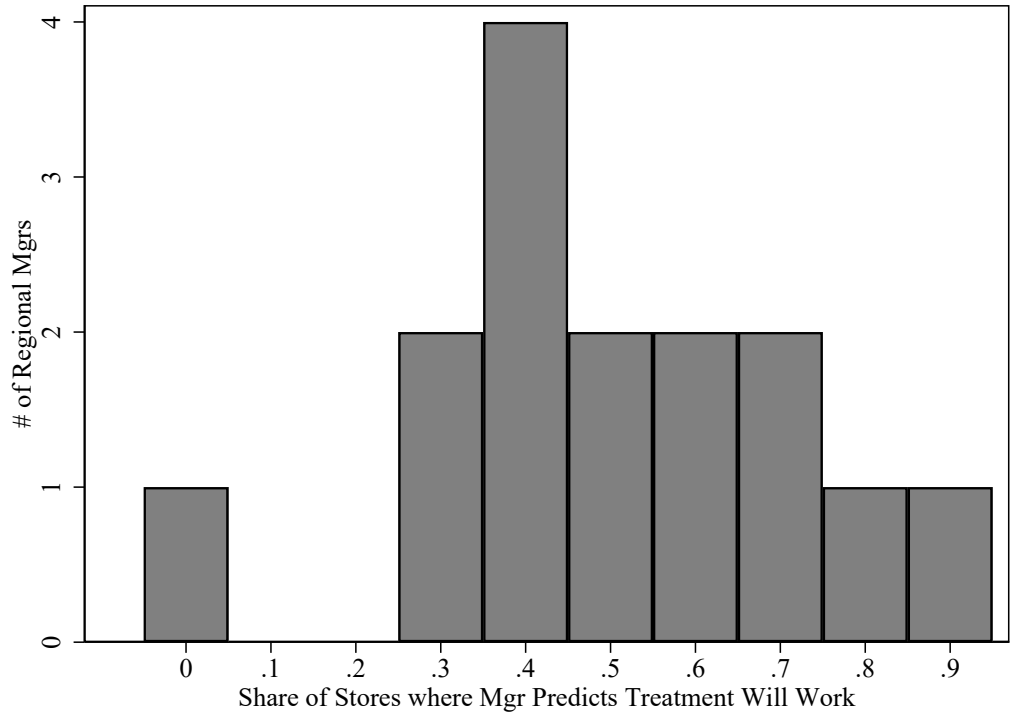
Notes: An observation is a store-month during the RCT. Standard errors clustered at the store level are in parentheses. A big team is defined as having a store head count above 10. Controls are the same as in Table 2. * significant at 10%; ** significant at 5%; *** significant at 1%

Web Appendix, "Is This Really Kneaded? Identifying and Eliminating Potentially Harmful Monitoring Practices", by Friebel, Heinz, Hoffman, Kretschmer, and Zubanov

Appendix A contains additional figures and tables. Appendix B provides additional discussion on various topics. Appendix C provides materials used by the firm in the RCT.

Appendix A Appendix Figures and Tables

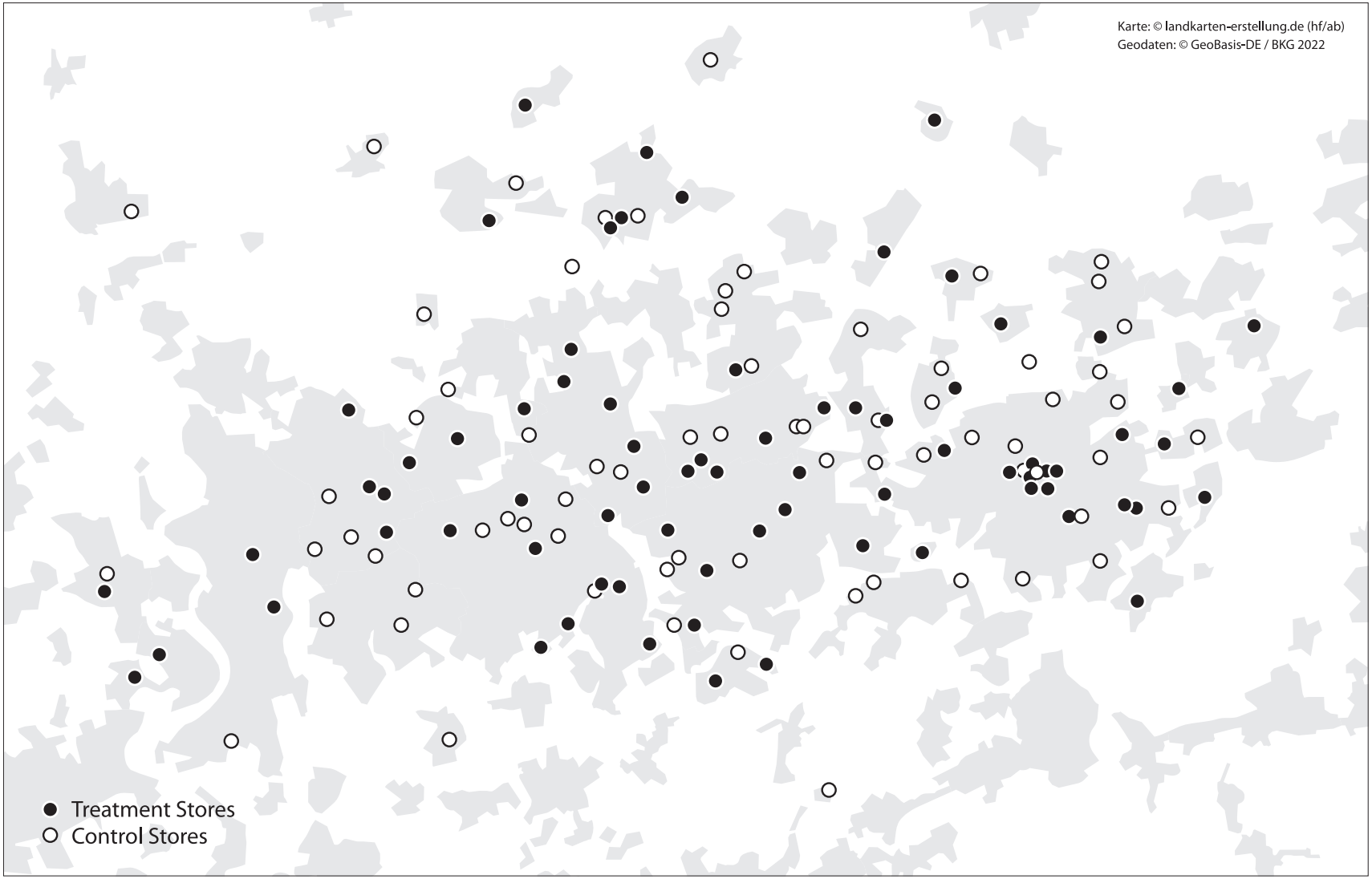
Figure A1: Variation in Manager-Level Rates of Predicting that the Treatment Will Work



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.

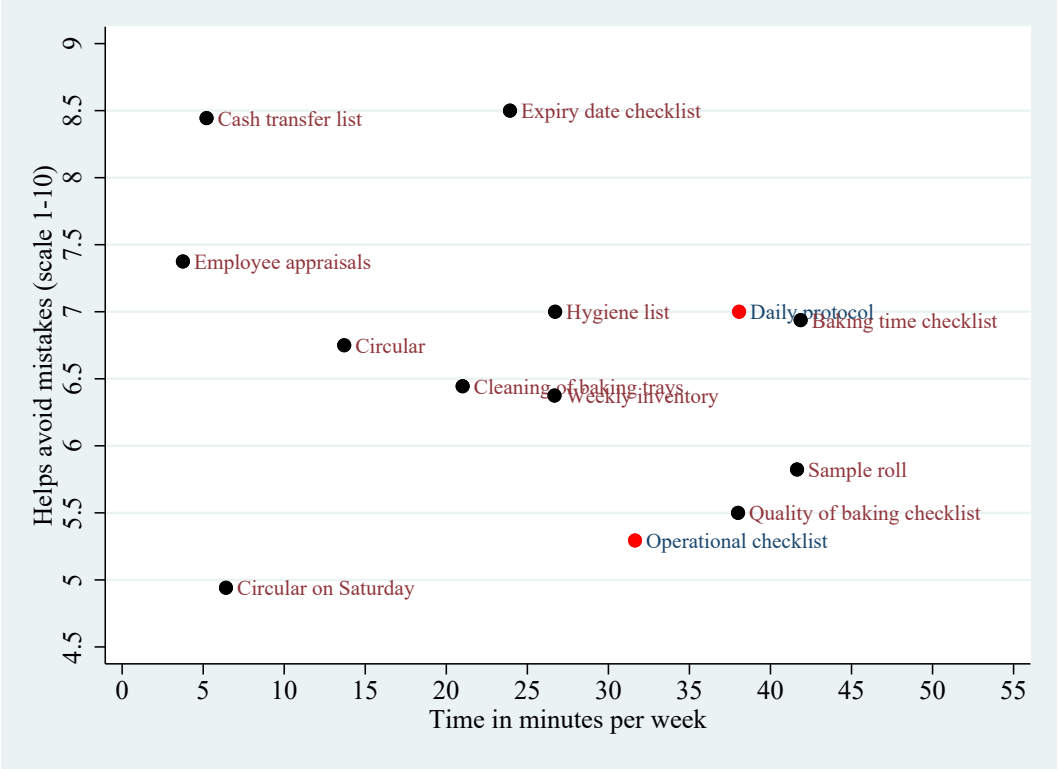
Figure A2: Picture of a Sample Bakery

Figure A3: Location of Treatment and Control Stores

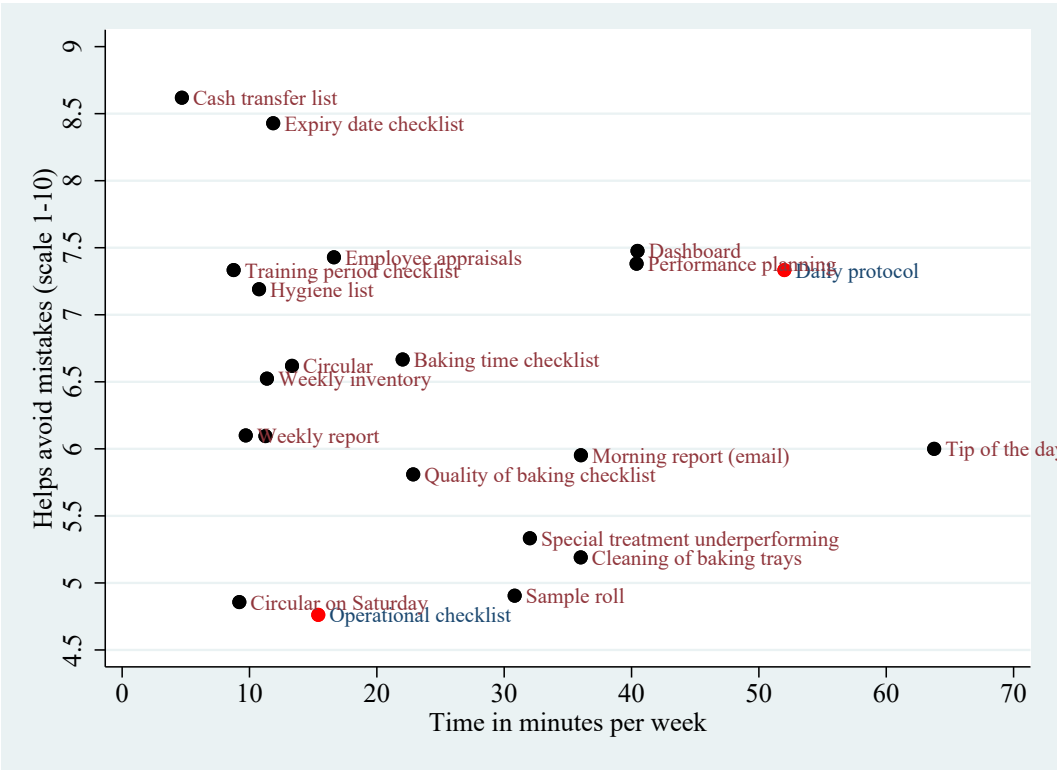


Notes: This figure shows the location of treatment and control stores on a map, with identifying information redacted.

Figure A4: Variation Across Checklists in Time per Week and Help in Avoiding Mistakes



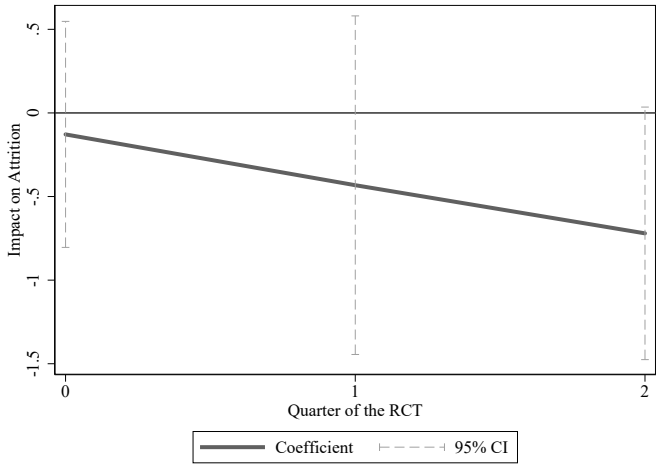
(a) Workers



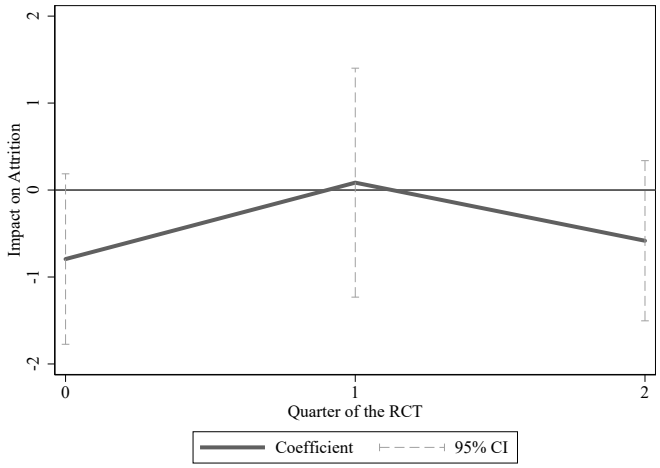
(b) Managers

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

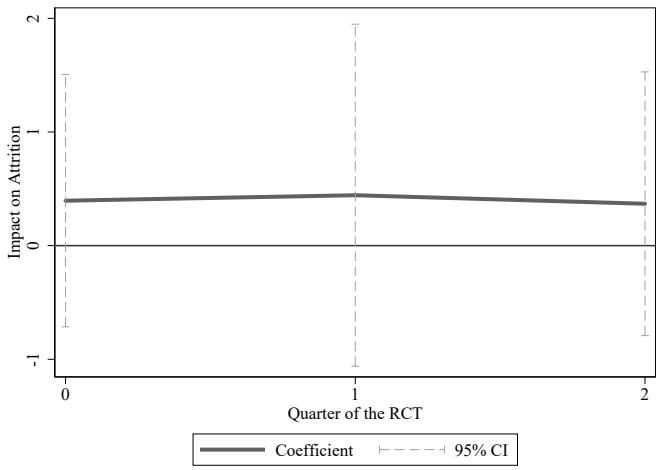
Figure A5: Treatment Effects on Trained Worker Attrition Estimated Separately by Quarter, All Stores and Split By Regional Manager Prediction



(a) All Stores



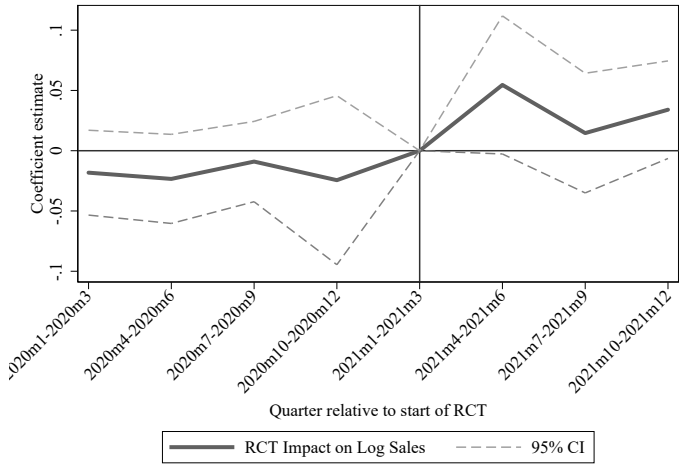
(b) Stores Where RCT Predicted to Work by Regional Managers



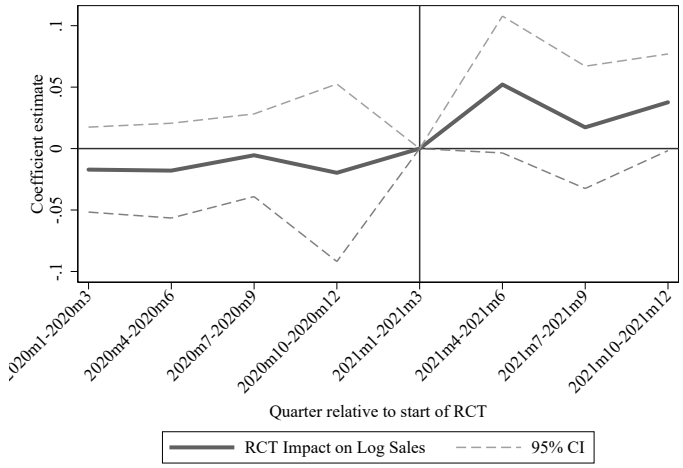
(c) Stores Where RCT Not Predicted to Work by Regional Managers

Notes: This figure shows that impacts on employee attrition of trained workers do not vary significantly by quarter. Panel (a) is similar to that in column 3 of Panel B of Table 2, but we split separately by quarter of the RCT. Likewise, panels (b) and (c) here are similar to column 3 of Table 5. Quarter 0 of the RCT is April-June 2021, Quarter 1 is July-September 2021, and Quarter 2 is October 2021-January 2022.

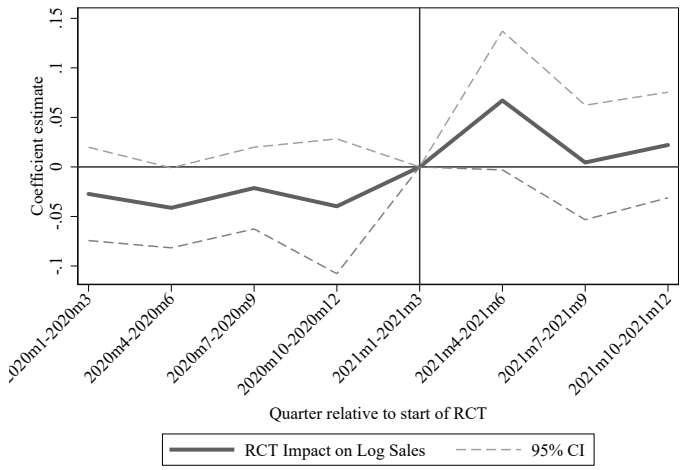
Figure A6: Event Study Impacts of the Treatment: Stores Where Treatment Expected to Have Effect



(a) Log Sales



(b) Log Busy Sales



(c) Log Slow Sales

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

Table A1: 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 A2: 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 A3: 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
0	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 A4: 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.

Appendix B Additional Discussion

Procedure for identifying list of all checklists and other documentation duties. We asked the top management to present all documentation duties. In the meeting, the sales director, who was part of the project team, presented step by step all documentation duties from the stores. He forgot one documentation duty—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 documentation duties were missing on our lit. It turned out that no documentation duties were missing.

"Missing" checklist. In the interviews in randomly selected stores, we only asked questions about 21 checklists. The "missing" checklist is called "Einverständniserklärung Sonntagsarbeit" or declaration of consent for working on Sundays. When employees have holidays on a Friday, it is not clear whether the weekend counts to the holidays or is already the start of the new working week. According to the top management of the firm, the weekend does not account towards the holidays (as long as workers do not have holidays on the upcoming Monday); according to the workers' council, the weekend is part of the holidays from the previous week. There was a big dispute between the firm and the workers' council about the case in the past. The compromise: Saturday counts as holidays and workers are not allowed to work on that day; Sunday counts as a holiday, workers are, however, allowed to work on that day and the hours are counted as overtime – however, workers have to explicitly declare that they are willing to work on the Sunday (i.e., sign the "Einverständniserklärung Sonntagsarbeit"). The top management and the workers' council also made a "work agreement" about the compromise – which is an Germany a legal binding agreement between the firm and its workers (i.e., the document is legally treated on the company level in the same way as a the German labor law). When we prepared the interviews, both the top management and the workers' council informed us that it is legally and politically impossible to drop this document. Also note that the document is rarely used (only if employees work on Sundays after their holidays) and it takes only one minute to sign it. Because of that, we did not include this checklist in our interviews.

Store league performance ranking. Inspired by the German Bundesliga, the firm uses different measures of performance to provide an overall scores to stores at the firms, and the stores are then ranked. The goal is to account for differences in possible sales and profitability.

Randomization procedure. 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.

Minijobbers. In Germany, there are workers who are allowed to work up to 12 hours per week who are known as minijobbers, 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.[1]

Framing on introduction of the treatment. In Section 2, we discuss how the framing of the treatment is not neutral. In particular, it is emphasized to workers that the firm trusts it workers, and that extra time freed up can be spent on customers and colleagues. 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.

As we discuss in the paper, 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 high unlikely that a pure framing effect could lead to very 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.[2] 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.

Use of review data. 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 reviews.

[1] For minijobbers, the treatment has no significant impact, consistent with the idea that their eventual attrition is expected.

[2] 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.

Mediation analysis. As described in Section 5.2, 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 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

ONLY FOR PILOT STORES:

Information for store managers and employees

At [FIRM NAME] 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 documentation duties (daily protocol, expiry date checklist, weekly report, etc.) at [FIRM NAME] 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[1] 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 paperwork is an opportunity, which we can use for training new colleagues and communicating with customers.

[1] 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 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.

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.