ROCKWOOL Foundation Berlin
Institute for the Economy and the Future of Work (RFBerlin)

DISCUSSION PAPER SERIES 065/26

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

Guido Friebel, Matthias Heinz, Mitchell Hoffman, Tobias Kretschmer, Nikolay Zubanov

www.rfberlin.com MARCH 2026

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

Authors

Guido Friebel, Matthias Heinz, Mitchell Hoffman, Tobias Kretschmer, Nikolay Zubanov

Reference

JEL Codes: M50

Keywords: Organizations; monitoring; checklists; respect

Recommended Citation: Guido Friebel, Matthias Heinz, Mitchell Hoffman, Tobias Kretschmer, Nikolay Zubanov (2026): Is This Really Kneaded? Identifying and Eliminating Potentially Harmful Forms of Workplace Control. RFBerlin Discussion Paper No. 065/26

Imprint

RFBerlin
ROCKWOOL Foundation Berlin –
Institute for the Economy
and the Future of Work

Gormannstrasse 22, 10119 Berlin
Tel: +49 (0) 151 143 444 67
E-mail: info@rfberlin.com
Web: www.rfberlin.com

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

Guido Friebel* Matthias Heinz† Mitchell Hoffman‡

Tobias Kretschmer§ Nick Zubanov¶

December 2025

Abstract

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

Keywords: Organizations; checklists; monitoring; respect. *JEL Code*: M50

*Goethe University of Frankfurt and CEPR and IZA and RF Berlin
†University of Cologne and CEPR and Max Planck Institute for Research on Collective Goods
‡UC Santa Barbara and NBER and CEPR and IZA
§Imperial College Business School and CEPR
¶University of Konstanz and IZA

Acknowledgments

For detailed comments, we are particularly grateful to Christian Dustmann, Paul Oyer, Andrea Prat, Jonah Rockoff, Kathryn Shaw, Lowell Taylor, John Van Reenen, and especially Wouter Dessein. We also thank numerous seminar and conference participants, including those at NBER Summer Institute Labor Studies & Personnel Economics (joint session) and NBER Organizational Economics. We thank the study firm and its management for their enthusiastic participation in this collaboration.

The RCT was pre-registered on 04/14/2021 with the AEA RCT registry under ID AEARCTR-0007550. We are grateful for the assistance of many RAs in Germany and the US for their help. IRB approval was received from the University of Cologne. The Works Council approved the project and was involved in all steps.

This paper is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation program for 2021-2027 (Grant agreement No. 101040134). Financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2126/1– 390838866) and SSHRC is also gratefully acknowledged.

Starting at least with Taylor (1919), research on workplace productivity emphasizes the importance of *workplace control*, broadly defined as firms' actions to structure and control employee behavior at work. Workplace control includes monitoring, communication, and other management practices. The degree and form of workplace control varies widely, e.g., many firms have extensive rules about how to handle situations whereas others do not.[1] Randomized controlled trials (RCTs) in field settings, detailed below, emphasize the value of employee monitoring and other forms of control in improving performance.

Firms commonly use checklists as a control tool. These are structured lists that workers fill out or work through. Checklists are celebrated as a powerful tool to help workers remember tasks and coordinate production (Gawande, 2010). Popularized in high-stakes settings such as surgery and aviation, their use has expanded far beyond these contexts to many industries, including retail, where they are widely used, including by Walmart, McDonalds, Starbucks, and many other leading firms.[2] Studies show large benefits of checklists, but little is known about whether and when checklists can be harmful, and this is for several reasons. First, as is true for many management practices (Bloom *et al.*, 2014), firms' use of checklists is highly non-random, making it difficult to estimate causal returns. Second, it seems likely that returns to monitoring through checklists are heterogeneous—some stores may benefit from extra structure, while others may find checklists useless and insulting. Better understanding this heterogeneity is critical for a full understanding of checklists. Third, modern firms often use numerous checklists, so even if one believes a workplace is "overmonitored," as often alleged by popular observers,[3] it is hard to know which checklists to modify.

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

Our partner is a major German bakery chain with 145 stores and over €100M of annual revenue. Prior to the RCT, the firm used checklists in many aspects of production. Workers needed to record extensive information, not only about products (e.g., when bread taken out of the oven), but also on customer interactions, such as whether they smiled. Drawing

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

[2] Others include Gap, Costco, Goodwill, Home Depot, and many more. Taqtics.co, citing a 2022 Deloitte report, states that 72% of high-performing retail stores utilize paper or digital checklists to increase team accountability and task accuracy. These are often daily checklists for store walks, opening/closing, and executing brand standards, but can also be weekly or monthly.

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

on a deep collaboration with the firm and top management, we conduct intensive pre-RCT interviews and surveys, and discover that several checklists are perceived as especially low-value (i.e., high time costs and limited benefits).

Our RCT is grounded in a simple conceptual framework of checklists, as laid out in Section 1. Monitoring through checklists helps firms address moral hazard problems, coordinate production, and remind workers of tasks. However, checklists also entail costs, both directly in terms of time and indirectly in terms of other factors (e.g., by reducing worker happiness or signaling mistrust). While the framework doesn't yield clear predictions about overall effects of checklists, it clarifies which stores may benefit most from checklist removal.

As detailed in Section 2, our bakery chain represents an ideal setting for our RCT. First, the sample is large. Second, we have access to granular administrative data, coupled with the ability to conduct high-quality surveys. The administrative data cover detailed aspects of sales, customers, and orders hour by hour, which is important for examining how workers and managers use time and how they substitute time on checklists for other tasks. Because of our deep collaboration with the firm's Works Council, most surveys achieve high response rates. Unusually, we survey not only store employees and managers, but also regional managers (RMs)—the supervisors of store managers—and have them predict in which stores the RCT will be effective. These beliefs are essential for analyzing treatment effect heterogeneity.

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

As has been documented for several other management practices, effects of checklist removal may be heterogeneous. Indeed, our initial discussion with RMs highlighted an informal understanding that treatment effects would likely vary considerably across stores. Thus, prior to randomization, we surveyed RMs about whether the treatment would be

effective for each store they oversee. While some dimensions of potential treatment effect heterogeneity (e.g., worker tenure) are observable, we also rely on the knowledge that RMs gather from interaction with worker teams in different stores. RMs predict that in about half of their stores the treatment would be effective, and in our RCT pre-registration, we focused on this aspect of heterogeneity. Splitting the sample based on whether RMs predicted the treatment would work, we observe vast differences in results (Section 4). Among stores where the RCT was predicted to be successful, removing checklists increases sales by 5% and substantially decreases trained worker attrition. Sales improve throughout the work day and there is an increase in customers. In contrast, in stores where the treatment was not predicted to work, the impact on both store-level outcomes and employee attrition is zero. More generally, when we ignore RM predictions and examine heterogeneity using only standard organizational characteristics like tenure and team size, we find no consistent predictors of treatment heterogeneity in sales or trained worker attrition.

To better understand heterogeneity by RM predictions, we dig into the free text of RMs' responses on why the treatment would work in particular stores. Among stores where RMs predict the treatment will work, in about one-third of cases, RMs mention something about workers enjoying the removal of checklists, consistent with a utility cost to excessive monitoring. In about two-thirds of cases, RMs mention something about the absence of problems, consistent with traditional views of monitoring to help detect and avoid problems. Heterogeneity by RM predictions appears to reflect RM information about store-level characteristics that are unobserved to the econometrician.

Section 5 estimates that the RCT substantially increases firm profits, especially among stores where the treatment is predicted to work. It also discusses threats to internal and external validity. On internal validity, we discuss spillovers across stores and the issue of our estimates being somewhat noisy overall. The latter point may reflect that we study a management practice with highly heterogeneous returns. Among stores where the treatment is predicted to work, estimates are quite precise. On external validity, we conduct an informal survey of German bakery chains to support that checklists are widespread in the industry.

The firm was highly satisfied with the results of the RCT, as discussed in Section 6. Unlike past interventions in the literature, our RCT subtracted instead of adding something, so the direct cost to implement the RCT was minimal. With estimated benefits of checklist removal roughly 60 times larger than costs in the RCT, the firm decided to make firmwide changes in checklists, removing the operational checklist from all stores. Interestingly, while the RCT removed two checklists, the firm restored the daily protocol in the firmwide rollout.

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.[4] 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, which have been previously detected by lab studies. Our RCT suggests that such insights extend into the field as well, and we offer a framework that rationalizes why monitoring may be good for some tasks, but bad for others.

In economics RCTs on monitoring, closest to ours is Nagin *et al.* (2002), who consider a field experiment where a call-center company exogenously varies its monitoring rate in some 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,[5] 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). 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 are Bandiera *et al.* (2021) and Belot & Schröder (2016). In an RCT in Pakistan where authority is transferred to tax collectors from their monitors, Bandiera *et al.* (2021) show that delegating to tax collectors boosts performance. Instead of changing authority, our study changes monitoring holding authority fixed. Hiring students to identify

[4]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.* (2024). With the exception of Nagin *et al.* (2002), in these studies, monitoring is added instead of removed or reduced. Many observational studies also document benefits to monitoring, especially in trucking (Hubbard, 2000, 2003), though some share of truckers may respond negatively (de Rochambeau, 2022). In contrast, some lab experiments and lab-in-the-field experiments point to potential overmonitoring (Dickinson & Villeval, 2008; Falk & Kosfeld, 2006; Herz & Zihlmann, 2022). As surveyed by Ravid *et al.* (2023), there is also a large psychology literature on monitoring which focuses on correlations of monitoring with survey outcomes.

[5]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.

coins, Belot & Schröder (2016) show that randomly added monitoring can backfire on some dimensions of performance. More broadly related is work by Ash & MacLeod (2015) and Coviello *et al.* (2014) who show that putting constraints on how people (in their case, judges) work can backfire if they are more motivated to do their work.

Second, it provides a clear example of a successful large-scale intervention that is "subtractive," i.e., involves taking something away. Psychologists argue that individuals and firms systematically overlook subtractive interventions (Adams *et al.*, 2021), perhaps because humans are hard-wired to look for new things (Klotz, 2021) or because it is easier to take credit for an addition. We are not familiar with prior economics RCTs that improve outcomes via subtraction. That our RCT is subtractive is substantively important, as it makes our RCT's cost quite low. Several additive management practices also yield performance improvements broadly similar to ours, but our RCT's benefit to cost ratio of about 60 is the largest that we are aware of in the management practice literature.

Third, our paper contributes to work in personnel and organizational economics on the heterogeneous returns to management practices and on the impact of managers. Amid substantial work on the importance of management practices in general (Bloom & Van Reenen, 2011; Bloom *et al.*, 2012, 2019), growing research emphasizes that management practices are complementary to one another (Ichniowski *et al.*, 1997; Dessein & Prat, 2022; Guadalupe *et al.*, 2024), and that their impact may be contingent on other within-firm factors (Blader *et al.*, 2020). We show that there is substantial heterogeneity in the return to a management practice, namely, checklists, based on RM beliefs. A growing literature examines what non-CEO managers do and know (Benson & Shaw, 2025). Our results support that RMs have information about how the RCT will affect particular stores, consistent with theories of dispersed information in firms (Dessein, 2002; Dessein & Santos, 2006). Having such information could be a way that managers add value separate from other channels such as motivation and teaching.[6]

Fourth, our paper makes a methodological contribution to RCTs. Beginning with DellaVigna & Pope (2018), work uses expert predictions to examine how the results of an RCT compare to priors of experts, i.e., to see to what extent a result is surprising. Rather than having outside experts predict the average results of the RCT (e.g., that the treatment will affect sales by a certain amount), our RCT has insider experts predict store by store

[6]RMs in our setting lack discretion over checklist use, so they cannot act on their information by tailoring practices to specific stores. However, our estimates suggest that the firm would not benefit from differentiating checklists by RM beliefs, as treatment effects are close to zero—rather than negative—in stores where RMs do not predict the treatment to work. Our contribution is to show that managers hold relevant information—thus, empirically grounding theories emphasizing informational roles of managers (Dessein, 2002; Dessein & Santos, 2006)—rather than analyzing mechanisms by which managers matter (Dessein & Santos, 2021).

whether the treatment will be effective in that particular store. We are aware of very limited prior work that uses expert predictions in RCTs in such a manner, but we believe this methodology may be useful in other contexts, especially in organizational economics.[7] We see two main advantages of incorporating upper manager beliefs into analysis of treatment heterogeneity. First, and foremost, it reveals whether upper managers have information about who will benefit most from a treatment. Second, analyzing text of beliefs can shed light on theory-relevant aspects of managerial predictions (e.g., frequency of problems, aversion to control). By asking insiders to make formal predictions about an RCT, our methodology complements earlier work in "insider econometrics" (Shaw, 2009; Ichniowski & Shaw, 2012) that rigorously uses administrative and survey data from inside of firms to study productivity.

1 Conceptual Framework

We suggest a framework about how removing checklists could affect store performance and worker attrition. Beyond average effects, the framework models across-store treatment heterogeneity according to RM predictions, and it provides intuition about which types of checklists are more beneficial to remove. Checklists are randomized in our RCT, so we focus on the impact of checklists and do not consider the decision to adopt them. For simplicity, the framework has a binary comparison of having checklists versus not, rather than removing some checklists and keeping others (as we do in the RCT). The unit considered is a store.

Problems. As in Garicano (2000), firm stores face *problems*. Thinking of problems broadly, these include memory problems—such as a surgery team forgetting steps (Gawande, 2010) or bakery workers forgetting the right amount of sugar—and coordination problems, such as failing to inform the next shift when the bread was made, which is important for keeping it fresh. Stores may differ not only in the frequency of problems but also in ability to solve them (Garicano, 2000; Garicano & Rossi-Hansberg, 2006, 2015). Let p be the probability that a problem occurs in a store and remains unsolved in the absence of checklists. We assume problems occur exogenously, though the logic of the framework can be extended to incorporate worker moral hazard as in Nagin *et al.* (2002). When a problem occurs, the cost to the firm is k. Thus, without checklists, store expected profits are $-pk$.

Benefits. Checklists help stores identify and solve problems.[8] The impact of checklists

[7] Broadly relatedly, Dal Bó *et al.* (2021) ask supervisors of government agricultural workers to rank which workers should get free phones, and Bryan *et al.* (2024) have loan officers predict how microfinance clients will fare under treatments. We differ by studying predictions of higher-up insiders in a private-sector firm.

[8] We think of checklists as highly structured forms of documentation. These include, of course, lists that workers check off. Checklists also include documentation where workers enter simple information in a structured fashion, e.g., cash amounts or IT issues. Checklists may be paper or digital, and done individually

on problem-solving is m, the probability that a problem is detected and solved in full.

Costs. Using checklists involves direct cost, c, which can include material costs, but in our setting is primarily the time cost of filling out checklists. In addition, using checklists entails an indirect cost θ to store performance. Many people dislike being monitored and controlled, perhaps because it is intrinsically unpleasant to fill out checklists, but also because monitoring and control can be viewed as a sign of disrespect (Ellingsen & Johannesson, 2007, 2008) or mistrust. Ellingsen & Johannesson (2008) model workplace respect in terms of second-order beliefs, i.e., a worker's belief about the firm's belief about whether she is altruistic or competent. Being respected can be important for store performance, both because it may reduce worker attrition and may motivate workers to work harder (Friebel *et al.*, 2023). Control could also crowd out intrinsic motivation to work hard (Benabou & Tirole, 2003; Rebitzer & Taylor, 2011; Ash & MacLeod, 2015).

Profits. Under checklists, profits in a store are $-(1-m)pk-c-\theta$. Thus, the returns from our treatment of removing checklists are $c+\theta-mpk$. This expression allows us to predict when the treatment will be beneficial for the firm, as well as to predict what are the stores and checklists where checklist removal will have the largest benefit. Our treatment is likely to be beneficial when direct and indirect costs of checklists are high (higher $c+\theta$); when problems are infrequent (lower p); when checklists detect problems less reliably (lower m); and when problem costs are small (lower k). In practice, variation across stores seems likely to be greatest in problem frequency p and in the indirect costs of checklists θ.[9]

Context dependence of θ. It is natural that θ could be related to p and k, i.e., the use of checklists may feel more onerous or disrespectful when they serve less of a purpose, such as when there are few problems to solve or when the cost of problems is small.[10]

Heterogeneity by RM predictions. This framework also raises the possibility that there could be heterogeneity across stores within a firm in the return to checklist removal. RMs may know that some stores experience coordination problems more frequently; stores may also vary in outcomes if some workers dislike checklists more than others (e.g., if some workers find them more disrespectful or wasteful than others), and store employees may differentially complain about these costs to RMs. Given there are multiple factors affecting whether monitoring is beneficial and that some factors (like frequency of coordination problems) are hard to observe in data, it is natural to ask RMs to make predictions about

or in groups (e.g., surgical teams).

[9]There could also be heterogeneity in m and k, e.g., some stores have greater costs when problems arise. Given the multiplicative term mpk, heterogeneity in p has the same effect in the model as that in m or k.

[10]Pilots may accept checklists because problem costs are high. In retail, where costs are lower, checklists may feel burdensome. θ depending on p and k can rationalize stores with fewer problems feeling controlled.

whether a treatment will work in a store.

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

Attrition. Our framework focuses on store performance, in line with our RCT pre-registration, but is easily extended to cover worker attrition. It is natural that the direct and indirect costs of control through checklists is not only reflected in performance, but also in worker utility from the job. Our treatment is likely to reduce attrition most for workers with higher personal costs of checklists, but could increase attrition for workers who benefit from checklists' added structure. Dube *et al.* (2022) find that workers care deeply about being respected and provide evidence that this matters for turnover. It is natural that more qualified workers would experience more positive effects on attrition of checklist removal.

Summing up. Theory does not make clear predictions about the *overall* effect of removing checklists, as there are competing costs and benefits. However, if RMs have private signals about returns to removal, the framework suggests two patterns that we test. First, if "will work" is understood in the usual way—meaning RMs use a non-negative prediction threshold ($z^* \geq 0$)—then effects should be positive in stores where RMs predict the treatment will work. Second, effects should be larger in stores where the treatment is predicted to work than in those where it is predicted not to work.

2 Study Background

Our study firm is one of the largest bakery chains in one densely populated region of Germany (in Germany, most bakery chains operate in particular regions). Like most bakery chains, the firm is family-owned. Many of the top executives, including the CEO, have been with the firm for decades and helped guide its growth. The firm has roughly 1,350 regular employees,

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

who work a median of 35 hours per week.[12] 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 chain.[13] The firm has a reputation for quality products, as evident, among other places, in online reviews.

Hierarchy. Most employees work in the stores with an average of about 10 regular employees per store (including the store manager). There is one store manager per store. Store managers are often experienced workers who are promoted from within. Above store managers are the 15 RMs, who each manage about 10 stores. The 15 RMs are supervised by three sales directors, who in turn report to the firm's top executives, whom we also refer to as the management team or top management. "Headquarters" refers to the central office where top management and firm-wide centralized functions are based.

Job tasks. 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. For store managers, their key role is to supervise and motivate their workers, as is common in many lower-skilled jobs (Benson & Shaw, 2025).[14]

Pay and control systems. Store employees are paid by the hour. Pay is slightly above minimum wage, as is common for low skill German jobs (Dustmann *et al.*, 2009). Besides hourly pay, only a minority of workers get performance pay and bonuses are low, as is common in German bakeries.[15] Employees have regular reviews and can get promoted to higher positions. The firm's culture is control-oriented. Detailed instructions, checklists, and regular top-down communication are used to ensure quality standards. Workers are also monitored by store managers and mystery shoppers (undercover shoppers who visit each store roughly each month). 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.

[12] The firm also uses about 500 "minijobbers" who work 7-8 hrs/wk (Appendix B.6). As they aren't regular employees, minijobbers are excluded from attrition analysis, but results are similar when including them.

[13] A typical bakery store is in the same building as a grocery store but outside its layout, with a separate entrance and different hours (e.g., groceries close Sundays in Germany, but bakeries remain open).

[14] Each store has a personnel budget that is decided centrally. Job ads are done centrally. RMs may give store managers more or less authority over whom to hire, but RMs represent the firm legally.

[15] In the month before the RCT, bonuses represented $< 2\%$ of total pay: 1.3% for workers and 4.5% for managers. Just 35% of workers received any bonus vs. 86% of managers. Bonuses depend on individual and store performance, but large ones were rare: 99% of both groups received less than 10% of pay in bonuses.

Why the firm did the RCT. Our collaboration with the firm started in 2020. In exploratory discussions, two signs indicated concerns about overcontrol and overmonitoring. First and foremost, we came across a 2018 employee survey which indicated broad dissatisfaction with firm checklists. Second, the head of HR was concerned about employee turnover, especially of trained workers, and separately expressed concern about overmonitoring (via feedback from the works council), and we thought the two could be linked. Trained workers are trained via apprenticeships paid in part by firms (Dustmann & Schoenberg, 2012). To jointly explore these two observations in greater detail, 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 *pre-RCT in-depth interviews*. The project team began by creating a comprehensive list of all 22 in-store checklists. This committee-based approach broadly follows idea-generation and process-optimization practices used in large firms like Toyota (Womack *et al.*, 2007). RAs randomly selected 22 shops to conduct in-depth interviews on checklists. Given the control-oriented culture, we were concerned there'd be issues of trust, so we opted for in-person instead of online surveys. To build trust, RAs were driven to stores by the head of the works council, who introduced them and emphasized they could be trusted. For each store, RAs generally asked to speak to the store manager and the first worker they came across. In one store, the store manager was not available, and in 4 stores, no worker besides the manager could be asked to do an interview. We thus interviewed 21 store managers and 18 quasi-randomly selected workers. No one asked for an interview refused (100% response rate), likely due to the works council's involvement. For 21 of the checklists, respondents were asked:[16]

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

[16] One of the 22 checklists—consent for working on Sundays—was omitted from interviews, as the works council stated it was legally mandatory and couldn't be dropped.

4. How many minutes do you spend each time filling out the checklist?

Figure 1 gives results on the *in-depth interviews*, focusing on the value in helping the firm achieve its goals (Q1) and the weekly time cost (Q3 and Q4 combined). In the worker interviews, five checklists stand out for having relatively low value and high time cost, indicated by location in the lower-right of panel (a). Three of these, all on baking, were considered "sacred cows" and, thus, impossible to remove, either for political reasons or because they relate to the firm's unique selling proposition.[17] The two remaining duties were the **operational checklist** (*Operative Liste*) and the **daily protocol** (*Tagesprotokoll*). The daily protocol and especially the operational checklist also score poorly on avoiding mistakes (Figure A3).

According to self-reports, workers spend an average of 319 minutes (about 5.5 hours) per week on all the checklists, while store managers report an average of 499 minutes (over 8 hours). This amounts to roughly 15-20% of their weekly work time. While employees may overstate these figures, this is not a concern for our study.[18]

In a meeting in October 2020, the researchers presented analyses on these interviews 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 that monitoring may signal disrespect. Thus, executives had pre-RCT debates which resembled tradeoffs in our conceptual framework.[19]

Operational checklist. The operational checklist is a detailed form where workers affirm completion of specific tasks (e.g., I touched the Berliner doughnuts correctly, I made eye contact and smiled). As seen in Figure 2, which shows the checklist from shortly before the RCT, it reminds workers about how they are supposed to do their jobs. In our initial focus groups, many workers view the list as somewhat insulting. Employees must sign each item of the checklist every day. Employees do the checklist at different times of day.

[17] These three are the sample roll, baking time, and baking quality checklists. For instance, the sample roll checklist requires stores to send five rolls from each batch they bake to headquarters for testing. Since bread rolls are central to the firm's unique selling proposition, these checklists couldn't be removed.

[18] Whether checklists use 15-20% of work time or a fraction of this, our survey shows *substantial* time on checklists. Our focus is not to measure time on checklists, but to assess tradeoffs in removing two of them.

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

Most items on the checklist are updated each month. Thus, employees spend some time reading it each day to know what they are signing. Workers spend an average of 32 minutes per week on the checklist (p25 = 14m, p50 = 24m, p75 = 35m). Store managers spend less (mean = 15m; p25 = 6m, p50 = 7m, p75 = 20m). Appendix C.6 provides two examples of older versions, one from Aug. 2019 and one from Jan. 2017.[20]

Stores receive similar information to what's in the checklist via a weekly newsletter (e.g., about correct placement of Berliner doughnuts). That is, employees are constantly reminded how to do their jobs—first in the newsletter, then again via the operational checklist, which requires signatures. However, prior to the RCT, some executives thought that without the checklist, stores would experience operational problems and that some workers would not follow company guidelines (e.g., not smile at customers).

Daily protocol. The second checklist we study is the daily protocol, where employees write down several key things that happen during the day. As seen in Figure 3, this includes how much money is in each cash register, items sold out, IT problems, and information to pass along to the next shift. In contrast to the operational checklist, some employees find more value in the daily protocol. Workers spend a mean of 38 minutes per week on the daily protocol (p25 = 14m, p50 = 18m, p75 = 70m). Store managers tend to spend even more time, a mean of 52 minutes per week (p25 = 35m, p50 = 35m, p75 = 70m), reflecting that it is often done by store managers. Unlike the operational checklist, the daily protocol does not change over time, but it still requires significant time to provide the required information. Employees do the daily protocol at end of shift.

How checklists are used and why employees take them seriously. The completed operational checklist and daily protocol forms are only infrequently examined by firm headquarters. However, employees can be held responsible and even fired if they don't complete a checklist or fill out a checklist falsely, because by German labor law, they must follow workplace instructions. Employees are, thus, motivated to take the checklists seriously.

RCT setup with RMs. RMs and sales directors were invited to a meeting on Feb. 16, 2021 with top executives and the research team. RMs were informed there would be a 6-month RCT and were given detailed guidelines about it. They were also given the chance to ask questions. In the meeting, several RMs spontaneously expressed strong views on the stores in which the treatment would be effective. This suggested possible heterogeneous treatment effects and that RMs may have strong knowledge on this heterogeneity.

[20] Comparing the old checklists with Figure 2 shows: (i) the format changes over time, so workers can't simply breeze through without reading; (ii) the content remains broadly similar; and (iii) the Dec. 2020 version is longer. The expanding checklist may help explain worker frustration and may reflect a "management by exception" approach—adding items when new problems arise.

RM predictions: ***The pre-RCT survey of RMs.*** In March 2021, before knowing which stores were in control or treatment, RMs made predictions by phone about in which stores the treatment would be effective. We used phone interviews because RMs tend to be on the road visiting stores they oversee. RMs are used to speaking by phone and all 15 were willing to talk to us (100% response rate). To show respect and elicit serious responses, interviews were conducted by a coauthor (a chaired German professor) rather than an RA. No incentives were used for predictions because they are subjective.

We motivated phone calls to RMs by reminding them that there was significant heterogeneity in RMs' informal predictions for whether the treatment would work during the Feb. 2021 meeting. To make predictions as natural as possible, we asked RMs for verbal responses, which we later convert into a binary response of whether it will work. Due to German privacy norms, the coauthor wrote down RMs' predictions by hand instead of recording calls. RMs gave responses of different styles, with some simply stating their belief about whether the treatment would work and others giving explanations. The classification of responses was done very shortly after all the calls by the coauthor conducting the interviews who also translated the responses to English. For almost all responses, the conversion to binary responses is clear and unambiguous.[21] Predictions are only counted as positive if the RM says unambiguously that the overall effect or effect on store operations will be positive. Appendix C.1.1 has the exact wording of what was told to RMs and the one question asked. Appendix B.4 gives further details on RM predictions. Figure A1 shows variation across RMs in the share of stores where they predict the treatment will work.

Why one treatment? We use a single treatment—removing two checklists—for three reasons. First, our 2020 pre-RCT survey identified two checklists that were both low-value to workers and politically feasible to remove, making it natural and managerially relevant to drop both. Second, pre-RCT power calculations showed we were well-powered to detect a 3% effect with one treatment but might be under-powered with multiple treatments. Third, we expected and pre-registered substantial treatment heterogeneity, which we would be underpowered to detect if treatment were split across multiple arms.

RCT setup with store managers and workers. Store managers and workers in treated stores were informed via the firm's weekly newsletter—both through a message on the store intranet on Tuesday, April 6, 2021 (after Easter) and in paper form with the weekly bundle of checklists. In contrast to RMs, workers and store managers were not informed that there was an RCT or that the change would last for a certain period of time. Given that the checklists were a routine part of the job, their removal was clearly noticed by employees.

[21] To validate coding accuracy, a second German-speaking coauthor independently translated and classified all RM predictions. Agreement was nearly perfect, with one difference between classifiers (Appendix B.4).

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

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

The message emphasizes two factors, paralleling our Section 1 discussion on direct and indirect effects. First, it emphasizes how the firm trusts workers (indirect effect). Second, it emphasizes the extra time (direct effect), and that workers should use the extra time for customers and colleagues. While one could worry that workers are being "led" to think a certain way, it would be artificial for a firm to make a change like removing 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 RM expectations.

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

We ensured the RCT was implemented as planned. Checklists are delivered to stores weekly in a physical bundle. An RA verified that bundles for treatment stores excluded the operational checklist and daily protocol, while control stores received them. In May 2021, we confirmed with RMs, the head of HR, and a sales director that the treatment was implemented as intended, with no reported issues.

RCT timing. The RCT began April 6, 2021. Checklists were removed in treatment stores. Results were presented to the firm in Dec. 2021. Given the RCT's success, the firm rolled out a modified version of the treatment to all stores in late Jan. 2022. In this rollout, all stores operated with the operational checklist removed while the daily protocol was reintroduced—some workers found it useful and less onerous. We registered the RCT on

the AEA Registry on 4/14/21. Our analyses closely follow the pre-registration (Appendix B.7). We registered the RCT to last for 6 months, but logistical delays extended it to 10.[22]

Administrative data. We use administrative data from the firm to create two main panel datasets. First, using hourly sales data, we create a store-month panel of store outcomes, including overall sales, sales at different points of the day, and number of customers (i.e., number of unique transactions). The panel also includes monthly store shrinkage (a joint measure of food rot and employee theft), as well as mystery shopping stores, which are measured in each store in most months by a mystery shopper. Second, we create a worker-month panel of regular employees regarding worker attrition. Details are in Appendix B.6.

Surveys and Google reviews. Besides the above-discussed (1) *pre-RCT in-depth interviews* regarding time spent on checklists and (2) *pre-RCT survey of RMs* (conducted to get beliefs on which stores would benefit from the treatment), several other surveys were conducted within the partner firm. Figure A2 summarizes these surveys. In all the surveys, RAs conducting the survey were never aware of which stores were treated.

- *Pre-RCT store manager survey:* Conducted by RAs via phone in March 2021, with a 96% response rate. This gives us information on time spent on the daily protocol and when it was completed.
- *During-RCT store manager survey:* Conducted by RAs via phone in Nov. 2021, with a 94% response rate. This gives us information on RM visit frequency and duration; perceived value of the treatment; informal checklist replacements; and whether control stores had heard about the treatment.
- *During-RCT worker survey:* Conducted in-store in Oct. 2021 using pen and paper, it gives us data on worker attitudes. RAs distributed and collected questionnaires during store visits. The response rate was 35%—lower than in manager surveys but typical for employee surveys (Appendix B.9). Appendix B.9 also shows that the treatment doesn't affect response, supporting that non-response doesn't bias the results.

We also collect data on Google reviews, which we discuss when analyzed in Section 3.

Randomization. Using "randtreat" in Stata, we use a stratified randomization with 4 stratification dimensions: pre-RCT head count (above or below mean), pre-RCT sales (above or below mean), pre-RCT store ranking in the firm performance league (above or below mean; described in Appendix B.3), and region (9 regions). This gives 46 strata. Appendix B.5 further justifies our randomization procedure. Table 1 shows strong balance on observables. A typical worker is female, about 40 in age, and has hourly base pay of €12.

[22]Parental leave by a coauthor and a vacation by a key firm contact delayed the endline survey by 4 months, so the RCT lasted 10 months. This fortuitous extension was clearly unrelated to statistical power.

3 Overall Results

To estimate the impact of the treatment on store-level outcomes, we use ANCOVA specifications following McKenzie (2012). Using data from the RCT period, we estimate OLS models focusing on T_s, a dummy for treatment in store s, and 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 the 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. Following Young (2019), we also perform randomization inference ("RandInf"). The resulting p-values in square brackets are generally extremely similar to those from conventional clustering-by-store inference. Furthermore, there are no estimates in the main text which are statistically significant under conventional clustering, but insignificant under RandInf. 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, which is the main pre-registered store outcome. Sales go up by 2.7%, statistically significant at the 10% level ($p = 0.07$ under both conventional clustering and RandInf). The number of customers increases by 2.3%, narrowly missing statistical significance, suggesting that more customers are coming through the door instead of solely upselling more.

How do effects on sales vary by time of day? As seen in Figure 4, which plots sales effects separately by hour of the day, effects are relatively constant over the day. Returning to Panel A of Table 2, sales increase significantly during the busier part of the day for bakeries (7am to 2pm) and in the less-busy time segment (not in 7am-2pm).[24]

[23] All our findings are unchanged to doing simple ANCOVA where we don't control for variables used in stratification (Table A1). Our conclusions are also robust to controlling for strata dummies (Table A2), though results are generally more imprecise, which we believe occurs because we have a high ratio of strata to stores (Bruhn & McKenzie, 2009), including 14 singleton strata. Including dummies for singleton strata is akin to excluding them from analysis. Our findings are also unchanged when covariates are selected using post-double selection LASSO (Table A3). Our approach of not controlling for strata dummies for a reason is consistent with Bruhn & McKenzie (2009) (p. 219), and follows recent econometrics work indicating that controlling for strata dummies can yield overly conservative inference (Cytrynbaum, 2024; Bai *et al.*, 2024).

[24] That effects are similar for busy and slow periods is broadly consistent with employees "working better" instead of "working faster," e.g., if the treatment makes employees feel more trusted, and this increases the chance that customers return to the store in the future. It also suggests about our firm's production function that store revenue is limited not just by speed of service, but also by other factors like whether customers come back. We also do not observe the length of the line at stores, and we do not know whether the length of the line and speed of service is actually worse in busy periods compared to slow ones. While we cannot reject that effects are the same for slow versus busy periods, we also cannot reject small to moderate differences in

While sales increase, a concern with removing checklists is that it could lead to drawbacks like an increase in employee misbehavior or a decrease in product quality. We see no evidence for this concern. Shrinkage and the mystery shopping score are both unchanged. With 95% confidence, we reject that shrinkage increases by more than 0.033 log points and we reject that mystery shopping score decreases by more than 0.13 standard deviations (σ), i.e., we can rule out moderate-sized deteriorations. Mystery shoppers are unaware there is an RCT, so they do not know which stores are treated.

In addition to the overall mystery shopping score, we also analyze individual components of mystery shopping. This helps illuminate whether there are operational aspects that suffer from our treatment. As seen in Panel A of Table A4, we see no evidence of harm on any component scored by mystery shoppers. This is true across simple process outcomes, like whether employees show their name badge and engage in upselling, but also in terms of subjective quality outcomes, like baked roll quality and friendliness of customer interactions.

Besides estimating overall effects for the full RCT, it is useful to show effects over calendar time. Panel (a) of Figure 5 estimates Equation (1) separately for each 5-month period, a natural division given the RCT lasts 10 months. Effects are similar across both periods: even 6–10 months after checklists are removed, sales increase by 3%, statistically significant at the 5% level. The stability of effects also appears when dividing the RCT into quarters (panel (a) of Figure A4) and when plotting treatment-control differences over time using two separate lines (panel (b) of Figure A4). Figure 5 also shows how stores that were treatment and control in the RCT differ in the post-RCT rollout, discussed in Section 6.

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

The treatment has no overall effect on attrition (column 1). However, this masks heterogeneity by skill. A critical distinction is between trained and untrained employees. Trained workers already did a 3-year apprenticeship, often from the bakery firm. The firm is keen to retain these workers receiving expensive training. In contrast, untrained workers have fewer skills and are less important to retain.[25] Trained worker attrition decreases by 0.44 percentage points (hereafter, "pp") per month, which is a 35% decrease relative to control. However, untrained worker attrition increases by 0.64pp per month, an increase of 20%. The attrition difference by worker training is statistically significant ($p = 0.02$ under

favor of busy periods, so the result should not be overinterpreted.

[25]Trained workers are central to the firm's quality strategy. They have much longer tenure (median 10.7 vs. 2.9 years) and 33% higher base pay than untrained workers in the month before the RCT (Table A6).

both conventional clustering and RandInf). Untrained workers know less and may benefit from added structure. For trained workers, this structure may be unnecessary and unwanted.

Within trained workers, some are store managers and some are not, and effects are driven by managers. Manager attrition decreases by over 1pp per month, a reduction by roughly half, and statistically significant at the 10% level. The decrease is seen in raw counts: there are 10 store manager quits in control stores, but only 4 in treatment stores. Why could there be especially large effects on manager attrition? One reason is that costs of checklists are especially strong for managers, particularly for the daily protocol. Managers spend almost an hour per week on the daily protocol, while workers spend half an hour. In pre-RCT focus groups and discussion with the firm, there was a feeling that checklists took managers away from high-value activities like mentoring and teaching workers. It is also possible that utility costs of checklists are especially bothersome for managers. Store managers are supposed to monitor and lead—by using extensive checklists, the firm may communicate that it doesn't trust store managers to monitor and lead by themselves.

As seen in panel (a) of Appendix Figures A5-A6, there is no evidence that impacts on trained worker attrition or manager attrition fade over time. If anything, treatment effects appear to grow over the RCT, but we cannot reject that effects are constant over time.

Our attrition results use linear probability models, as is common for analyzing treatment effects in RCTs (Angrist & Pischke, 2008). They are highly robust to using Cox models, and, in fact, become stronger statistically, with effects on trained worker attrition and store manager attrition having p-values of 0.05 and 0.03, respectively (see Panel A of Table A7).

Magnitudes. How large are the effects? In a different German bakery chain, Friebel *et al.* (2017) find that introducing a team performance bonus increases sales and customers by 3%—similar to our estimates. However, their treatment raises pay by 2.2%, while compensation remains constant in our RCT, making our RCT more cost-effective. The seminal monitoring RCT by Nagin *et al.* (2002) examines suspicious calls but lacks sales data.

Another study of a specific management practice is Bloom *et al.* (2014). Randomizing Chinese employees to work from home, the study finds a 4% increase in calls per minute—similar to our effect on sales. It also finds a 50% drop in attrition, echoing our results for store managers. Alan *et al.* (2023) conduct an RCT with Turkish firms to evaluate a module delivered by a consulting company, designed to improve the relational atmosphere at work. This RCT reduces manager attrition by roughly 80%, with much smaller effects on workers. The impact of checklist removal on manager attrition is thus broadly similar.

In sum, removing two perceivedly low-value checklists in our setting yields treatment effects of the same order as some of the most promising and highly regarded management interventions. At the same time, we believe our effect sizes are plausible. While it may be

surprising that removing checklists has quantitatively substantial effects, we emphasize that the ones removed are perceived as particularly onerous and ineffective.

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

During-RCT worker survey. Panel A of Table 3 shows that the treatment increased workers' sense of trust between headquarters and workers by 0.28σ, as well as increased workers' commitment to their store by 0.21σ, both statistically significant at the 5% level. 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.[26]

Online customer reviews. To further understand the impact on sales, Panel B performs a text analysis of Google reviews scraped using the developer tool `Apify.com`. Using the text of reviews scraped for 2019m1–2022m1, 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 focus on positive comments in reviews, as negative ones are relatively rare. Our outcomes of interest are the share of reviews in each store-month saying something positive about different attributes (e.g., about speed of service). Because we aim to understand mechanisms, we focus primarily on the text of the reviews. (We analyze star levels in Appendix B.10, where we show there is no effect.) Appendix B.10 discusses the classification procedure in detail, including the issue that many reviews don't contain text.

The text analysis yields several findings. First, as shown in Panel B of Table 3, the share of reviews mentioning something positive about speed of service increases by 1.0pp—from 0.74% in control to 1.75% in treatment stores (column 4). This more than doubling from a low baseline is statistically significant at the 5% level under conventional clustering and at

[26] A concern with interpreting results on worker trust is trust was mentioned by the COO in his message introducing checklist removal. However, the *During-RCT worker survey* was conducted half a year after the COO's message, making Hawthorne Effects unlikely to explain our findings. Section 5.2 discusses further.

the 1% level under RandInf. Second, the share of reviews with positive comments about shop appearance rises by 1.4pp (from 1.6% to 3.0%), an 86% increase, also statistically significant at the 5% level (column 3). Positive comments on shop appearance include remarks on cleanliness, product presentation, and overall ambiance. Employees who feel more committed and trusted, and who have more time, may invest greater effort into serving clients quickly and improving store appearance. Comments on good customer service increase by 2.6pp (from 19% to nearly 23%), though this effect is not statistically significant. As expected, the treatment has no effect on product availability, which is standardized across stores.

In sum, customers in treatment stores are more likely to report being served fast and to experiencing a better store appearance (such as a cleaner store). Research indicates that improvements in speed of service and store appearance are highly plausible channels by which sales can increase in bakeries (Friebel *et al.*, 2017).

Rare problems. Mystery shopping scores indicate that many aspects of rule-following and operational quality, from store appearance to the quality of baked rolls, do not decline with checklist removal. But could removing checklists increase rare problems, as might happen if surgical checklists were eliminated in operating rooms? We examine this using Google reviews. One-star reviews, which reflect strongly negative customer experiences, show no effect of the treatment: the coefficient is near zero, and with 95% confidence we can rule out increases above 2.7pp. A German-speaking RA also manually reviewed all one-star reviews for highly negative experiences (e.g., food poisoning) and found no treatment effect. In sum, there is no evidence that checklist removal increases rare problems in our setting.

4 Treatment Effect Heterogeneity

RMs had strong beliefs about in which stores the treatment would be successful. Thus, we focus our analysis of treatment heterogeneity on RM predictions. In stores where RMs predict the treatment to work, the treatment increases sales and decreases trained worker and manager attrition. Moreover, effects on sales and attrition are differentially stronger, in economic and statistical significance, in stores where the treatment is predicted to work.

Heterogeneity by RM predictions. Table 4 separates treatment effects on store outcomes by RM predictions, showing that effects are much stronger in stores where RMs predict the treatment to be beneficial. In such stores (Panel A), sales increase by 5.2%, which is highly statistically significant ($p = 0.010$ under conventional clustering and $p = 0.005$ under RandInf). There are similar increases among busy and slow sales. The number of customers increases by 4.8%, and shrinkage—a combination of wasted product and theft—goes down

2.4%, though this decrease is not statistically significant. In contrast, for stores where the treatment is not predicted to work (Panel B), the effects on sales are zero and shrinkage *increases* by 2.4%, though this decrease is also insignificant.

Panel C of Table 4 shows p-values regarding equality of the treatment effect in stores where RMs predict the treatment to work and not to work. Beyond showing two-sided p-values, which are common for analyzing interaction terms, we also present one-sided p-values, which we believe are more appropriate given the explicitly one-sided prediction of store managers (i.e., dividing stores in the ones where the treatment will work and ones where it will not work). Under one-sided p-values, all interactions with respect to sales, customers, and shrinkage are statistically significant at $p = 0.06$ or less for conventional clustering.

Returning to Figure 5, panels (b) and (c) show results over time by RM predictions. Restricting to stores where RMs predict the treatment will work, effects are similar over both 5-month halves of the RCT (panel (b)). We can't reject constant effects over time.

Table 5 shows treatment effects on attrition separating by RM predictions. Overall attrition decreases by 0.44pp in stores where the treatment is predicted to work and increases by 0.48pp in stores where it is predicted not to work. Both effects are not significantly different from zero, but the difference between the two effects is significant under a one-sided test ($p = 0.03$). Turning to trained workers, who are the firm's main focus for attrition, attrition decreases by 1pp or roughly 2/3 in stores where the treatment is predicted to work, and this effect is significant for trained non-managers and managers. The drop in trained worker attrition is entirely driven by stores where treatment is predicted to work.

The drop in store manager attrition is also fully driven by stores where the treatment is predicted to work. In those stores, attrition decreases by 2.2pp per month, essentially a complete reduction relative to control. In the raw data, in stores where the treatment is predicted to work, there are 8 store manager quits in control stores, but only 1 in treatment stores. In contrast, in stores where the treatment is not predicted to work, the effect is zero. This difference is statistically significant at the 5% level.[27]

Robustness of RM predictions as a source of heterogeneity, and other heterogeneity dimensions. To assess the robustness of RM predictions as a source of heterogeneity, we apply two alternative approaches to estimating heterogeneous treatment effects, both of which strongly support robustness. We analyze effects on our main outcomes of sales and trained worker turnover. Our first approach is to run linear regressions containing interactions of the treatment variable with six pre-RCT store characteristics. These six include

[27]Store manager attrition is 3x higher in stores where the treatment is predicted to work. Two factors may explain this: (i) RMs may have information about which store managers are likely to quit——possibly due to overmonitoring——and expect the treatment to help; (ii) RMs may believe that stronger store managers are the ones quitting and see the treatment as especially helpful for more capable store managers.

our three main pre-registered dimensions of heterogeneity (RM predictions, tenure, head count). We supplement these with three pre-RCT measures of store performance: baseline outcome (i.e., sales or attrition), mystery shopping, and shrinkage. These capture store quality in different ways. As seen in Table A9, the interaction of treatment with RM prediction is highly robust. This occurs whether other interactions are added one at a time or all at once. No other pre-RCT characteristic analyzed is a consistent predictor of heterogeneity, either across specifications or across both primary outcomes.

A concern with simple interaction terms is that correlation among pre-RCT characteristics can reduce efficiency and, in small samples like ours, bias estimates. As a second approach, we use two machine learning methods to assess treatment heterogeneity: sorted effects (Chernozhukov *et al.*, 2018) and causal random forests (Wager & Athey, 2018).[28] Appendix Table A10 shows that in high-benefit stores, RMs overwhelmingly predict the treatment will work, while in low-benefit stores, they do not.

Mechanisms for the RM predictions. Why are RM expectations predictive of the treatment effect? To explore this, we do a text analysis of RMs' pre-RCT predictions. Appendix Tables A12-A15 provide the full raw text (translated into English) of the notes recorded during the phone calls with RMs.

There are two salient features of responses for stores where RMs predict the treatment would work. First, RMs often say that workers will enjoy having fewer checklists—for instance, for one store, an RM notes that workers "Would be very happy about less bureaucracy; less work as a result; do not like to work with notes and strict rules; will work." This aligns with the utility cost of monitoring in Section 1. Second, RMs frequently mention that teams are unlikely to face problems because communication is already good. For example, one store "Could live without bureaucracy; very communicative store manager." Some predictions combine both themes, such as an RM saying "Team will be glad when operational list is gone. No problems expected. Will work out!"

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

Table 7 examines correlates of RM predictions using the same observables as in our analysis of treatment effect heterogeneity (minus RM predictions). It reveals, first, that only

[28] Sorted effects parameterize and rank conditional treatment effects, while causal random forests estimate them nonparametrically. We adapt both to account for clustering and multiple testing.

two observable characteristics significantly correlate with RM predictions. In addition, such characteristics explain only a modest share of RM predictions ($R^2 \approx 0.14$). The largest predictor of RM predictions is a store's pre-RCT mystery shopping score: RMs believe removing checklists will be more effective in stores with higher pre-RCT scores. Pre-RCT Log Sales and pre-RCT mean worker tenure are not significant predictors of RM predictions.

Could our results on RM predictions be due to RMs behaving differently in treatment vs. control stores? As discussed below in Section 5.2, this is highly unlikely, as RMs have no incentive to behave differently and we see no effect of the treatment on RM store visits.

Observed vs. unobserved store characteristics? Does the estimated treatment effect heterogeneity reflect RMs combining standard observable characteristics in their predictions, or does it reflect RMs having information about characteristics that are unobserved by the econometrician? We address this question using the selection model framework of Dal Bó *et al.* (2021), which strongly supports the latter, namely, information about store characteristics unobserved by the econometrician (see Appendix B.11 for details). This is consistent with Table 6, as a store's tendency to avoid problems or enjoy not having checklists may not be well-captured by standard observable store characteristics.

The role of RM tenure. Because the personnel data contain no information on RMs, we manually collected a coarse measure of RM tenure using web searches: 11 RMs have served since before 2019, and 4 began in 2019 or later. Table A11 tests whether the predictive power of RM beliefs differs by RM tenure. For store-level outcomes, the interaction terms show that beliefs of higher-tenure RMs are significantly more predictive of treatment effects than beliefs of lower-tenure RMs. For attrition, the interaction terms are insignificant, though 4 of 5 have the expected sign. While this analysis faces clear limitations of statistical power (due to having only 15 RMs), these patterns suggest that RMs learn over time which stores are most likely to benefit from checklist removal.

5 Profits, Further Analyses, and Validity Threats

After analyzing the profit implications of our results in Section 5.1, Section 5.2 discusses threats to internal validity. Section 5.3 discusses issues of external validity. Section 5.4 provides further discussion and analysis related to the interpretation of our results.

5.1 Profit Implications of Results

The RCT is highly profitable for the firm, with an estimated benefits to cost ratio of 59:1. To estimate the benefit, the treatment effect on sales of 2.7% implies that stores receive an extra

roughly €2,050 per month, given average monthly sales per stores of €75k, that is, exp(.027)-1)*75k ≈ €2,050. Aggregated as if applied over all 145 stores and the 10 months of the RCT, the total revenue benefit is almost €3M. Using a share of value added in bakery chains of 0.56 (Friebel *et al.*, 2022), the gains from the RCT are .56*3M = €1.7M. In addition, assuming a turnover cost of €6000 for trained workers and €720 for untrained workers following Blatter *et al.* (2012), the turnover benefit of the RCT is an additional €150,000, and the total benefit of the RCT is €1.85M. The assumption of a much larger turnover cost for trained workers follows directly from Blatter *et al.* (2012), and also follows our conversations with executives at the firm, who told us that it is fairly cheap to find untrained workers, while trained workers are much more expensive to hire and train. Appendix B.13 gives further details.

Turning to costs, checklists seem inexpensive to remove, as there was no apparent decrease in operational quality or increase in rare problems. Counting hours spent by the project team, as well as time spent by executives, RMs, and others in implementing the RCT, total time cost is unlikely to exceed €31k (details in Appendix B.13).

Comparing total benefits to costs yields an estimated benefit to cost ratio of 59:1, the largest ratio that we are aware of for any management practice. For example, Friebel *et al.* (2022) and Ashraf *et al.* (2025) each have a benefit to cost ratio of roughly 2:1. Another way to assess profitability is to look at profit per store-month. Using a profit margin of 1% (as indicated by top management), we estimate that our RCT more than doubles the profit margin. This indicates high profitability for the RCT.

Finally, there is striking heterogeneity. In stores where RMs predict the treatment to work, the benefit–cost ratio is 112:1, while in stores where they do not, the benefit is effectively zero. This highlights how profitability varies sharply with RM predictions.

5.2 Threats to Internal Validity

Control store frustration. Could treatment effects stem from negative reactions in control stores rather than positive effects in treated ones? For instance, control employees might resent not being selected. To minimize this risk, no store employees (workers or store managers) were told they were part of an RCT in treatment or control stores, and control store employees were not informed by the firm about any potential checklist removal. Still, people may talk, and in designing the RCT, the HR head expected some store managers to share information. Figure A7 shows the geographic location of treatment and control stores.

To mitigate possible contamination, RMs received written guidelines (Appendix C.7) on what to say if asked. They were to explain that some stores were part of a pilot with University of Cologne researchers, randomly selected for fairness, and coordinated with the

works council. Workers were told to contact the council with questions.[29]

To assess contamination, we asked store managers in the *During-RCT store manager survey* in 2021m11 (8 months in) whether the store managers or their workers know about a pilot involving checklist removal. The response was that about 3/4 of store managers and 1/2 of workers in control stores were aware of the pilot (i.e., the RCT). However, reported annoyance was minimal: among those aware of the RCT, the average level of annoyance or disappointment was only about 2.5 on a 1–10 scale. All our results are robust to dropping the small number of control stores where there was believed to be even a modest level of annoyance or disappointment from workers and store managers (Table A18).

This low annoyance is unsurprising. Neither researchers nor the works council head received complaints. Employees are accustomed to pilots—like testing new coffee or prices—run in some stores but not others. That the RCT's existence did not trigger frustration is consistent with studies like Bloom *et al.* (2014, 2024), where participants knew they were randomized into remote work, yet this awareness is not seen as driving results.

RM effort. Could the overall effects we observe be driven by RMs reallocating effort between control and treatment stores (e.g., RMs stop spending time on control stores to focus on helping treatment stores)? Anecdotally, the firm believes this is very unlikely because RMs had other key concerns during the RCT, such as covid. Also, using the Nov. 2021 *During-RCT store manager survey*, we see no impact of the treatment on the number of visits RMs make to each store or the amount of time spent on such visits (Table A19).

Separate from the overall treatment effect, could RM effort drive the fact that the treatment effect is entirely concentrated in stores where RMs 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 RMs of predicting correctly. Finally, in the same *During-RCT store manager survey*, there is no impact of the treatment on time with RMs (measured in number of visits or time spent on visits) even when restricting to stores where RMs expect the treatment to work (Appendix Table A19).

Hawthorne effects. A separate concern in any RCT is whether subjects alter behavior to please researchers. As noted above, workers and store managers were not informed they were part of an RCT, though some information did leak. We have two responses to this concern. First, treatment effects persist throughout the 10-month RCT, which seems inconsistent with evidence that Hawthorne effects are generally short lived (Levitt & List, 2011). Second, Hawthorne effects can't easily explain key heterogeneity results by RM beliefs.[30]

[29]This builds trust, as German workers have faith in elected councils. The works council once contacted us when a store manager didn't get a survey voucher, suggesting they'd raise broader concerns if needed.

[30]Hawthorne effects could only drive heterogeneity by RM predictions if RMs had information about the

Multiple hypothesis testing. In a study addressing heterogeneous treatment effects and multiple outcomes, one worries that treatment effects could be spurious due to multiple hypothesis testing. We address this first via rigorous pre-registration. Our main outcomes were listed in the pre-registration before the RCT began, and we say that our primary outcome is store sales, and that we will study heterogeneity according to RM expectations. In addition, we address multiple hypothesis testing using the Westfall & Young (1993) step-down procedure. As seen in Appendix Table A10, our machine learning treatment heterogeneity results are robust to a multiple testing correction.

Table A20 addresses multiplicity of outcomes. We consider the two main outcomes, store sales and trained worker attrition. Using all stores, p-values increase from 0.07–0.08 to 0.15. Using stores where RMs predict the treatment will work, p-values increase only from 0.005–0.01 to 0.026. Thus, results using all stores are more sensitive to addressing multiplicity of outcomes, but results for RM-predicted-to-work stores remain highly robust.

Average treatment effects using all stores can be somewhat noisy. The average treatment effect estimates of checklist removal in Section 3 are sometimes somewhat noisy, and we cannot always reject null effects with high confidence. This is unsurprising: theory predicts substantial treatment effect heterogeneity, yet our overall estimates pool stores where the treatment is expected to work with those where it is not. In the subset of stores where the treatment is expected to work, effects are precisely estimated—even after adjusting for multiple hypothesis testing. We hope our approach of eliciting managerial predictions prior to random assignment can be useful for studying other management interventions.

Contemporaneous policy changes. There were no contemporaneous policy changes in treatment stores. Consistent with this, we see no evidence the treatment affected worker earnings (results not shown)—whether overall, for particular types of workers, or in stores where RMs predicted the treatment would work—suggesting no changes in pay policies.

5.3 External Validity

Covid. The RCT ran in April 2021-Jan. 2022. However, we do not believe there are significant external validity concerns related to covid. The German covid lockdown was almost over by the start of the RCT and was fully over by May 2021, and food retail establishments including bakery stores like ours were exempt from the lockdown. All stores were fully open during the RCT, including store coffee areas. Both the operational checklist and daily protocol were used before, during, and after the pandemic. The operational checklist often had an

extent of Hawthorne effects across stores. No RM mentioned Hawthorne Effects in their explanations about why the treatment would work in particular stores.

item or two related to covid (Figure 2 has an example), but these were otherwise unaffected.

External validity with respect to other firms. Like all RCTs, our results are specific to our context, namely, a leading firm in the German bakery chain industry. Do other industry firms use checklists and in what form? To address this, we conducted informal surveys of comparable firms—German bakery chains with more than 50 branches. We surveyed five chains in a large metropolitan area in Germany (distinct from that of our study firm), conducting 21 in-person interviews in total (3–6 per firm) using a research assistant. All five bakery chains report using checklists, and this was consistent across the employees interviewed. While checklists vary by firm, they tend to focus on operational issues, broadly similar to the issues in the operational checklist. The interviews also suggested that workers spent significant time on checklists. These findings, drawn from the industry we study, reinforce the point made in our Introduction that checklists are pervasive in large retail firms.

The heterogeneity of effects within our firm suggests that returns to treatments like ours may vary across firms. Where problems arise frequently or are costly, checklists may be important and their removal could be harmful. By contrast, in settings where checklists are time-consuming or perceived as mistrustful, removing them may be more beneficial.

External validity with respect to tasks. Our RCT removed two low-value checklists. Should one be concerned that these were not "typical" monitoring tasks or that we did not randomly select which checklists to remove? We think not. Our goal is **not** to estimate the effect of removing a typical or randomly chosen checklist. Such an exercise would likely yield negative effects, as economists generally think firms try to optimize and most duties serve some purpose. Instead, our contribution is to provide a methodology for identifying potentially low-value tasks and studying the effects of removing them.

5.4 Additional Analyses and Issues of Interpretation

Testing for the time use channel using the *Pre-RCT store manager survey*. Checklist removal increases worker trust and commitment in the worker survey, and utility benefits are reflected in RM predictions. Separate from improving morale, another possible mechanism is that checklist removal frees up time for workers to spend on production. Table A21 studies treatment effect heterogeneity based on store time on the daily protocol in the pre-RCT period, using the *Pre-RCT store manager survey*. As seen in column 1, there is no evidence that the effect on sales varies with pre-RCT time spent on the protocol. Rather than focusing on total time, columns 2–4 test whether effects are larger for sales during the times of day when stores tended to perform the protocol pre-RCT, but again we find no such pattern. Recall that the daily protocol takes more time than the operational checklist.

In sum, unlike the utility benefit channel, we see no direct evidence here for the time use channel, though the evidence is necessarily indirect. A descriptive check based on manager perceptions during the RCT is provided in Appendix B.14 and yields a consistent conclusion.

Mediation analysis. We explore mediation analysis to better understand mechanisms for effects. First, we examine to what extent treatment effects are mediated by the increase in trust observed in the *During-RCT worker survey*. A mediation analysis (Table A22) shows that 14% of the treatment effect on trained worker attrition and 21% of the effect on sales is mediated through an increase in trust. However, estimates (including the share mediated) are quite noisy, reflecting there are only 395 workers and 100 stores in the *During-RCT worker survey*, so employee trust is noisily measured. A mediator will be biased down in magnitude due to classical measurement error, so it could be that the true share mediated by trust is higher. Overall, we view this exercise as inconclusive. Second, we examine to what extent sales effects are due to attrition. A mediation analysis provides no evidence that the sales effect is mediated by the turnover of trained workers or store managers (Appendix B.15). This suggests our treatment affects store outcomes directly instead of through attrition.

Authority as a mechanism. Our interpretation emphasizes utility benefits of checklist removal, supported by worker and manager surveys. A conceivable alternative is that removing checklists gave workers more authority to make better decisions—e.g., choosing how to greet particular customers (e.g., "Good morning" vs. "Hey") or display products. Two points argue against this. First, the RCT didn't change formal authority: workers were still required to perform the same tasks in the same way, just without signing checklists, and checklist contents were still reinforced in internal newsletters. Second, relevant behaviors continued to be monitored through mystery shopping.

Could the checklists have been initially useful? While removing the two checklists is beneficial on average in the RCT, this does not imply they were always harmful for the firm. Perhaps the checklists were valuable when first introduced by the firm, e.g., for instilling good routines. Supporting this, the benefit in terms of employee attrition is larger for trained workers (Table 2), and there is some suggestive evidence that the benefit may increase with worker tenure (Table A9). However, we find no analogous pattern for store-level outcomes. The treatment effect on store outcomes does not vary with average worker tenure or store age. If the checklists helped establish routines, their removal might have been less beneficial in younger stores—but we find no such evidence. Even if checklists were originally valuable and became less so over time, this does not mean the firm was wrong to remove them. It just means the benefits may fade once many new workers join the firm.

Is it "obvious" that RMs would be informed? As surveyed by Hoffman & Stanton

(2025), prior work in personnel and organizational economics shows that middle managers often struggle with core performance predictions, such as which applicants will perform best if hired or which employees will succeed after promotion. Thus, prior to our RCT, it was extremely non-obvious that RMs would make accurate predictions about which stores the treatment would work in. This was especially true because we provided no guidance on how many stores RMs should predict as successes; one RM predicted the treatment would work in none of their stores, while two RMs predicted it would work in 80% of stores or more.

RMs are informed, but is it "private information?" One view of the firm is that the principal (the CEO/owner) has the same information as all agents within the firm. Another is that information is dispersed, with agents having some information that the principal does not (Hayek, 1945). Given the CEO's high cost of time, it wasn't feasible to have the CEO predict whether the treatment would work in particular stores, let alone for all 145 stores. However, we think it is extremely unlikely the CEO would be able to make predictions like the RMs. 145 stores is too many for the CEO to keep accurate tabs on.[31]

Is the contribution of our study merely to show that it is beneficial for firms to remove "bad" practices? Certainly not. First, prior to our study, it was not clear that checklists can sometimes be harmful for performance. Second, our results do not suggest that the widespread use of checklists in retail is bad. We had to discover which checklists were harmful, and effects varied massively across stores, and this was not known *ex ante*.

6 Firmwide Rollout

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

However, in the firmwide rollout, only the operational checklist was removed; the daily protocol was reinstated. A key reason was that feedback from both workers and store managers supported some value in retaining the daily protocol. Some workers—and especially some store managers—viewed it as useful for coordinating production (Alonso *et al.*, 2008).

[31] It was impossible for us to have workers and store manager predict treatment effects because they didn't know there was an RCT. It is quite possible, however, that workers and store managers could predict whether the treatment would be effective in their store. Such a possibility does not threaten whether RMs have "private information," which we view as information unobserved to the CEO (following a principal/agent framework). Future work could consider which hierarchy layers are most informed.

In treated stores, workers and store managers were asked in their respective *During-RCT surveys* whether they agreed it was a good decision to remove each checklist. Both groups strongly supported removing the operational checklist, with mean agreement levels of 5.7 and 6.0, respectively, on a 1–7 scale (1 = strongly disagree, 7 = strongly agree). However, support for removing the daily protocol was weaker: workers gave an average of 4.9, while managers gave just 3.1. As of September 2022—eight months after the rollout—when we had our final formal discussion with the firm, the operational checklist remained discontinued.

Sales and trained worker attrition. Figure 5 shows the difference between treatment and control stores in sales during the post-RCT period. Results also appear in Appendix Table A23, which repeats Table 2 while restricting to the post-RCT period. Under the firmwide rollout, the difference in sales between treatment and control stores drops from 2.7% to 1.6%. This is a drop in the coefficient of almost half, and the difference is no longer statistically significant. The difference in trained worker attrition is starker, going from a coefficient of -0.44pp during the RCT to a coefficient of 0.46pp during the rollout.

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

Why wasn't the rollout differentiated by store? Given the heterogeneity results, one might ask why the firm did not implement checklist removal only in stores where RMs expected it to work. While checklist removal generates sizable positive effects in those stores, it generates no sizable negative effects in stores where RMs predicted it would not. In these stores, checklist removal was roughly neutral. Thus, even if the firm had differentiated the rollout using RM beliefs, our estimates imply that overall outcomes wouldn't have improved. Also, although differentiation was feasible during the RCT, the firm believed there would be longer-run costs to maintaining different procedures across stores. The firm frequently opens new stores, and differentiating practices would require regularly eliciting RM predictions for new locations, often before RMs have meaningful information. In short, we can't reject that the firm was optimizing by choosing a uniform rollout rather than tailoring to RM beliefs.

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

7 Conclusion

Checklists are an extremely common management practice in retail and other industries. Scholarship in economics and other fields often focuses on benefits of checklists and monitoring in general. In a large German bakery chain, surveys of workers and store managers indicate wide variation across checklists in time cost and perceived benefit. Removing two of the perceivedly lowest-value checklists improves average store performance as measured by sales and store manager attrition, and without a reduction in mystery shopping scores. Performance benefits are comparable in size to those from introducing well-known costly management practices like team bonuses, but the costs are much smaller. The benefit to cost ratio from our intervention of roughly 60 is the largest (to our knowledge) in the management practice literature. Online reviews indicate improvements in speed of service and shop appearance. In surveys, relative to control workers, treated workers perceive greater trust between workers and managers at the firm and feel more committed to their stores. Even apart from heterogeneity analysis, we believe that cleanly estimating these overall effects is a major contribution to scientific understanding of checklists and monitoring.

Pre-RCT conversations with RMs suggested that treatment effects may vary substantially across stores—consistent with theory (Benabou & Tirole, 2003; Ellingsen & Johannesson, 2008) indicating that the effects of monitoring may depend heavily on context (e.g., because it may signal negative information to some employees). Thus, we asked RMs to predict treatment efficacy for all their stores before treatment assignment, and we find that positive performance effects are entirely concentrated in stores where RMs predict the treatment to be effective. This is not due to RMs spending more time with, or being partial to, those stores. Rather, the evidence suggests that RMs have knowledge about which stores are most likely to benefit from checklist removal, and the relevant knowledge appears to exist regarding unobserved-to-the-econometrician characteristics of stores. Text analysis of RM predictions indicates that RMs focused on which stores would benefit from the structure of checklists and which stores' workers would experience utility benefits from checklist removal.

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

Our evidence suggests that removing these checklists may be a "win–win," improving firm profits and worker welfare, at least for trained workers. The worker-welfare conclusion is supported by lower trained turnover and workers' stated satisfaction with checklist removal. Thus, this RCT illustrates how management interventions can benefit both sides and raise welfare. The positive sales effects also suggest possible gains for customers, though impacts

on customer welfare are more speculative.

The RCT lasted 10 months before checklist removal was rolled out firmwide, a long period relative to most management practice RCTs (Bloom *et al.*, 2020). Impacts on sales and store manager attrition are durable, remaining strong in months 6–10 of the RCT. The rapid firmwide rollout further underscores the durability of the treatment effects.

Our RCT suggests the firm was not fully optimizing before the RCT, raising the question: why? One possibility is top management underestimated how burdensome employees found the checklists. While they likely had a rough sense of how much time checklists take, they may not have realized how many workers found them distasteful. Employees and RMs may have felt uncomfortable voicing concerns to top management (Fornasari *et al.*, 2025).

We hope future RCTs will continue to examine when reductions in checklists or workplace control generate positive returns and when they do not.

References

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

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

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

Angrist, Joshua D., & Pischke, Jörn-Steffen. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Ash, Elliott, & MacLeod, W. Bentley. 2015. Intrinsic Motivation in Public Service: Theory and Evidence from State Supreme Courts. *Journal of Law and Economics*, **58**(4), 863–913.

Ashraf, Nava, Bandiera, Oriana, Minni, Virginia, & Zingales, Luigi. 2025. *Meaning At Work*. Working Paper 33843. National Bureau of Economic Research.

Bai, Yuehao, Jiang, Liang, Romano, Joseph, Shaikh, Azeem, & Zhang, Yichong. 2024. Covariate Adjustment in Experiments with Matched Pairs. *J. Econometrics*, **241**(1), 105740.

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.

Belloni, Alexandre, Chernozhukov, Victor, & Hansen, Christian. 2014. Inference on Treatment Effects After Selection Among High-dimensional Controls. *Rev. Econ Studies*, **81**(2).

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.

Benson, Alan, & Shaw, Kathryn. 2025. What Do Managers Do? An Economist's Perspective. *Annual Review of Economics*, Forthcoming.

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

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

Bloom, Nicholas, & Van Reenen, John. 2011. Human Resource Management and Productivity. *Handbook of Labor Economics*, **1**, 1697–1767.

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.

Bloom, Nicholas, Han, Ruobing, & Liang, James. 2024. Hybrid working from home improves retention without damaging performance. *Nature*, 1–6.

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, Karlan, Dean, & Osman, Adam. 2024. Big Loans to Small Businesses: Predicting Winners and Losers in an Entrepreneurial Lending Experiment. *AER*, **114**(9).

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

Coviello, Decio, Ichino, Andrea, & Persico, Nicola. 2014. Time Allocation and Task Juggling. *American Economic Review*, **104**(2), 609–623.

Cytrynbaum, Max. 2024. Covariate Adjustment in Stratified Experiments. *Quantitative Economics*, **15**(4), 971–998.

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. 2022. *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.

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

Dessein, Wouter, & Prat, Andrea. 2022. Organizational Capital, Corporate Leadership, and Firm Dynamics. *Journal of Political Economy*, **130**(6), 1477–1536.

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

Dessein, Wouter, & Santos, Tano. 2021. Managerial Style and Attention. *American Economic Journal: Microeconomics*, **13**(3), 372–403.

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.

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

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

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

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

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

Fornasari, Margherita, Rasul, Imran, Rogger, Daniel, & Williams, Martin J. 2025. *Ideas Generation in Hierarchical Bureaucracies: Evidence from a Field Experiment and Qualitative Data*. CEPR Discussion Paper No. DP20388.

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

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

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

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

Garicano, Luis, & Rossi-Hansberg, Esteban. 2006. Organization and Inequality in a Knowledge Economy. *Quarterly Journal of Economics*, **121**(4), 1383–1435.

Garicano, Luis, & Rossi-Hansberg, Esteban. 2015. Knowledge-based Hierarchies: Using Organizations to Understand the Economy. *Annual Review of Economics*, **7**(1), 1–30.

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.

Guadalupe, Maria, Rappoport, Veronica, Salanie, Bernard, & Thomas, Catherine. 2024 (Dec). *The Perfect Match: Assortative Matching in Mergers and Acquisitions*. LSE Research Online Documents on Economics 126749. LSE.

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

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

Hayek, Friedrich August. 1945. The Use of Knowledge in Society. *AER*, **35**(4), 519–530.

Herz, Holger, & Zihlmann, Christian. 2022. *Adverse Effects of Control: Evidence from a*

Field Experiment. CESifo Working Paper No. 8890.

Hoffman, Mitchell, & Stanton, Christopher T. 2025. People, Practices, and Productivity: A Review of New Advances in Personnel Economics. *Handbook of Labor Economics*.

Holmstrom, Bengt. 1979. Moral Hazard and Observability. *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. 2012. Insider Econometrics: A Roadmap for Estimating Empirical Models of Organizational Design and Performance. *Handbook of Organizational Econ.*

Ichniowski, Casey, Shaw, Kathryn, & Prennushi, Giovanna. 1997. The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines. *American Economic Review*, **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. 2024. Monitoring in Small Firms: Experimental Evidence from Kenyan Public Transit. *AER*, **114**(10), 3119–60.

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

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

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

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

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

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

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

Shaw, Kathryn. 2009. Insider Econometrics: A Roadmap with Stops Along the Way. *Labour Economics*, **16**(6), 607–617.

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

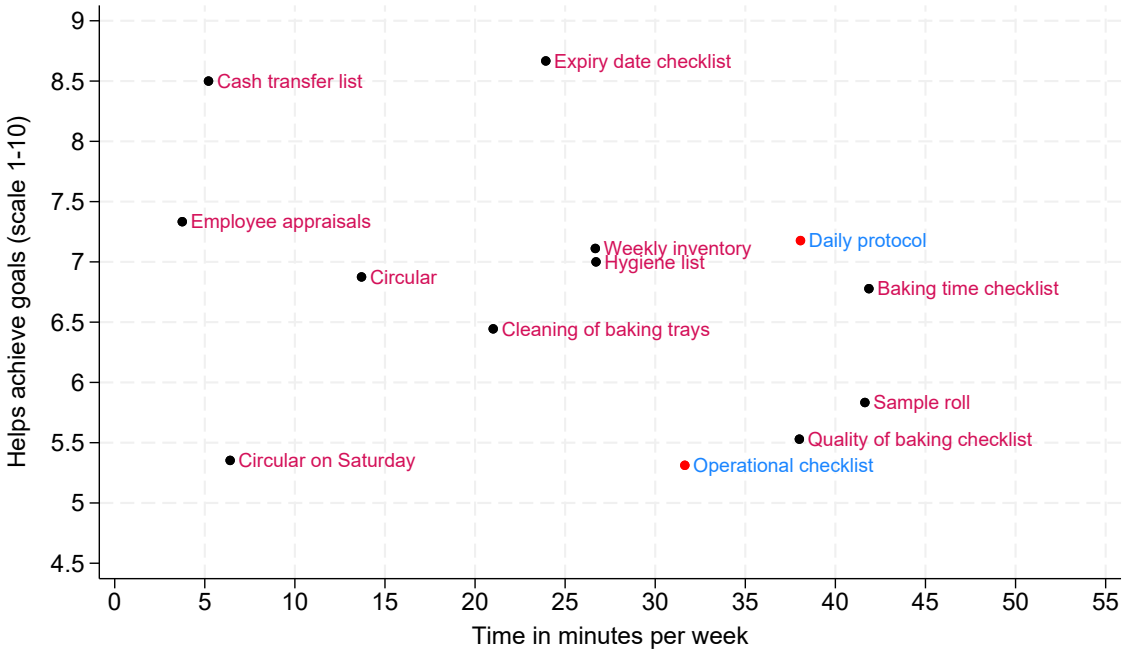
Wager, Stefan, & Athey, Susan. 2018. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *JASA*, **113**(523), 1228–1242.

Westfall, Peter, & Young, S. Stanley. 1993. *Resampling-based Multiple Testing*. Vol. 279.

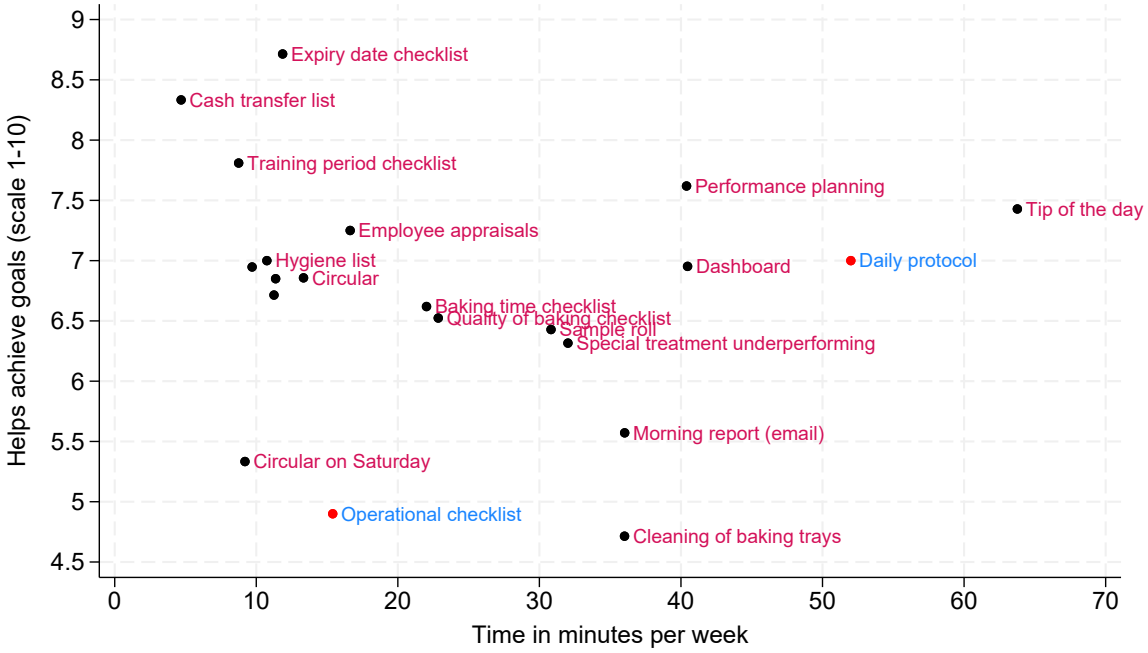
Womack, James P, Jones, Daniel T, & Roos, Daniel. 2007. *The Machine That Changed the World: The Story of Lean Production*. Simon and Schuster.

Young, Alwyn. 2019. Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results. *QJE*, **134**(2), 557–598.

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



(a) Workers (N=18 workers)



(b) Store Managers (N=21 store managers)

Notes: This figure uses data from the *Pre-RCT in-depth interviews* of 18 workers and 21 store managers described in Section 2. For each checklist, the figure plots the mean amount of time across workers spent on the checklist, as well as the mean level of agreement with the statement: "The checklist helps (FIRM) to get better and reach company goals." Our pre-RCT interviews ask separately about reaching company goals and avoiding mistakes. Results on avoiding mistakes are similar, and are shown in Appendix Figure A3. As seen in the lower-right of panel (a), there are five checklists for workers that stand out for relatively high time per week and relatively low value in achieving goals. The two checklists removed in the RCT are colored in blue.

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

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

Notes: This figure provides an example of the operational checklist. It is translated from German and has firm-identifying information redacted. Appendix C has more examples of the operational checklist, including ones that focus more on communication with customers (e.g., eye contact, smiling).

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

Date: _____ Store: _____

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

Sales (€)		Working hours		Performance	

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

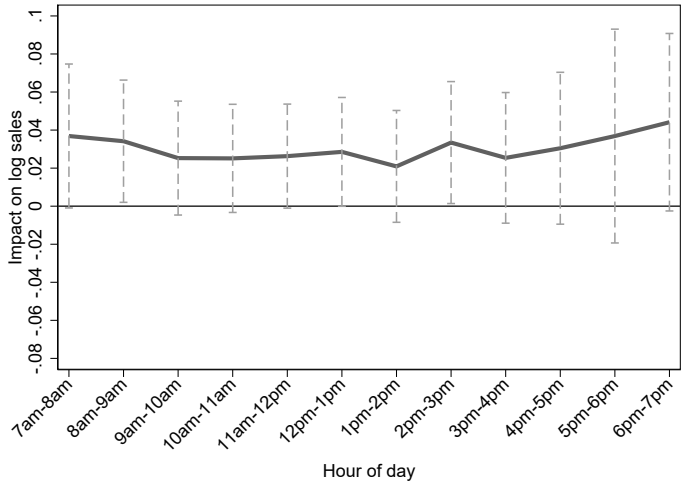
Facility or IT problems, etc.

Shift changes

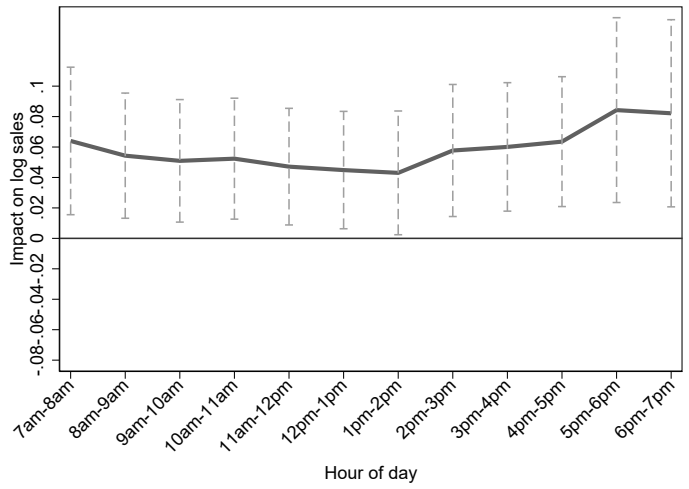
Additional information for the next shift

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

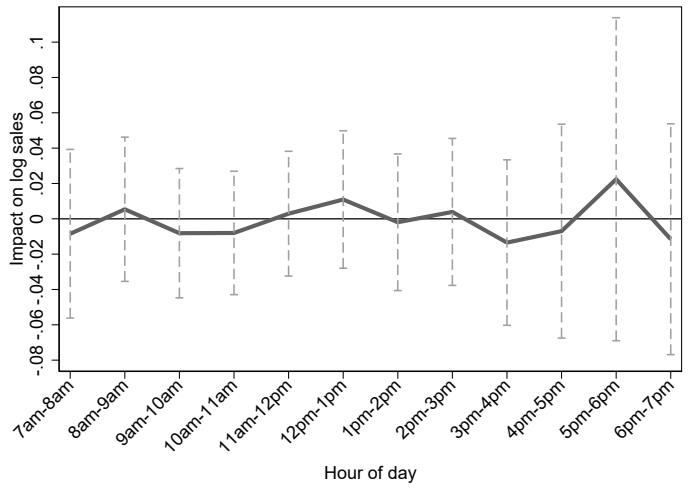
Figure 4: Treatment Effects on Store Sales at Different Hours of the Day



(a) All Stores



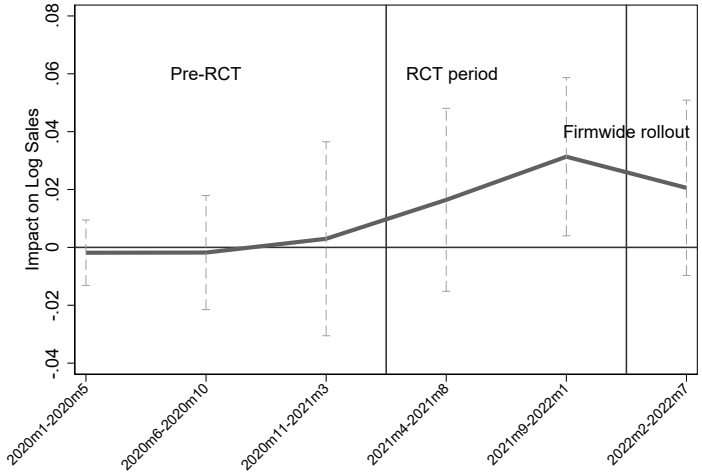
(b) Stores Where RCT Predicted to Work by Reg. Mgrs.



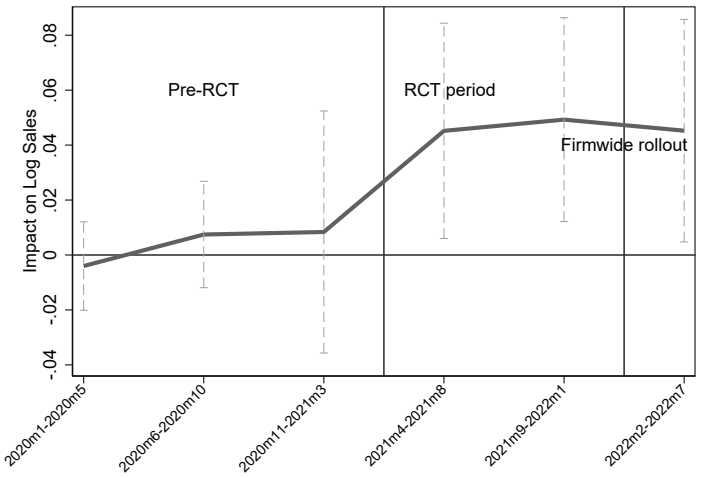
(c) Stores Where RCT Predicted Not to Work by Reg. Mgrs.

Notes: This figure examines how impacts on sales vary by the hour of the day. We estimate regressions similar to those in Panel A of Table 2, but the outcome is log(sales), where sales is monthly sales measured in a certain hour of the day. The line shows the coefficient estimates, and the dotted line shows the 95% confidence intervals. The 95% confidence intervals are calculated using conventional clustering by store.

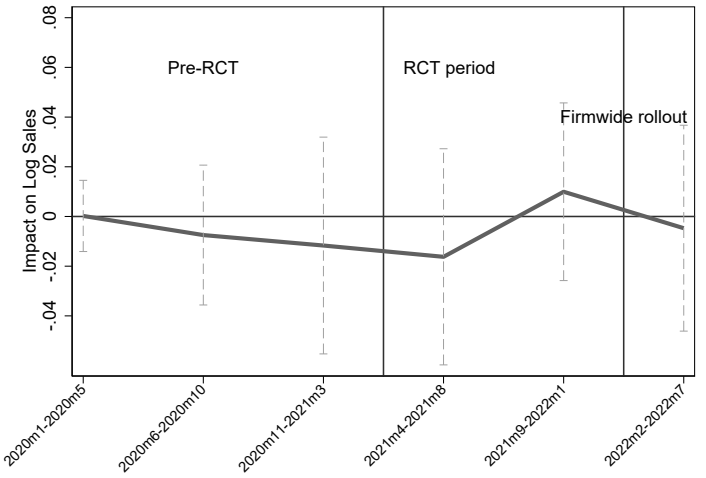
Figure 5: Differences Between Treatment and Control Stores Over Time in Sales



(a) All Stores



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



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

Notes: Panel (a) is similar to that in column 1 of Panel A of Table 2, but we split separately by period of the RCT. The first period of the RCT is April-August 2021 and the second period is September 2021-January 2022. We also show three periods before the RCT, as well as the post-RCT rollout, where we have data for six months. Likewise, panels (b) and (c) here are similar to column 1 in panels (b) and (c) of Table 4. The line shows the coefficient estimates, and the dotted line shows the 95% confidence intervals. Figure A4 shows robustness to other ways of presenting the results.

Table 1: Comparing Pre-RCT Variables between Control and Treatment Stores: Balance Check

	N	Control mean	Treatment mean	p-value of difference
Panel A: Monthly Store Variables				
Log Sales	145	11.16	11.14	.73
Log Busy Sales	145	10.83	10.82	.76
Log Slow Sales	145	9.86	9.84	.72
Log Customers	145	9.85	9.84	.97
Log (Shrinkage as % of Sales)	145	-2.06	-2.06	.9
Mystery Score (normed)	145	.01	-.03	.52
Head count	145	13.27	13.8	.49
Panel B: Monthly Attrition Rates				
All workers	2057	1.69	1.61	.65
Untrained workers	1114	2.58	2.5	.82
Trained workers	943	.92	.84	.62
Trained non-manager	757	.95	.9	.79
Trained manager	186	.81	.63	.5
Panel C: Employee Characteristics				
Female	2057	.97	.97	.34
Age	2057	39.06	38.98	.91
Base hourly wage in euros	2057	12.45	12.44	.92
Monthly bonus in euros	2057	31.78	34.75	.13
Total monthly pay in euros	2057	1655	1637	.47
Tenure in yrs	2057	7.83	8.1	.55
Tenure of 1yr or less	2057	.16	.15	.77
Tenure of 1-2yrs	2057	.11	.1	.53
Tenure of 2-5yrs	2057	.19	.18	.45
Tenure of 5-10yrs	2057	.2	.23	.26
Tenure more than 10yrs	2057	.34	.34	.92

Main notes: This table compares pre-RCT store- and worker-level characteristics across treatment and control stores using data from January 2019 to March 2021 (i.e., the 27 months before the start of the RCT). The p-values of difference are based on a regression of each variable on a treatment dummy. The p-values account for clustering by store. * significant at 10%; ** significant at 5%; *** significant at 1%

Panel A: An observation is a store. The table compares stores in terms of mean pre-RCT values of different store characteristics measured at the monthly level. Store "head count" means the number of employees per store and includes minijobbers.

Panel B: An observation is an employee in the pre-RCT period. The number presented are mean monthly attrition rates for different populations. For example, among all workers in our data in the pre-RCT period, the mean monthly attrition rate is 1.69% in Control stores, meaning 12*1.69 $\approx$ 20% of workers quit the firm each year in control stores. As noted in footnote 15, the monthly bonus is a very small share of total monthly pay.

Panel C: An observation is an employee in the pre-RCT period. As noted in footnote 12, minijobbers are excluded from the employee panel used to analyze employee attrition, and so they are not included here.

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

Panel A: Store Outcomes	(1)	(2)	(3)	(4)	(5)	(6)
Dep. var.:	Log Sales	Log Busy Sales	Log Slow Sales	Log Customers	Log Shrink -age	Mystery Shopping Score (normed)
Treatment	0.027*	0.026*	0.034*	0.023	0.002	0.003
	(0.015)	(0.014)	(0.019)	(0.015)	(0.016)	(0.070)
	[0.074]	[0.064]	[0.079]	[0.122]	[0.888]	[0.982]
Observations	1,431	1,431	1,431	1,431	1,431	1,161
Mean dep. var. if Treat=0	11.17	10.86	9.838	9.762	-2.099	-0.0284
Stores	145	145	145	145	145	144

Panel B: Worker Turnover	(1)	(2)	(3)	(4)	(5)
Sample of workers:	All	Untrained Workers	Trained workers	Trained Non-Mgrs	Trained Managers
Treatment	0.07	0.64*	-0.44*	-0.23	-1.07*
	(0.24)	(0.39)	(0.25)	(0.27)	(0.60)
	[0.777]	[0.095]	[0.085]	[0.393]	[0.082]
Observations	13,271	6,489	6,782	5,403	1,379
Mean dep. var. if Treat=0	2.038	2.806	1.254	1.159	1.647
Workers	1637	863	774	624	150
2-sided p-val: trained v. untrained			0.02		
			[0.02]		
2-sided p-val: manager v. non-mgrs					0.04
					[0.07]

Main notes: Standard errors clustered by store in parentheses. "Rand-t" randomization inference p-values (Young, 2019) are in square brackets (1,000 replications) and account for clustering by store. Stars are based on clustered standard errors in parentheses, with * significant at 10%; ** at 5%; *** at 1%

Panel A: An observation is a store-month during the RCT. Busy sales are sales between 7am and 2pm, whereas slow sales are sales after 2pm or before 7am. Shrinkage is the share of product lost as a share of total sales revenue. Each regression controls for the mean of the dependent variable in the pre-RCT period, year-month dummies, and several pre-RCT store characteristics (above/below median sales, above/below median head count, above/below median store league performance ranking, and region). The number of observations is smaller for mystery shopping scores compared to the other outcome variables, as mystery shopping occurs each month in most stores instead of all stores.

Panel B: An observation is a worker-month during the RCT. Coefficients are multiplied by 100 for ease of exposition. Column 1 includes all workers. Columns 2–3 split the sample into untrained and trained workers. Columns 4–5 further divide trained workers into non-managers and store managers. All regressions include the same controls as in Panel A, plus a quadratic in tenure, gender, and an indicator for being hired during the RCT. Since an observation is a worker-month (instead of a store-month), we additionally control for the store-level mean attrition rate in the pre-RCT period. The "trained v. untrained" and "manager v. non-mgrs" p-values are two-sided tests of whether treatment effects vary by worker skill (e.g., are effects different for trained vs. untrained workers), and come from regressions interacting worker skill (trained/manager) with all regressors. Corresponding randomization inference p-values are shown in brackets below the clustered p-values (e.g., 0.04 vs. 0.07 for the manager comparison).

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

Panel A: During-RCT Worker Survey

Dependent variable: (all normed)	(1) Trust bwn. HQ & workers	(2) Commitment to one's store	(3) Last new hire was well-trained	(4) Basic quality control
	(1)	(2)	(3)	(4)
Treatment	0.280**	0.212**	0.016	-0.074
	(0.138)	(0.095)	(0.117)	(0.108)
	[0.053]	[0.028]	[0.887]	[0.520]
Workers	394	390	368	394
Stores	100	100	99	100

Panel B: Google Reviews

Dep. var.: Whether there is a positive comment regarding:	(1) The product	(2) Service	(3) Shop appearance	(4) Speed of service	(5) Value for money	(6) Product availability
Treatment	0.018	0.026	0.014**	0.010**	0.007	0.004
	(0.022)	(0.018)	(0.007)	(0.004)	(0.007)	(0.009)
	[0.416]	[0.170]	[0.038]	[0.008]	[0.303]	[0.661]
Observations	1,023	1,023	1,023	1,023	1,023	1,023
Stores	142	142	142	142	142	142
Mean DV if Treat=0	0.286	0.189	0.0162	0.00741	0.0194	0.0526

Main notes: Standard errors clustered by store in parentheses. "Rand-t" randomization inference p-values following Young (2019) in square brackets (1,000 replications). Stars are based on clustered standard errors in parentheses, with * significant at 10%; ** significant at 5%; *** significant at 1%

Panel A: An observation is a worker. We control for the pre-RCT store characteristics listed in Table 2. The data are from the *During-RCT worker survey*. Further details on the survey are in Appendix B.9. Figure A2 summarizes all the surveys within the partner firm.

Panels B: An observation is a store-month during the RCT. We control for the mean of the dependent variable in the pre-RCT period, plus year-month dummies and 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. We use all Google reviews regardless of whether they contain text. Reviews without text get a zero for all 6 dependent variables listed here. Results are similar if we restrict to Google reviews with text, as seen in Table A8. Appendix B.10 provides further information about the Google reviews data.

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

Dep. var.:	Log Sales	Log Busy Sales	Log Slow Sales	Log Customers	Log Shrink -age	Mystery Shopping Score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Stores Where RCT Predicted to Work by RMs						
Treatment	0.052**	0.050**	0.058***	0.048**	-0.024	0.080
	(0.020)	(0.019)	(0.022)	(0.019)	(0.021)	(0.089)
	[0.005]	[0.006]	[0.005]	[0.004]	[0.272]	[0.364]
Observations	744	744	744	744	744	597
Mean dep. var. if Treat=0	11.09	10.77	9.761	9.684	-2.063	0.0410
Stores	76	76	76	76	76	75
Panel B: Stores Where RCT Not Predicted to Work by RMs						
Treatment	-0.003	-0.003	0.004	-0.006	0.024	-0.068
	(0.020)	(0.019)	(0.027)	(0.020)	(0.020)	(0.109)
	[0.881]	[0.902]	[0.893]	[0.810]	[0.245]	[0.550]
Observations	687	687	687	687	687	564
Mean dep. var. if Treat=0	11.27	10.96	9.929	9.852	-2.142	-0.108
Stores	69	69	69	69	69	69
Panel C: Comparing Treatment Effects by RM Predictions						
1-sided p-val: Panels A v. B	0.02	0.03	0.06	0.03	0.05	0.15
	[0.02]	[0.03]	[0.07]	[0.03]	[0.05]	[0.15]
2-sided p-val: Panels A v. B	0.05	0.05	0.12	0.05	0.10	0.29
	[0.05]	[0.05]	[0.14]	[0.06]	[0.10]	[0.30]

Panels A and B 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 (RM) predicted the treatment would work in each store. "Rand-t" randomization inference p-values (Young, 2019) are in square brackets (1,000 replications) and account for clustering by store. Stars are based on two-sided p-values and based on the clustered standard errors in parentheses, with * significant at 10%; ** significant at 5%; *** significant at 1%

Panel C notes: We report p-values from tests of whether treatment effects differ by RM prediction. The p-values come from regressions in which the RM prediction dummy is interacted with all regressors (namely, the treatment dummy and the controls in Panels A and B). The p-values without brackets are based on conventional clustering-by-store inference; p-values in square brackets are from randomization inference ("rand-t" p-values with 1,000 replications).

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

Workers:	All	Untrained Workers	Trained workers	Trained Non-Mgrs	Trained Managers
	(1)	(2)	(3)	(4)	(5)
Panel A: Stores Where RCT Predicted to Work					
Treatment	-0.44	0.25	-1.05***	-0.67*	-2.17**
	(0.31)	(0.52)	(0.36)	(0.39)	(0.84)
	[0.167]	[0.574]	[0.004]	[0.084]	[0.009]
Mean dep. var. if Treat=0	2.056	2.667	1.505	1.306	2.228
Observations	6,595	3,126	3,469	2,691	778
Workers	829	422	407	320	87
2-sided p-val: trained v. untrained			0.04		
			[0.04]		
2-sided p-val: manager v. non-mgrs					0.06
					[0.06]
Panel B: Stores Where RCT Not Predicted to Work					
Treatment	0.48	0.98*	0.08	0.10	0.55
	(0.37)	(0.57)	(0.37)	(0.37)	(0.87)
	[0.217]	[0.080]	[0.840]	[0.777]	[0.453]
Mean dep. var. if Treat=0	2.019	2.931	0.967	1	0.806
Observations	6,676	3,363	3,313	2,712	601
Workers	878	483	395	328	67
Panel C: Comparing Treatment Effects by RM Predictions					
1-sided p-val: Panels A v. B	0.03	0.17	0.02	0.08	0.01
	[0.03]	[0.17]	[0.02]	[0.07]	[0.02]
2-sided p-val: Panels A v. B	0.06	0.34	0.03	0.15	0.03
	[0.07]	[0.33]	[0.03]	[0.15]	[0.04]

Panels A and B 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 (RM) predicted the treatment would work in each store. Stars are based on two-sided p-values and based on the clustered standard errors in parentheses, with * significant at 10%; ** significant at 5%; *** significant at 1%

Panel C notes: We report p-values from tests of whether treatment effects differ by RM prediction. The p-values come from regressions in which the RM prediction dummy is interacted with all regressors (namely, the treatment dummy and the controls in Panels A and B). The p-values without brackets are based on conventional clustering-by-store inference; p-values in square brackets are from randomization inference ("rand-t" p-values with 1,000 replications).

Table 6: Responses from the RM 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: Based on the pre-RCT regional manager (RM) prediction survey. This table summarizes themes from free-text explanations provided by RMs for stores where they predicted checklist removal would be beneficial. We restrict to the 78 such stores. For 57 of these stores, RMs provided explanatory comments; the remaining 21 stores are excluded from this breakdown because RMs indicated the treatment would help but did not elaborate further (e.g., saying "Yes, will work"). Although these cases lack detailed explanations, RMs provided predictions for all stores they oversaw. Full translated responses appear in Appendix Tables A12–A15.

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

Specification:	OLS	Lasso-selected regressors
	(1)	(2)
Treatment store	-0.023	
	(0.079)	
Pre-RCT Log Sales	0.089	
	(0.374)	
Pre-RCT mystery shopping score	0.344***	0.339***
	(0.101)	(0.093)
Pre-RCT mean head count	-0.022	-0.022***
	(0.017)	(0.007)
Pre-RCT Log (Shrinkage as % of Sales)	0.220	
	(0.432)	
Pre-RCT mean tenure of workers in years	0.004	
	(0.018)	
Observations	145	145
R-squared	0.146	0.143

Notes: An observation is a store. Robust standard errors in parentheses. Column 1 performs simple OLS. Column 2 uses the regressors selected by lasso, where λ is selected by cross-validation, and it is implemented in Stata 19 using "lasso." The pre-RCT mystery shopping score is based on data normalized at the store-month level and then averaged by store. If the resulting store-level averages are instead standardized, the coefficient on pre-RCT mystery shopping in Column 1 is 0.14—i.e., a one standard deviation increase in mystery shopping predicts a 14pp increase in the chance an RM predicts the treatment to work in that store. If one drills into particular components of mystery shopping matter most into predicting RM predictions, the one that matters the most is the quality of customer interactions, though the overall predictive power remains modest (see Table A16). * significant at 10%; ** significant at 5%; *** significant at 1%

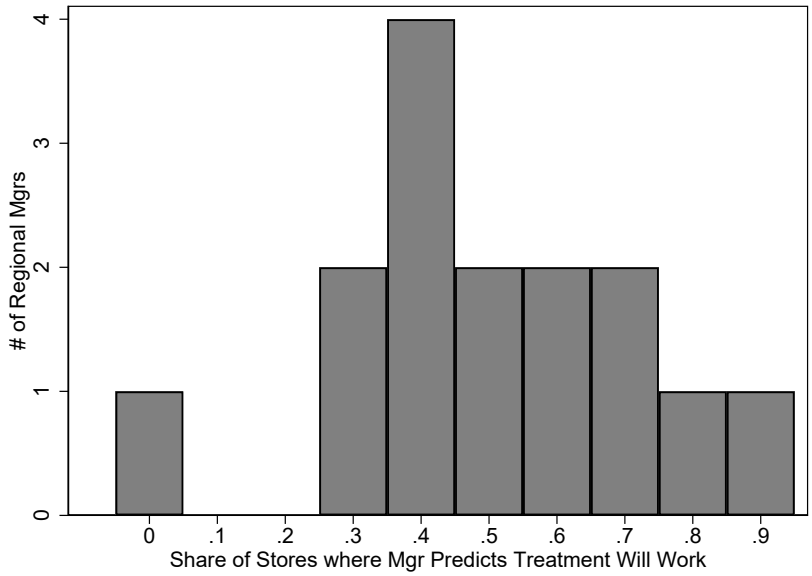
Web Appendix, "Is This Really Kneaded? Identifying and Eliminating Potentially Harmful Forms of Workplace Control", by Friebel, Heinz, Hoffman, Kretschmer, and Zubanov

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

Appendix A Appendix Figures and Tables

As in the main text, for the analyses in the Web Appendix, we estimated treatment effects using both conventional clustered-by-store standard errors and randomization inference. To keep the tables readable—some contain many rows and specifications—and because randomization inference is computationally intensive for certain models, we report randomization inference p-values only for Tables A1, A2, and A8 in the Web Appendix. As in the main text, the results are virtually identical regardless of inference method, making the choice of inference approach largely immaterial for interpreting our findings.[1]

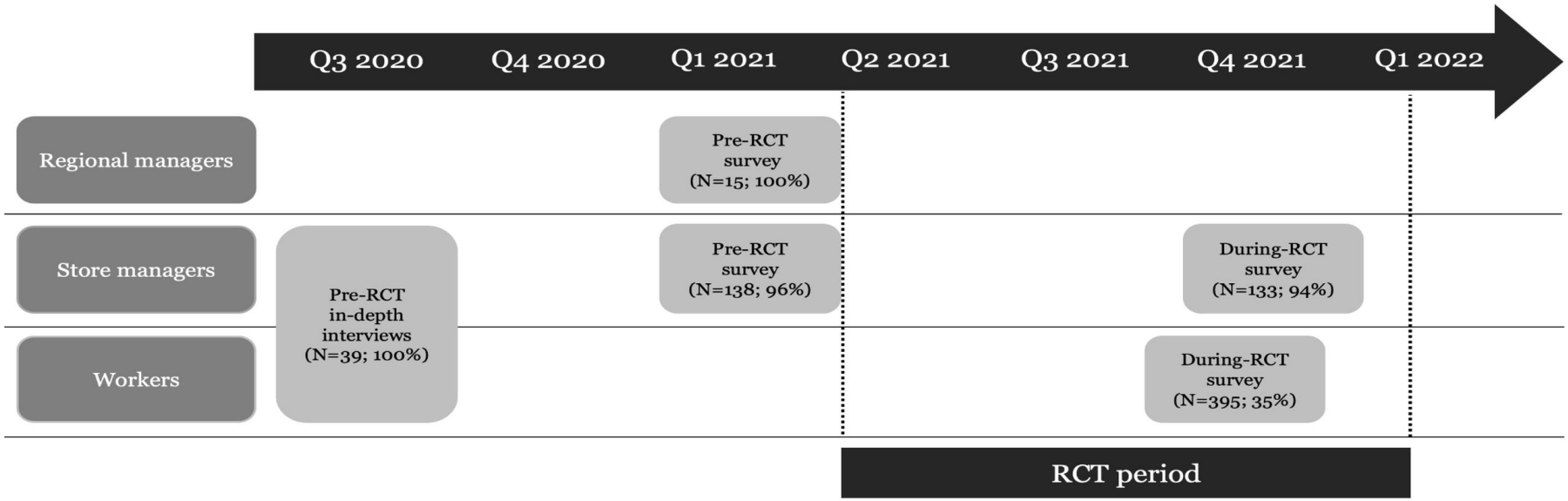
Figure A1: Variation in RM-Level Rates of Predicting that the Treatment Will Work



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

[1] This aligns with Young (2019), who finds the largest differences arise in RCTs with multiple treatments or few clusters. Our RCT has one treatment and 145 clusters, so the similarity is unsurprising.

Figure A2: Summary of Surveys Within the Partner Firm



Notes: This figure summarizes the surveys conducted within the partner firm. For each survey, we list the sample size of people who responded ("N") and the response rate. Appendix C provides the text of survey questions analyzed in the paper.

The *Pre-RCT in-depth interviews* are discussed in Section 2 of the paper. They measure how much time workers and store managers spend on checklists, as well as how much value they perceive in all of the checklists. These interviews were conducted verbally and in-person in the stores.

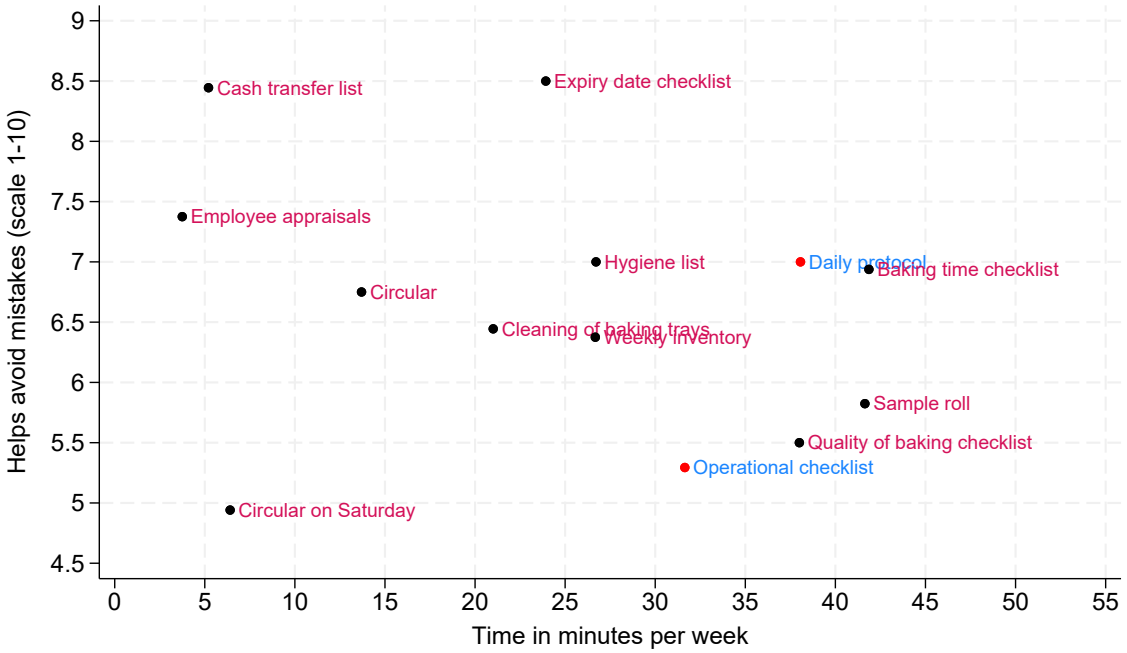
The *Pre-RCT survey of regional managers* collects regional manager predictions about whether the treatment will work for each store they supervise. This was a one question survey, and is discussed further in Appendix B.4.

The *Pre-RCT survey of store managers* collects information from store managers from before the RCT.

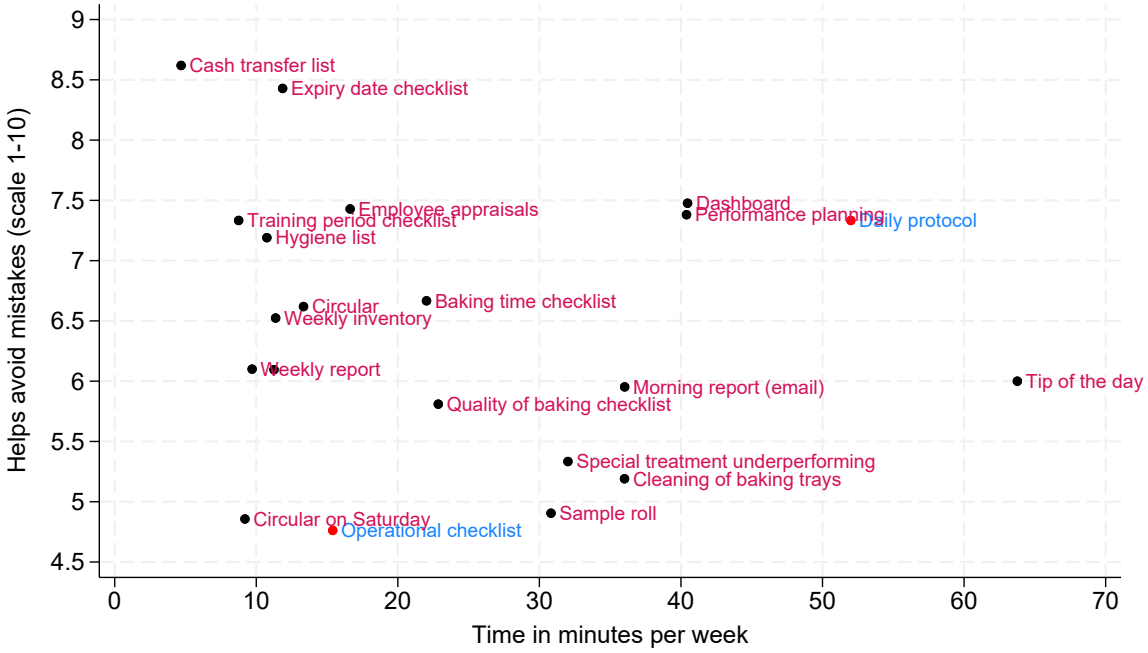
The *During-RCT survey of workers* measures workers' attitudes toward the firm. Results from this survey are presented in Panel A of Table 3. The response rate is calculated based on 395 responses from roughly 1100 workers contacted. Our paper only analyzes regular employees. This survey is discussed further in Appendix B.9.

The *During-RCT survey of store managers* was performed using store managers during the RCT.

Figure A3: Variation Across Checklists in Time per Week and Help in Avoiding Mistakes



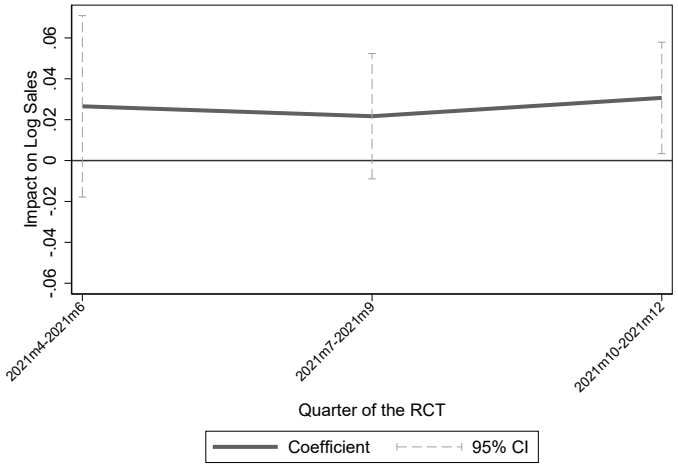
(a) Workers (N=18 workers)



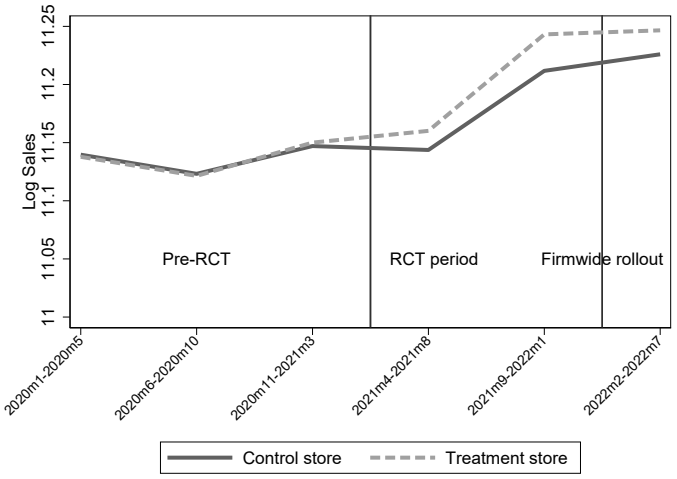
(b) Store Managers (N=21 store managers)

Notes: This figure is similar to Figure 1 in the main text, but focuses on checklists' perceived help in avoiding mistakes (instead of help in obtaining goals). Help in avoiding mistakes is measured using: "The checklist helps (FIRM) avoid mistakes." This figure uses data from the in-depth, pre-RCT interviews of 18 workers and 21 store managers described in Section 2.

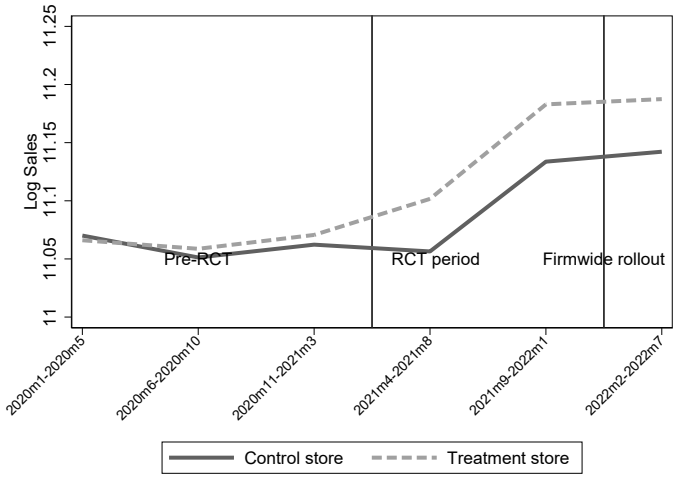
Figure A4: Further Figures Summarizing the Effect on Sales



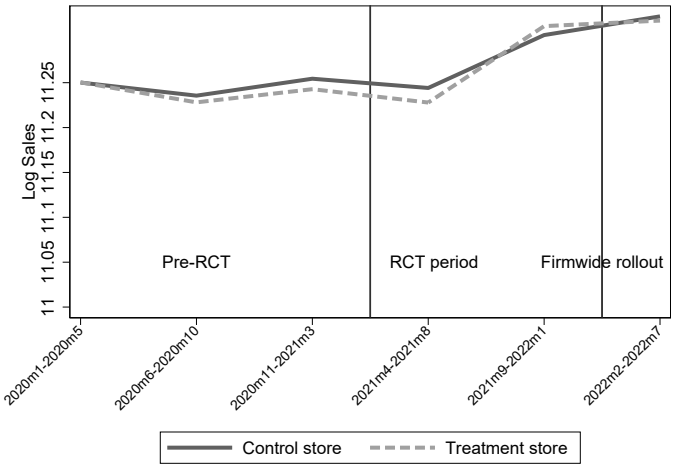
(a) Effects on Sales by Quarters of the RCT



(b) Differences Between Treatment and Control Stores Shown Using Two Lines



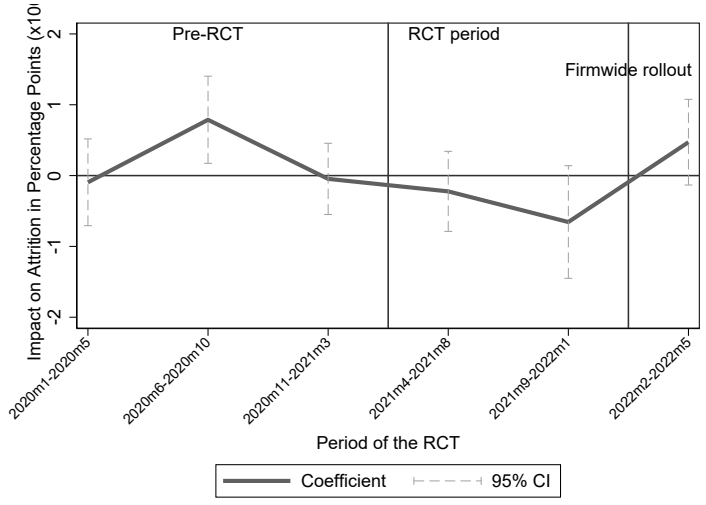
(c) Two Lines, Stores where RCT Predicted to Work



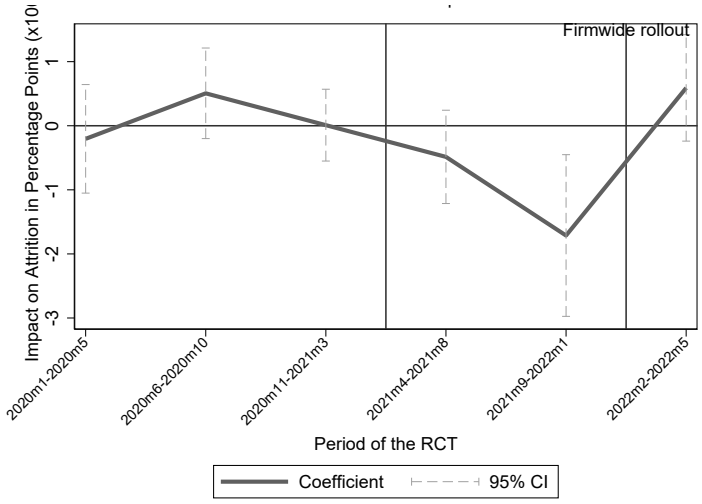
(d) Two Lines, Stores where RCT Predicted Not to Work

Notes: This figure provides robustness for our effects on sales. Panel (a) here is similar to the "RCT Period" in panel (a) of Figure 5, but shows effects using quarter of the RCT instead of 5-month periods. Panels (b)-(d) here compare treatment versus control stores with two separate lines. The control line shows average sales in control stores, while the treatment line shows regression-adjusted means (control averages plus the estimated treatment effect in each period). 95% confidence intervals based on conventional clustering by store.

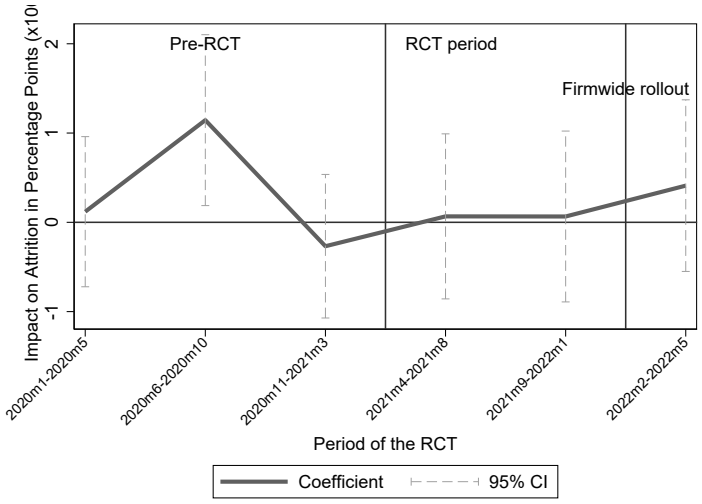
Figure A5: Differences Between Treatment and Control Stores Over Time in Trained Worker Attrition



(a) All Stores



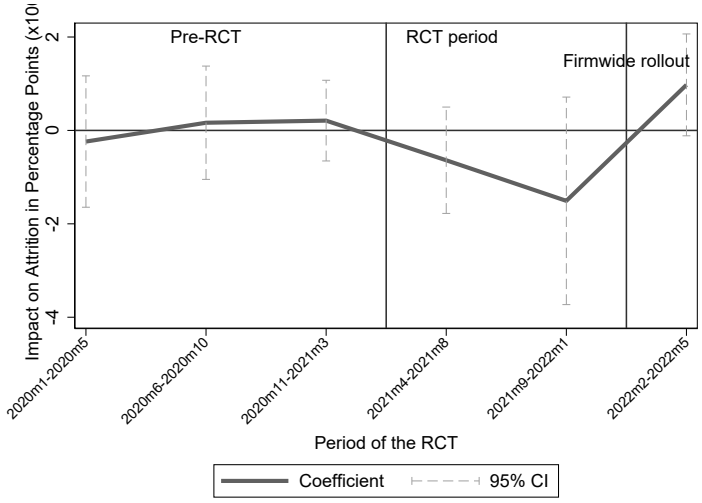
(b) Stores Where RCT Predicted to Work by Reg. Mgrs.



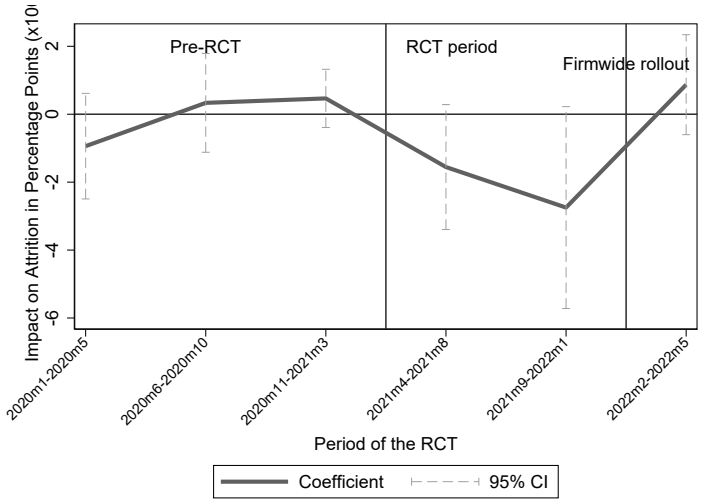
(c) Stores Where RCT Predicted Not to Work by Reg. Mgrs.

Notes: Panel (a) is similar to that in column 3 of Panel B of Table 2, but we split separately by 5-month period of the RCT. Likewise, panels (b) and (c) here are similar to column 3 of Table 5. 95% confidence intervals based on conventional clustering by store.

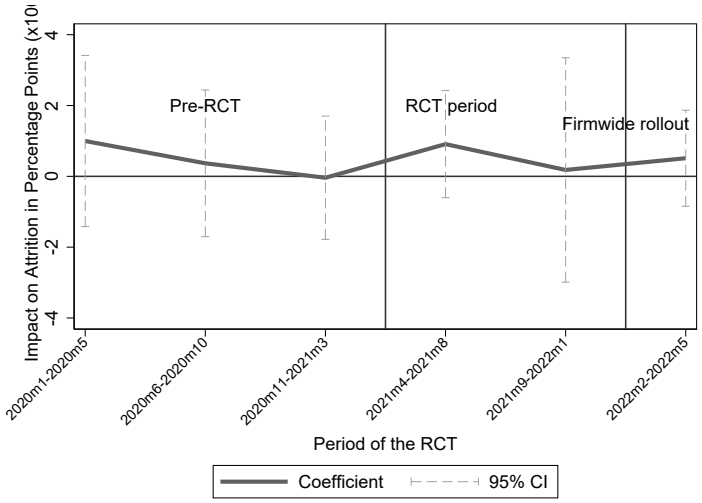
Figure A6: Differences Between Treatment and Control Stores Over Time in Store Manager Attrition



(a) All Stores



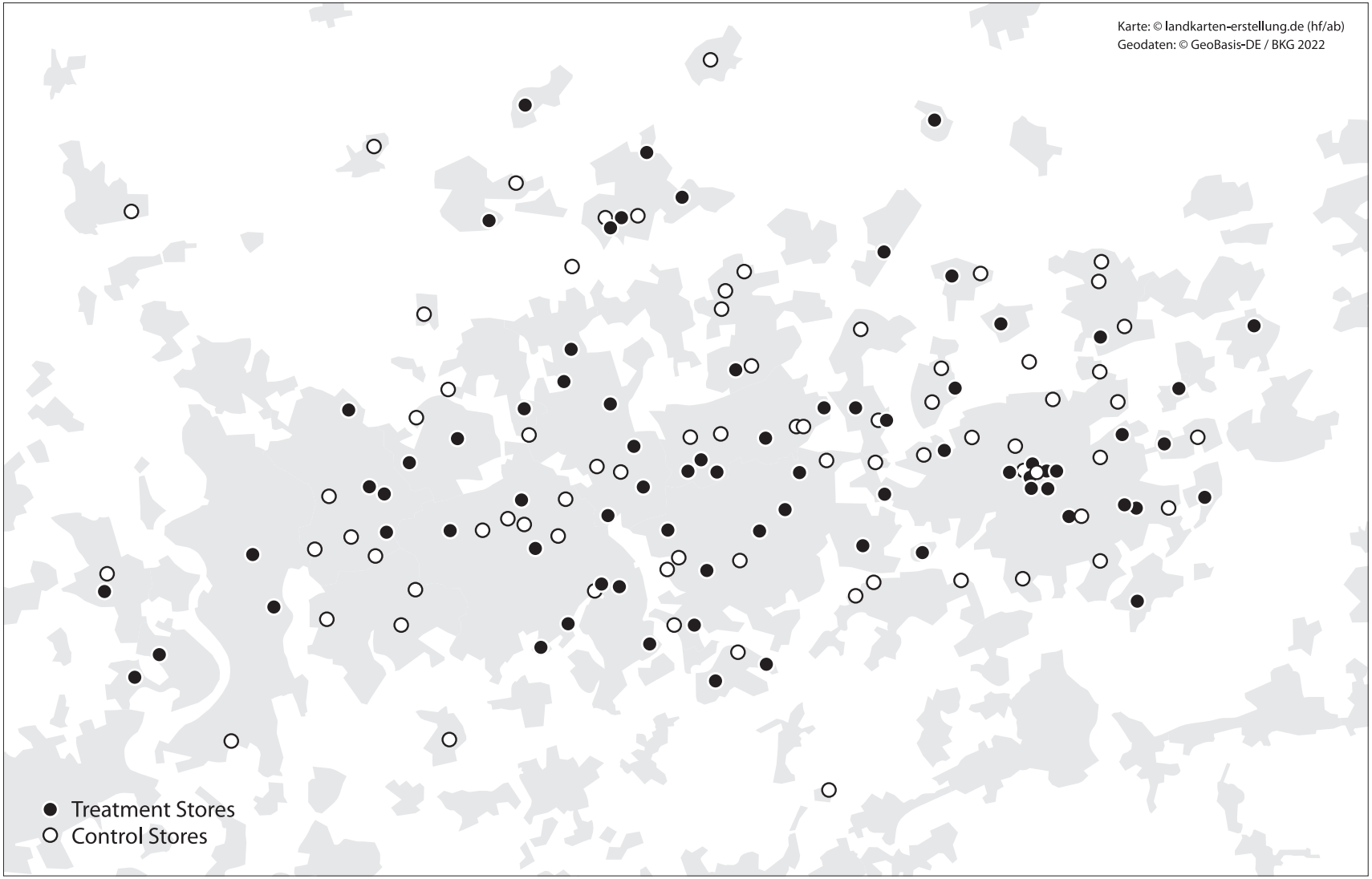
(b) Stores Where RCT Predicted to Work by Reg. Mgrs.



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

Notes: Panel (a) is similar to that in column 5 of Panel B of Table 2, but we split separately by 5-month period of the RCT. Likewise, panels (b) and (c) here are similar to column 5 of Table 5. 95% confidence intervals based on conventional clustering by store.

Figure A7: Location of Treatment and Control Stores



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

Table A1: Robustness: Simple ANCOVA (i.e., Do Not Control for Strata Characteristics or Strata Dummies)

Dep. var.:	Log Sales	Log Busy Sales	Log Slow Sales	Log Custo-mers	Log Shrink-age	Mystery Shopping Score (normed)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Store Outcomes, All Stores						
Treatment	0.028*	0.028*	0.034*	0.024	0.000	0.022
	(0.016)	(0.015)	(0.020)	(0.015)	(0.017)	(0.073)
	[0.055]	[0.053]	[0.084]	[0.117]	[0.997]	[0.775]
Panel B: Store Outcomes, Stores Where RCT Predicted to Work						
Treatment	0.053**	0.051**	0.060***	0.049**	-0.033	0.062
	(0.021)	(0.021)	(0.023)	(0.020)	(0.022)	(0.086)
	[0.003]	[0.008]	[0.004]	[0.002]	[0.146]	[0.457]
Panel C: Store Outcomes, Stores Where RCT Not Predicted to Work						
Treatment	-0.001	-0.000	0.009	-0.005	0.034	-0.020
	(0.021)	(0.020)	(0.031)	(0.021)	(0.023)	(0.119)
	[0.974]	[0.984]	[0.803]	[0.836]	[0.144]	[0.867]
1-sided p-val: predicted to work vs. not	0.04	0.04	0.09	0.03	0.02	0.29
	[0.03]	[0.03]	[0.10]	[0.03]	[0.02]	[0.29]
2-sided p-val: predicted to work vs. not	0.07	0.08	0.19	0.06	0.04	0.57
	[0.07]	[0.07]	[0.20]	[0.06]	[0.03]	[0.58]

Panel D: Worker Turnover, All Stores	(1)	(2)	(3)	(4)	(5)
Sample of workers:	All	Untrained Workers	Trained workers	Trained Non-Mgrs	Trained Managers
Treatment	-0.05	0.47	-0.43*	-0.26	-1.06*
	(0.26)	(0.43)	(0.24)	(0.27)	(0.58)
	[0.850]	[0.263]	[0.077]	[0.327]	[0.081]
Panel E: Worker Turnover, Stores Where RCT Predicted to Work					
Treatment	-0.44	0.14	-0.97***	-0.67*	-1.97**
	(0.34)	(0.57)	(0.35)	(0.38)	(0.81)
	[0.215]	[0.769]	[0.006]	[0.071]	[0.010]
Panel F: Worker Turnover, Stores Where RCT Not Predicted to Work					
Treatment	0.36	0.82	0.14	0.15	0.18
	(0.40)	(0.66)	(0.34)	(0.37)	(0.82)
	[0.402]	[0.203]	[0.696]	[0.665]	[0.976]
1-sided p-val: predicted to work vs. not	0.07	0.22	0.01	0.06	0.03
	[0.07]	[0.22]	[0.01]	[0.05]	[0.03]
2-sided p-val: predicted to work vs. not	0.13	0.43	0.02	0.13	0.06
	[0.14]	[0.43]	[0.03]	[0.11]	[0.06]

Notes: Standard errors clustered by store are in parentheses. "Rand-t" randomization inference p-values following Young (2019) in square brackets (1,000 replications). Panels A-C here are similar to the analyses of Panel A of Table 2 and to Table 4. Panels D-F here are similar to Panel B of Table 2 and to Table 5. The difference from these tables is that we run simple ANCOVA, i.e., we don't control for strata characteristics or year-month dummies. Observation counts are the same as in the main text. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A2: Robustness: Include Strata Dummies

Dep. var.:	Log Sales	Log Busy Sales	Log Slow Sales	Log Custo-mers	Log Shrink-age	Mystery Shopping Score (normed)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Store Outcomes, All Stores						
Treatment	0.024	0.023	0.032	0.018	0.018	-0.022
	(0.016)	(0.016)	(0.020)	(0.015)	(0.013)	(0.067)
	[0.178]	[0.173]	[0.126]	[0.283]	[0.199]	[0.765]
Panel B: Store Outcomes, Stores Where RCT Predicted to Work						
Treatment	0.051*	0.048*	0.059*	0.046	-0.023	0.018
	(0.028)	(0.028)	(0.031)	(0.028)	(0.021)	(0.091)
	[0.078]	[0.094]	[0.06]	[0.118]	[0.295]	[0.847]
Panel C: Store Outcomes, Stores Where RCT Not Predicted to Work						
Treatment	0.004	0.008	0.005	0.006	0.035**	-0.098
	(0.021)	(0.020)	(0.030)	(0.021)	(0.014)	(0.105)
	[0.876]	[0.708]	[0.879]	[0.802]	[0.02]	[0.371]
1-sided p-val: predicted to work vs. not	0.09	0.12	0.11	0.13	0.01	0.20
	[0.11]	[0.14]	[0.12]	[0.15]	[0.01]	[0.21]
2-sided p-val: predicted to work vs. not	0.18	0.24	0.21	0.26	0.02	0.41
	[0.22]	[0.28]	[0.25]	[0.30]	[0.03]	[0.42]

Panel D: Worker Turnover, All Stores	(1)	(2)	(3)	(4)	(5)
Sample of workers:	All	Untrained Workers	Trained workers	Trained Non-Mgrs	Trained Managers
Treatment	0.18	0.78**	-0.29	-0.07	-1.09
	(0.22)	(0.35)	(0.28)	(0.29)	(0.76)
	[0.447]	[0.025]	[0.293]	[0.801]	[0.148]
Panel E: Worker Turnover, Stores Where RCT Predicted to Work					
Treatment	-0.31	0.60	-1.31**	-1.11*	-2.65*
	(0.40)	(0.61)	(0.62)	(0.60)	(1.43)
	[0.439]	[0.301]	[0.033]	[0.058]	[0.067]
Panel F: Worker Turnover, Stores Where RCT Not Predicted to Work					
Treatment	0.85**	1.39***	0.16	0.06	1.11
	(0.33)	(0.48)	(0.42)	(0.44)	(1.21)
	[0.014]	[0.008]	[0.697]	[0.886]	[0.336]
1-sided p-val: predicted to work vs. not	0.01	0.15	0.02	0.06	0.02
	[0.01]	[0.15]	[0.02]	[0.05]	[0.02]
2-sided p-val: predicted to work vs. not	0.03	0.310	0.05	0.12	0.05
	[0.03]	[0.30]	[0.04]	[0.11]	[0.05]

Notes: Standard errors clustered by store are in parentheses. "Rand-t" randomization inference p-values following Young (2019) in square brackets (1,000 replications). Panels A-C here are similar to the analyses of Panel A of Table 2 and to Table 4. Panels D-F here are similar to Panel B of Table 2 and to Table 5. The difference from these tables is that we include strata dummies. Observation counts are the same as in the main text. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A3: Robustness: Covariates Selected Using Post-double Selection LASSO

Dep. var.:	Log Sales	Log Busy Sales	Log Slow Sales	Log Custo-mers	Log Shrink-age	Mystery Shopping Score (normed)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Store Outcomes, All Stores						
Treatment	0.027*	0.026*	0.036*	0.023	0.001	-0.001
	(0.015)	(0.014)	(0.019)	(0.015)	(0.016)	(0.070)
Panel B: Store Outcomes, Stores Where RCT Predicted to Work						
Treatment	0.053***	0.051***	0.062***	0.048**	-0.026	0.045
	(0.019)	(0.020)	(0.021)	(0.019)	(0.021)	(0.085)
Panel C: Store Outcomes, Stores Where RCT Not Predicted to Work						
Treatment	-0.001	-0.000	0.008	-0.003	0.024	-0.045
	(0.021)	(0.020)	(0.029)	(0.022)	(0.021)	(0.112)
1-sided p-val: predicted to work vs. not	0.03	0.04	0.06	0.04	0.05	0.30
2-sided p-val: predicted to work vs. not	0.06	0.07	0.12	0.08	0.10	0.60

Panel D: Worker Turnover, All Stores	(1)	(2)	(3)	(4)	(5)
Sample of workers:	All	Untrained Workers	Trained workers	Trained Non-Mgrs	Trained Managers
Treatment	0.08	0.63	-0.43*	-0.19	-1.04*
	(0.24)	(0.39)	(0.24)	(0.26)	(0.57)
Panel E: Worker Turnover, Stores Where RCT Predicted to Work					
Treatment	-0.34	0.31	-0.93***	-0.61	-1.97**
	(0.32)	(0.56)	(0.35)	(0.38)	(0.80)
Panel F: Worker Turnover, Stores Where RCT Not Predicted to Work					
Treatment	0.57*	1.09**	0.14	0.23	0.18
	(0.34)	(0.54)	(0.34)	(0.35)	(0.81)
1-sided p-val: predicted to work vs. not	0.03	0.17	0.01	0.06	0.03
2-sided p-val: predicted to work vs. not	0.06	0.34	0.02	0.12	0.06

Notes: Standard errors clustered by store are in parentheses. Panels A–C correspond to the analyses in Panel A of Table 2 and Table 4; Panels D–F correspond to Panel B of Table 2 and Table 5. The key difference is that control variables are selected using Post-Double Selection LASSO (Belloni *et al.*, 2014), implemented in Stata via the pdslasso command. Each regression begins with the controls from Table 2, and selects covariates via LASSO. Treatment and the pre-RCT mean of the dependent variable are included as fixed regressors in all specifications. For the p-values comparing treatment effects between stores where the treatment is predicted to work or not, we additionally include the RM prediction and its interaction with the pre-RCT mean of the dependent variable as fixed regressors, but do not interact RM predictions with the full set of controls; this is done to avoid a Stata error when using pdslasso. Penalty levels are selected using the theoretical formula from Belloni *et al.* (2014), with heteroskedastic loadings and the default penalty grid. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A4: Impacts of the Treatment on Individual Components of the Mystery Shopping Score

Dep. var.: (normed)	Name badge	Sales procedure	Product presentation	Free sample	Advertising	Customer interaction	Sales questions	Upsell	Golden roll	Other roll	Store appearance
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Panel A: All Stores											
Treatment	0.002	-0.009	0.056	0.000	-0.028	-0.008	0.001	0.030	-0.045	0.026	0.056
	(0.062)	(0.080)	(0.062)	(0.000)	(0.057)	(0.070)	(0.003)	(0.029)	(0.071)	(0.053)	(0.055)
Observations	1,161	1,161	1,161	1,161	1,161	1,161	1,161	1,161	1,161	1,161	1,161
Stores	144	144	144	144	144	144	144	144	144	144	144
Panel B: Stores Where RCT Predicted to Work by Regional Mgrs											
Treatment	0.157**	-0.071	0.076	0.000	-0.037	-0.121	0.000	0.048	0.157*	-0.015	0.049
	(0.079)	(0.124)	(0.078)	(0.000)	(0.076)	(0.095)	(0.000)	(0.034)	(0.092)	(0.078)	(0.072)
Observations	597	597	597	597	597	597	597	597	597	597	597
Stores	75	75	75	75	75	75	75	75	75	75	75
Panel C: Stores Where RCT Not Predicted to Work by Regional Mgrs											
Treatment	-0.137	0.022	0.061	0.000	-0.034	0.128	0.006	-0.003	-0.206*	0.053	0.057
	(0.098)	(0.107)	(0.091)	(0.000)	(0.086)	(0.094)	(0.005)	(0.038)	(0.104)	(0.062)	(0.079)
Observations	564	564	564	564	564	564	564	564	564	564	564
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 instead of the overall score. Each component score is normalized. "Name badge" measures whether an employee shows their name badge. "Sales procedure" rates the quality of workers' sales procedures, such as saying good morning. "Product presentation" rates the quality of the way in which products are presented. "Free sample" measures whether the free sample is present, and has a standard error of 0 since the unnormalized outcome always equals 1 in our data period. "Advertising" measures whether the correct advertising is being carried out in the store. "Customer interaction" measures the quality of customer interaction, i.e., how friendly are workers to customers. "Sales questions" measures whether workers are able to answer questions about the product. "Upsell" measures whether employees did an upselling. "Golden roll" measures the overall quality and presentation of the golden rolls. Golden rolls are small, crusty wheat-based bread rolls. Often consumed at breakfast or as a sandwich base, they are considered a staple in German bakeries. "Other roll" measures the quality and presentation of rolls besides the golden rolls. "Store appearance" measures the quality of a store's appearance. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A5: Predicting Response to the *During-RCT Worker Survey*

	(1)
Treatment	0.031
	(0.052)
Log sales total	-0.074
	(0.084)
Female	0.725**
	(0.293)
Age at time of survey	0.019**
	(0.007)
Worker tenure in years	-0.006
	(0.011)
Observations (stores)	144

Notes: This table predicts variation in response rates across stores to the *During-RCT worker survey*. For each store, we regress the store-level response rate on various store-level characteristics. Robust standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A6: Comparing Trained vs. Untrained Workers in Mean Characteristics

	Untrained	Trained	Trained Non-mgr	Trained Manager
Female	.95	.99	.99	1
Age	36.7	42.37	41.51	45.66
Base hourly wage in euros	10.44	13.88	13.53	15.24
Monthly bonus in euros	15.05	45.21	27.85	112.01
Total monthly pay in euros	1337	1877	1754	2347
Tenure in yrs	4.84	11.64	10.84	14.7
Tenure of 1yr or less	.19	.05	.06	0
Tenure of 1-2yrs	.19	.06	.07	.01
Tenure of 2-5yrs	.29	.11	.14	.01
Tenure of 5-10yrs	.16	.26	.25	.29
Tenure more than 10yrs	.18	.52	.47	.68
N	654	698	554	144

Notes: This table compares workers of different types using data from March 2021, which is the month before the RCT began. The N in the last row is the number of workers of each type.

Table A7: Robustness: Cox Models for Employee Attrition

Panel A: All Stores	(1)	(2)	(3)	(4)	(5)
Sample of workers:	All	Untrained Workers	Trained workers	Trained Non-Mgrs	Trained Managers
Treatment	0.03	0.19	-0.49*	-0.37	-1.33**
	(0.12)	(0.13)	(0.25)	(0.26)	(0.62)
Observations	13,271	6,489	6,782	5,403	1,379
Workers	1637	863	774	624	150
Panel B: Stores Where RCT Predicted to Work					
Treatment	-0.21	0.04	-1.08**	-0.80*	-2.33**
	(0.17)	(0.19)	(0.43)	(0.46)	(0.93)
Observations	6,595	3,126	3,469	2,691	778
Workers	829	422	407	320	87
Panel C: Stores Where RCT Not Predicted to Work					
Treatment	0.22	0.33**	-0.05	-0.03	Convergence
	(0.15)	(0.15)	(0.36)	(0.36)	issue
Observations	6,676	3,363	3,313	2,712	601
Workers	878	483	395	328	67

Notes: This table is a robustness check to our main analyses on employee attrition (Panel B of Table 2, as well as Table 5). The difference is we analyze Cox proportional hazard models instead of linear probability models. The failure event is whether an employee attrites in a given month and we show coefficients (not odds ratios). For example, the coefficient of -0.49 in column 3 of Panel A means that the treatment reduced trained worker attrition by 39% (i.e., exp(-0.49)-1 = -0.39), which is similar to the 35% reduction in Panel B of Table 2. The controls are the same as in our main analyses on employee attrition, except (1) tenure is controlled for non-parametrically via the Cox model (instead of with a quadratic) and (2) there are no calendar time controls. In column 5 of Panel C, the model experiences convergence issues, reflecting that the number of attrition events for store managers in stores where the treatment is predicted not to work is small. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A8: Further Analysis of Google Reviews

Panel A: Robustness to Panel B of Table 3, Only Reviews with Text

Dep. var.: Whether there is a positive comment regarding:	(1) The product	(2) Service	(3) Shop appearance	(4) Speed of service	(5) Value for money	(6) Product availability
Treatment	-0.006	0.019	0.025**	0.023***	0.003	0.010
	(0.031)	(0.028)	(0.011)	(0.007)	(0.011)	(0.016)
	[0.852]	[0.512]	[0.024]	[0.001]	[0.775]	[0.505]
Observations	855	855	855	855	855	855
Stores	138	138	138	138	138	138
Mean DV if Treat=0	0.546	0.354	0.0276	0.0130	0.0361	0.101

Panel B: Google Review Scores

Dep. var.:	(1) Average rating	(2) Share 1s	(3) Share 2s	(4) Share 3s	(5) Share 4s	(6) Share 5s	(7) Number of ratings
Treatment	0.030	-0.002	-0.005	0.004	0.011	-0.003	0.434
	(0.068)	(0.018)	(0.008)	(0.010)	(0.018)	(0.023)	(0.419)
	[0.675]	[0.891]	[0.547]	[0.687]	[0.529]	[0.905]	[0.297]
Observations	1,023	1,023	1,023	1,023	1,023	1,023	1,023
Stores	142	142	142	142	142	142	142
Mean DV if Treat=0	4.234	0.0802	0.0317	0.0658	0.218	0.604	3.848

Main notes: Standard errors clustered by store in parentheses. "Rand-t" randomization inference p-values following Young (2019) in square brackets (1,000 replications). Stars are based on clustered standard errors in parentheses, with * significant at 10%; ** significant at 5%; *** significant at 1%

Panel A: This panel presents a robustness check for Panel B of Table 3, restricting the sample to Google reviews that contain text. The number of stores is smaller here than in Panel B of Table 3, as some stores only have reviews without text during the RCT period.

Panels B: An observation is a store-month during the RCT. Columns 1–6 show that the treatment has no significant effect on the quantitative score in Google reviews. Column 7 shows the treatment has no effect on the number of ratings that a store receives. All regressions control for the pre-RCT mean of the dependent variable, year-month fixed effects, and 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 A9: Examining Alternative Explanations for Larger Effects in Stores where RMs Predict Treatment to Work

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Sales							
Treatment X Predict success	0.055**	0.053*	0.051*	0.075***	0.043*	0.059**	
	(0.027)	(0.030)	(0.028)	(0.027)	(0.025)	(0.027)	
Treat*(Pre-RCT Log Sales)	-0.051					-0.120	-0.091
	(0.052)					(0.162)	(0.160)
Treat*(Pre-RCT mean head count)		-0.002				0.001	-0.001
		(0.004)				(0.006)	(0.006)
Treat*(Pre-RCT mean tenure of workers)			-0.001			-0.007	-0.007
			(0.005)			(0.006)	(0.006)
Treat*(Pre-RCT mystery shopping score)				-0.087**		-0.073**	-0.057
				(0.037)		(0.036)	(0.037)
Treat*Pre-RCT Log (Shrinkage as % of Sales)					0.127	0.003	0.047
					(0.094)	(0.156)	(0.155)
1-sided p-val, Treat X Predict success	0.02	0.04	0.04	0	0.04	0.01	NA
2-sided p-val, Treat X Predict success	0.04	0.08	0.07	0.01	0.09	0.03	NA
Panel B: Trained Worker Attrition							
Treatment X Predict success	-1.042**	-0.904*	-1.114**	-1.098*	-1.158**	-0.912	
	(0.517)	(0.532)	(0.501)	(0.561)	(0.556)	(0.557)	
Treat*(Pre-RCT turnover rate)	0.191					0.147	0.214
	(0.394)					(0.434)	(0.460)
Treat*(Pre-RCT mean head count)		0.049				0.017	0.019
		(0.054)				(0.059)	(0.057)
Treat*(Pre-RCT mean tenure of workers)			-0.208**			-0.102	-0.127
			(0.103)			(0.140)	(0.142)
Treat*(Pre-RCT mystery shopping score)				-0.275		-0.736	-0.826
				(0.646)		(0.746)	(0.677)
Treat*Pre-RCT Log (Shrinkage as % of Sales)					-0.207	0.383	-0.563
					(2.036)	(2.548)	(2.742)
1-sided p-val, Treat X Predict success	0.02	0.05	0.01	0.03	0.02	0.05	NA
2-sided p-val, Treat X Predict success	0.05	0.09	0.03	0.05	0.04	0.10	NA

Notes: This table accompanies the discussion in Section 4. It displays how key interaction term coefficient varies as we include regressors for an additional characteristic, as well as the interaction of treatment times the characteristic. RM is an abbreviation for regional manager. "Predict success" is a dummy for whether an RM predicts the treatment will work in a given store. As in Panel C of Tables 4-5, each heterogeneity dimension is fully interacted with control variables. Stars are based on two-sided p-values, with * significant at 10%; ** significant at 5%; *** significant at 1%

Table A10: Applying Machine Learning Toward Understanding Treatment Heterogeneity: RM Predictions by Quartile of Affected Stores

	Sorted effects	Random forests	Share of stores in common between two methods
	(1)	(2)	(3)
Panel A: Sales			
Q1 (least affected)	0.114	0.307	0.785
	(0.319)	(0.462)	
Q2	0.389	0.447	0.705
	(0.488)	(0.498)	
Q3	0.754	0.570	0.656
	(0.431)	(0.496)	
Q4 (most affected)	0.811	0.756	0.806
	(0.392)	(0.430)	
Unadjusted p-value equal	0.000	0.000	
WY p-value equal	0.000	0.001	
Panel B: Trained Worker Attrition			
Q1 (least affected)	0.000	0.095	0.875
	(0.000)	(0.293)	
Q2	0.439	0.509	0.671
	(0.496)	(0.500)	
Q3	0.716	0.672	0.683
	(0.451)	(0.470)	
Q4 (most affected)	0.894	0.772	0.792
	(0.308)	(0.420)	
Unadjusted p-value equal	0.000	0.000	
WY p-value equal	0.000	0.000	

Notes: This table applies two machine-learning approaches—sorted effects and causal random forests—to estimate store-level conditional treatment effects (CATEs). Each method yields a CATE estimate for every store. We rank stores by these estimates and assign them to quartiles, where Q1 contains the least affected stores and Q4 contains the most affected stores (for trained worker attrition, "most affected" refers to the strongest negative effect). For each quartile, we report the mean of a binary indicator for whether RMs predict the treatment will work in that store—that is, the share of stores for which RMs predict the treatment to work. Column 3 reports the share of stores assigned to the same quartile by both methods. We also report p-values of the mean-equality tests across quartiles, both unadjusted and Wesfall-Young (1993)-adjusted for multiple hypothesis testing. Standard errors are clustered by store. For illustration, in column 1 (sorted effects for sales), only 11% of least-affected stores (Q1) are predicted by RMs to benefit, compared to 81% of the most-affected stores (Q4).

Table A11: Heterogeneity in the Predictiveness of RM Beliefs Based on RM Tenure

Panel A: Store Outcomes	(1)	(2)	(3)	(4)	(5)	(6)
Dep. var.:	Log Sales	Log Busy Sales	Log Slow Sales	Log Customers	Log Shrink -age	Mystery Shopping Score (normed)
Treatment X Predicted to Work	-0.094	-0.099	-0.097*	-0.088	0.079**	-0.061
	(0.070)	(0.080)	(0.049)	(0.069)	(0.029)	(0.229)
	[0.374]	[0.386]	[0.024]	[0.416]	[0.069]	[0.819]
Treatment X Predicted to Work X Seasoned RM	0.200**	0.203**	0.203**	0.189**	-0.171***	0.268
	(0.088)	(0.092)	(0.094)	(0.084)	(0.043)	(0.277)
	[0.035]	[0.017]	[0.041]	[0.034]	[0.009]	[0.439]

Panel B: Worker Turnover	(1)	(2)	(3)	(4)	(5)
Sample of workers:	All	Untrained Workers	Trained workers	Trained Non-Mgrs	Trained Managers
Treatment X Predicted to Work	-0.006	0.446	-0.812	-1.021	0.262
	(0.872)	(1.445)	(0.640)	(0.676)	(2.960)
	[0.986]	[0.386]	[0.242]	[0.240]	[0.923]
Treatment X Predicted to Work X Seasoned RM	-1.314	-1.723	-0.420	0.360	-4.312
	(1.026)	(1.641)	(1.055)	(1.253)	(3.109)
	[0.440]	[0.498]	[0.768]	[0.820]	[0.316]

Notes: Standard errors clustered by RM are in parentheses. Since there are only 15 RMs and thus only 15 clusters, we also present wild bootstrap p-values (calculated using "wildbootstrap" with 5,000 replications in Stata 19) in brackets. Stars are based on clustered standard errors in parentheses, with * significant at 10%; ** at 5%; *** at 1%

The regressions here analyze the triple interaction "Treatment X Predicted to Work X Seasoned RM." As such, they include all singleton and double interaction terms. For Panel A here, we also include all the control variables used in Panel C of Table 4. For Panel B here, we also include all the control variables used in Panel C of Table 5. Observation counts are the same as in Table 2, with an observation being a store-month in Panel A and being a worker-month in Panel B. As noted in Section 4 in the main text, "Seasoned RM" is a dummy variable referring to an RM who was hired into the role before 2019. RM tenure is absent from the personnel data, but we gathered it manually using web searches, the most important source being Xing (a German network similar to LinkedIn).

As seen in Panel A, the predictiveness of RM predictions (measured by interaction term Treatment X Predicted to Work) is statistically significantly larger for seasoned RMs compared to novice RMs for all store outcomes except for mystery shopping. As seen in Panel B, the predictiveness of RM predictions for attrition outcomes tends to be greater for seasoned RMs, but the differences are not statistically significant.

Table A12: Regional Manager Predictions: Regional Managers 1-4

Store	Yes	Prediction
Regional Manager 1		
1	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.
2	0	Unclear: Some employees are happy about fewer guidelines, others need strict rules.
3	0	Will have a positive impact on employee satisfaction; but: poor communication of initiative by store manager expected; might have negative impact on sales.
4	1	Great, well-coordinated team in the store; everything fits in the store; would appreciate less bureaucracy.
5	0	Unclear: Employees will be happy, but you have to take individual employees by the hand from time to time and tell them what they should do.
6	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!
7	0	Negative effects, as the team is still very fresh; new manager in place; processes not yet internalised; negative sales.
8	0	Unclear. Employees will be glad; mixed team with some old and many young employees.
9	1	Would perhaps miss the list; but: no negative consequences in the store; on the contrary: positive impact!
Regional Manager 2		
10	0	Will be glad; but: implementation of processes not secure; chaotic store; internal evaluations (e.g., strawberries on a cake) are usually negative. Chaos may result without clear guidelines.
11	1	Would implement this very well; would also get along well without paper and clear structure; employee satisfaction will increase.
12	0	Many new staff members; store is a bit chaotic; need structure and guidance; want guidance.
13	1	Get along without bureaucracy; would feel more comfortable if there was less pressure because of less bureaucracy. Will work.
14	1	Get along without bureaucracy; nothing would change in the operational processes without bureaucracy; staff already understood important things.
15	0	Mixed picture; have too high return rates on baked goods; returns will get worse. Unclear how it will work.
16	1	Get along without bureaucracy; nothing would change. Therefore, will work.
17	0	Need structure; will not work without it; otherwise the store will sink into chaos and lose focus.
18	0	Need structure; haven't been around long; bureaucracy is important support; return rates for baked goods are poor.
19	0	Need structure and bureaucracy; otherwise staff will have problems.
Regional Manager 3		
20	1	Yes, will work.
21	1	Yes, will work.
22	1	Yes, will work.
23	0	No, will not work.
24	1	Yes, will work.
25	0	No, will not work.
26	0	No, will not work.
27	1	Yes, will work. Clear yes.
28	0	No, will not work. No way.
29	0	No, will not work.
Regional Manager 4		
30	1	Will work. Good and organized store manager; very conscientious and tidy. Implementation will work.
31	0	Need assistance. Complicated without lists; young store manager; young team needs guidance.
32	0	Undecided. Maintain documentation obligations, as other structure is difficult to implement; old store manager, who wants to maintain habits.
33	1	Store team does not need lists. Committed, thoughtful and conscientious.
34	0	Store desperately needs structure which is provided by bureaucracy; organized store manager; bad team. Will not work without lists.
35	0	Good leadership; bad team. Would work partially.
36	0	Would be good if lists remained. Recent change of store manager. Large store.
37	1	Would work. Complete confidence in the team.
38	1	No documentation requirements needed. Good team. Good store.
39	1	No documentation requirements needed. Good team and store manager. Well organized.

Notes: This table gives the predictions of several regional managers. The predictions here are notes that a coauthor wrote down in pen form during the phone calls with regional managers. Due to local norms on recording phone calls in Germany, it was not feasible to record the phone calls. Appendix B.4 gives further details and discussion on the elicitation and classification of regional manager predictions.

Table A13: Regional Manager Predictions: Regional Managers 5-8

Store	Yes	Prediction
Regional Manager 5		
40	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 not everything is done correctly: problems.
41	1	This store doesn't run in the same way as Store 45, but will work well here too; some structure may be necessary here, but they can manage it autonomously. It will work well.
42	1	Similar to Store 41; team will be glad; actually need list to get routine; would also work out without list.
43	1	Will work out without any requirements; team is confident in their performance; happy if there are no lists.
44	1	Like in store 41. Team will manage it, but need to stay focused. Problem: When there is a mystery shopping visit and expectations are not met, there will be trouble in the team. But will work out.
45	1	Team does not need lists. Can manage without lists. Strength in implementing processes.
46	1	No lists needed; works out without lists. However, when the store manager is not on duty, they sometimes do not meet expectations.
47	0	List needed for orientation. Does not work without it.
48	1	Definitely do not need lists; will implement everything in any case.
49	1	Do well without a list.
Regional Manager 6		
50	1	In general: will work out.
51	1	Will work out.
52	1	If treated and lists are dropped, would do well and without any problems. Would potentially like to keep the daily protocol.
53	0	Focus store; cannot work without clear guidelines, may result in chaos.
54	1	There won't be any problems with less bureaucracy, even if daily protocol is important from time to time.
55	0	Focus store; cannot work without clear guidelines, may result in chaos.
56	0	Cannot work without it; cash differences.
57	1	Can do without it; store runs great.
58	1	Can do without documentation requirements; runs great, but still relatively new store manager.
59	0	Can't do without it even if they would like to do without it; large cash register and other store differences and problems with sales.
Regional Manager 7		
60	1	Will work out without checklists.
61	1	Will work out without checklists.
62	1	Will work out without checklists.
63	1	Will work out without checklists.
64	1	Will work out without checklists.
65	1	Will work out without checklists.
66	1	Will work out without checklists.
67	0	They need structure; won't work without checklists.
68	1	Will work out without checklists.
Regional Manager 8		
69	0	Staff will be glad; procedures are sometimes problematic, often not implemented; therefore bureaucracy and structure needed.
70	0	Store manager wants to maintain bureaucracy; but it could work as well. Unclear if it works out.
71	0	Store manager wants to keep bureaucracy; unclear if it works out.
72	0	Store manager wants to maintain bureaucracy; clear structures important for training and coaching. Unclear what happens.
73	0	Store manager wants to maintain bureaucracy; clear structures important for training and coaching; mixed effects.
74	0	Store manager wants to maintain bureaucracy; clear structures important for training and coaching; manager has issues with cash register discrepancies and managing personnel.
75	0	Store manager wants to maintain bureaucracy; clear structures important for training and coaching; manager has issues with cash register discrepancies; mixed effects.

Notes: Same notes as in Table A12. In the predictions of regional manager 6, "focus store" refers to a small number of stores marked by top management as needing improvement.

Table A14: Regional Manager Predictions: Regional Managers 9-12

Store	Yes	Prediction
Regional Manager 9		
76	0	Sometimes help needed; large store; operationally strong, so could also work out.
77	1	Can be left out; very strong store manager; store manager trains employees very well.
78	1	Can be left out; small store; few employees; can also be trained in person.
79	0	New store manager; old established team; store manager needs guidance; won't work in the short-run but might in the medium turn.
80	0	Not a good store manager, not good at training staff; clear guidance and lists are important.
81	1	Works out without; small store; staff are well trained and guided by store manager.
82	0	Please don't remove the lists here. Big team; some difficult cases among the employees; information does not flow well downward from the store manager.
83	1	Training on important processes is a viable alternative to checklists; control can be omitted; will work out.
84	0	New store manager; lists are needed.
85	0	New store manager; lists are needed, but store manager is probably good; best case: keep first, leave out later.
Regional Manager 10		
86	1	Independent store; will work out without lists; employee satisfaction will improve.
87	0	Downtown store; no positive or negative developments on sales or performance; high employee satisfaction anyway.
88	1	Similar to other stores that are doing well; team will be happy when lists are gone; no loss of sales (rather the opposite!); save time through less bureaucracy; will work out.
89	0	Similar to other stores: they will be personally happy when the list is gone; no loss or increase in sales; save time, but also no increase in any kpi.
90	1	If operational list is gone, it's good for the team; it will work.
91	1	Always enjoyed making lists and bureaucracy, but will also work out well without restrictions.
92	0	Always enjoyed bureaucracy. Old employees and therefore difficulties without it.
93	1	Team will be glad when operational list is gone. No problems expected. Will work out!
94	0	Rather neutral. Mixed effects. No operational list is good, more time for employees.
Regional Manager 11		
95	0	Will not be received well. Daily protocol and operational lists are popular; employees like bureaucracy.
96	0	They like bureaucracy; if you remove they'll find another way to keep bureaucracy at the store; will neither be happy nor sad; neutral effects.
97	0	Bureaucracy needed.
98	1	Will work out without.
99	1	Will work out without.
100	0	Documentation requirements are needed.
101	1	Could live without bureaucracy; very communicative store manager.
102	0	Daily protocol needed; operational list not necessarily. Therefore mixed effects.
103	0	Bureaucracy needed; will not work out without.
Regional Manager 12		
104	1	Strong store manager; high revenues store; employee satisfaction is around 50-50; store manager will smile nicely about abolishing the lists because there are so many other lists and there is a fear that more lists will be added. But abolishing the lists will work without operational problems.
105	1	Strong store manager, been there for a long time; high employee satisfaction; it will work out very well without documentation requirements.
106	0	Currently closed; strong store manager; employee satisfaction high and will improve.
107	1	Small store, on a positive trajectory; new store manager, will accept bureaucracy reduction and implement successfully. It's an opportunity!
108	0	Very strong store manager; employee satisfaction will not change. Large store. But: operational implementation will work partially, no big problems.
109	1	Strong store manager, open to everything; high employee satisfaction; omitting lists will be successful.
110	0	Small store; will take a positive view; new store manager; effects: partly positive, partly negative.
111	0	Very strong store manager; employees been there for many years. Effects unclear.
112	0	Will meet with resistance; will not accept anything new; will only reluctantly, if at all, let themselves be dragged into it; store manager communicates this way to the team. Black box. Will not work out.
113	1	Strong store manager; open to everything and can implement everything well; already been there a few years.
114	1	Employee satisfaction will improve with less bureaucracy; strong store manager; will work out.

Notes: Same notes as in Table A12.

Table A15: Regional Manager Predictions: Regional Managers 13-15

Store	Yes	Prediction
Regional Manager 13		
115	1	Interested store manager; will be happy about it; positive emotional response; higher employee satisfaction; omitting will work out.
116	1	Top motivated store manager; positive emotional response; store manager takes on many tasks themselves; less bureaucracy will be supportive.
117	1	Focus store; motivated store manager; store manager already takes over a lot of bureaucracy from employees; employee satisfaction may not necessarily improve, but overall, the store will function well without the lists.
118	0	I'm skeptical about the team at this store; employee satisfaction will not get better; will not work out.
119	1	Mini store, hardly any bureaucracy; will work out.
120	1	Mini store, hardly any bureaucracy; only 3 employees; will be happy when there is less bureaucracy.
121	1	Store manager will be happy that lists/bureaucracy are gone, but will nonetheless say it doesn't help them much. Dominant store manager; employee satisfaction will not increase, but it will work overall.
122	1	Highly motivated store team, very communicative; maybe no increase in sales or staff satisfaction, because store is already productive; will work without lists.
123	0	Old store manager; if it is up to them they will continue to run lists; no change in sales; whether or not there are checklists, store will be ok.
124	1	Great store manager, will work hard on it and implement it well; will analyze whether it is successful. Will work. Positive influence; employees very satisfied, will increase.
125	0	Employees are dissatisfied with the situation in the store; there are grumblings; relief from less bureaucracy could help, but it is unclear what happens.
Regional Manager 14		
126	1	Will work; good store and well-organized store manager.
127	0	Problem team, a bit chaotic. Won't work without guidelines and clear guidelines.
128	1	Most likely will work. Well-organized store manager, therefore also well-organized team.
129	1	Will work, even though store manager is bureaucratic and likes bureaucracy.
130	1	Store manager retiring soon. If treated, would work out, as they have a well-functioning team; unclear if open to changes, but will work out overall.
131	1	Could work, or rather: will work!!
132	0	No, will not work.
133	1	Will work. But team needs to know why.
134	0	At the moment, no. Will not work.
135	1	Yes, the employees are implementing well; they always want to understand why things change. But: If the explanation makes sense, which will be the case [for removing checklists], it will work in the store.
Regional Manager 15		
136	1	Bureaucracy costs time; more time has a positive effect on satisfaction; will work out.
137	0	Older employees; very bureaucratic; keep handwritten lists; love bureaucracy; unclear.
138	1	Less bureaucracy saves time; more time = positive for employee satisfaction; young team, easy-going.
139	1	Less bureaucracy saves time; more time = positive for satisfaction; young team; more relaxed and more free time.
140	0	Structures and control needed.
141	0	Will improve the general mood; are often overwhelmed with bureaucracy; employee satisfaction and sales will not improve.
142	0	Neutral, mixed bag.
143	0	Store manager has been there for over 20 years. Unclear what happens.
144	0	Less bureaucracy will improve the general mood; but: employee satisfaction and sale will not improve. Unclear what happens.
145	0	Neutral. Unclear.

Notes: Same notes as in Table A12.

Table A16: Robustness to Table 7: Breaking Out the Components of Mystery Shopping

Specification:	Lasso-selected regressors (1)
Treatment store	
Pre-RCT Log Sales	
Pre-RCT mystery shopping: customer interaction	0.362***
	(0.133)
Pre-RCT mystery shopping: name badge	0.174
	(0.138)
Pre-RCT mystery shopping: sales procedure	
Pre-RCT mystery shopping: product presentation	
Pre-RCT mystery shopping: free sample	
Pre-RCT mystery shopping: advertising	
Pre-RCT mystery shopping: sales questions	
Pre-RCT mystery shopping: upsell	
Pre-RCT mystery shopping: golden roll	
Pre-RCT mystery shopping: other roll	
Pre-RCT mystery shopping: store appearance	
Pre-RCT mean head count	-0.024***
	(0.008)
Pre-RCT Log (Shrinkage as % of Sales)	
Pre-RCT mean tenure of workers in years	
Observations	145
R-squared	0.133

Notes: This table is a robustness check to Table 7. The regression here is similar to column 2 of Table 7, but it replaces the overall pre-RCT mystery shopping score with the 11 components of the mystery shopping score. The lasso procedure removes most of the mystery shopping components from the regression. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A17: Robustness: Excluding the One Store where the Two Classifiers of RM Predictions Didn't Agree on the Classification

Dep. var.:	Log Sales	Log Busy Sales	Log Slow Sales	Log Custo-mers	Log Shrink-age	Mystery Shopping Score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Store Outcomes, All Stores						
Treatment	0.028*	0.027*	0.036*	0.024	0.002	0.004
	(0.015)	(0.015)	(0.020)	(0.015)	(0.016)	(0.070)
Observations	1,421	1,421	1,421	1,421	1,421	1,154
Stores	144	144	144	144	144	143
Panel B: Store Outcomes, Stores Where RCT Predicted to Work						
Treatment	0.055***	0.052**	0.064***	0.050***	-0.023	0.082
	(0.020)	(0.020)	(0.022)	(0.019)	(0.021)	(0.090)
Observations	734	734	734	734	734	590
Stores	75	75	75	75	75	74
Panel C: Store Outcomes, Stores Where RCT Not Predicted to Work						
Treatment	-0.003	-0.003	0.004	-0.006	0.024	-0.068
	(0.020)	(0.019)	(0.027)	(0.020)	(0.020)	(0.109)
Observations	687	687	687	687	687	564
Stores	69	69	69	69	69	69
1-sided p-val: predicted to work vs. not	0.02	0.02	0.04	0.02	0.06	0.14
2-sided p-val: predicted to work vs. not	0.04	0.05	0.09	0.04	0.11	0.29

Panel D: Worker Turnover, All Stores	(1)	(2)	(3)	(4)	(5)
Sample of workers:	All	Untrained Workers	Trained workers	Trained Non-Mgrs	Trained Managers
Treatment	0.11	0.69*	-0.43*	-0.22	-1.07*
	(0.24)	(0.39)	(0.26)	(0.27)	(0.60)
Observations	13,197	6,449	6,748	5,369	1,379
Workers	1630	859	771	621	150
Panel E: Worker Turnover, Stores Where RCT Predicted to Work					
Treatment	-0.39	0.32	-1.03***	-0.65	-2.17**
	(0.31)	(0.51)	(0.37)	(0.39)	(0.84)
Observations	6,521	3,086	3,435	2,657	778
Workers	821	418	403	316	87
Panel F: Worker Turnover, Stores Where RCT Not Predicted to Work					
Treatment	0.48	0.98*	0.08	0.10	0.55
	(0.37)	(0.57)	(0.37)	(0.37)	(0.87)
Observations	6,676	3,363	3,313	2,712	601
Workers	878	483	395	328	67
1-sided p-val: predicted to work vs. not	0.04	0.19	0.02	0.08	0.01
2-sided p-val: predicted to work vs. not	0.07	0.39	0.04	0.17	0.03

Notes: Standard errors clustered by store are in parentheses. Panels A-C here are similar to the analyses of Panel A of Table 2 and to Table 4. Panels D-F here are similar to Panel B of Table 2 and to Table 5. The difference from these tables is that we exclude the one store (Store 113) where the second classifier classified differently from the original classifier. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A18: Robustness: Excluding Stores with Annoyance About the RCT

Dep. var.:	Log Sales	Log Busy Sales	Log Slow Sales	Log Custo-mers	Log Shrink -age	Mystery Shopping Score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Store Outcomes, All Stores						
Treatment	0.025*	0.025*	0.029	0.022	-0.000	0.019
	(0.015)	(0.015)	(0.018)	(0.015)	(0.017)	(0.074)
Observations	1,334	1,334	1,334	1,334	1,334	1,084
Stores	135	135	135	135	135	134
Panel B: Store Outcomes, Stores Where RCT Predicted to Work						
Treatment	0.054**	0.052**	0.062**	0.053**	-0.034	0.120
	(0.021)	(0.021)	(0.024)	(0.020)	(0.023)	(0.089)
Observations	697	697	697	697	697	562
Stores	71	71	71	71	71	70
Panel C: Store Outcomes, Stores Where RCT Not Predicted to Work						
Treatment	-0.005	-0.002	-0.007	-0.008	0.027	-0.047
	(0.020)	(0.020)	(0.024)	(0.020)	(0.021)	(0.119)
Observations	637	637	637	637	637	522
Stores	64	64	64	64	64	64
1-sided p-val: predicted to work vs. not	0.02	0.03	0.02	0.02	0.02	0.13
2-sided p-val: predicted to work vs. not	0.04	0.06	0.04	0.03	0.05	0.26

Panel D: Worker Turnover, All Stores	(1)	(2)	(3)	(4)	(5)
Sample of workers:	All	Untrained Workers	Trained workers	Trained Non-Mgrs	Trained Managers
Treatment	0.07	0.72*	-0.55*	-0.34	-1.13
	(0.25)	(0.39)	(0.29)	(0.30)	(0.70)
Observations	12,435	6,125	6,310	5,025	1,285
Workers	1547	819	728	586	142
Panel E: Worker Turnover, Stores Where RCT Predicted to Work					
Treatment	-0.64**	0.13	-1.32***	-0.86*	-2.78**
	(0.32)	(0.52)	(0.42)	(0.44)	(1.07)
Observations	6,251	2,976	3,275	2,564	711
Workers	794	405	389	308	81
Panel F: Worker Turnover, Stores Where RCT Not Predicted to Work					
Treatment	0.66*	1.27**	0.10	0.07	0.83
	(0.39)	(0.58)	(0.43)	(0.45)	(0.83)
Observations	6,184	3,149	3,035	2,461	574
Workers	819	452	367	302	65
1-sided p-val: predicted to work vs. not	0.01	0.07	0.01	0.07	0.00
2-sided p-val: predicted to work vs. not	0.01	0.14	0.02	0.14	0.01

Notes: Standard errors clustered by store are in parentheses. Panels A–C correspond to Panel A of Table 2 and Table 4; Panels D–F to Panel B of Table 2 and Table 5. The difference is that we exclude control stores where perceived annoyance about the RCT exceeded 6 on a 2–20 scale. In the *During-RCT store manager survey*, store managers rated annoyance separately for workers and for themselves, from 1 (not annoyed) to 10 (very annoyed); we sum the two to form the 2–20 scale. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A19: No Evidence that the Treatment Caused RMs to Allocate Effort to Treatment Stores or Stores where Predict Treatment to Work

Dep. var.:	Time spent on visits (minutes) (1)	Number of visits per week (2)
Panel A: All Stores		
Treatment	3.184	0.081
	(8.861)	(0.184)
Mean dep. var. if Treat=0	38.77	1.268
Stores	128	129
Panel B: Stores Where RCT Predicted to Work		
Treatment	2.520	0.147
	(8.916)	(0.186)
Mean dep. var. if Treat=0	28.27	0.974
Stores	69	70
Panel C: Stores Where RCT Not Predicted to Work		
Treatment	0.831	-0.065
	(16.070)	(0.324)
Mean dep. var. if Treat=0	52.66	1.657
Stores	59	59

Notes: This table shows that there is no evidence that the treatment caused RMs to allocate more effort to treatment stores or stores where they predict treatment to work. An observation is a store manager. Store managers were asked about the the frequency of visits and length of visits of RMs to their store in the *during-RCT store manager survey.** significant at 10%; ** significant at 5%; *** significant at 1%

Table A20: Accounting for Multiple Hypothesis Testing for Multiple Outcomes

Outcome:	Log Sales	Trained worker attrition
Panel A: All Stores		
Treatment	0.027*	-0.44*
	(0.015)	(0.25)
Conventional clustered p-vals	{0.07}	{0.08}
Westfall-Young p-vals	{0.15}	{0.15}
Bonferroni p-vals	{0.15}	{0.15}
Panel B: Stores Where RCT Predicted to Work by RMs		
Treatment	0.052**	-1.05***
	(0.020)	(0.36)
Conventional clustered p-vals	{0.010}	{0.005}
Westfall-Young p-vals	{0.026}	{0.026}
Bonferroni p-vals	{0.010}	{0.010}
Panel C: Stores Where RCT Not Predicted to Work by RMs		
Treatment	-0.003	0.091
	(0.020)	(0.37)
Conventional clustered p-vals	{0.87}	{0.81}
Westfall-Young p-vals	{0.96}	{0.96}
Bonferroni p-vals	{1.00}	{1.00}

Notes: The "Westfall-Young p-vals" are family-wise error rate adjusted p-values based on the Westfall & Young (1993) free step-down procedure (5,000 replications). In each panel, the family of hypotheses includes one for log sales and one for trained worker attrition. The Westfall-Young p-vals account for clustering by store by using a clustered bootstrap and are implemented using "wyoung.ado" in Stata (Jones *et al.*, 2019). Stars are based on the conventional clustered-by-store standard errors in parentheses, with * significant at 10%; ** significant at 5%; *** significant at 1%

Table A21: No Evidence to Support the Time Use Channel. DV = Log Sales

Time period:	All hours together	Between 12-1pm	Between 7-8pm	All hours separately
	(1)	(2)	(3)	(4)
Treatment	0.031	0.029*	0.015	0.021
	(0.031)	(0.015)	(0.038)	(0.014)
Time spent by store on daily protocol, in hours	0.030			
	(0.030)			
Treatment X Time spent on daily protocol	-0.013			
	(0.048)			
Hour where store generally does daily protocol				0.007
				(0.012)
Treatment X hour where store generally does daily protocol				-0.002
				(0.014)
Observations	1,355	1,431	1,304	17,424
Stores	137	145	136	137

Notes: An observation is a store-month during the RCT in columns 1-3, and is a store-month-hour of the day in column 4. 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, and column 4 additionally controls for hour of the day fixed effects. The data on time spent on daily protocol by store is from the *pre-RCT store manager survey*. Column 1 shows there is no evidence that the treatment effect of checklist removal varies by the amount of self-reported time that stores spent on the daily protocol in the pre-RCT period. Column 2-3 examines treatment effects of checklist removal on sales during 12-1pm and 7-8pm. These are the periods of the day when stores are most likely to complete the daily protocol. The treatment effects here are statistically identical to our overall treatment effect in column 1 of Panel A of Table 2. Column 4 shows there is no evidence that the treatment effect in a given hour of the day varies by whether it is an hour of the day where the store generally does the daily protocol. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A22: Mediation Analysis

Panel A: Employee Trust as Mediator for Effects on Trained Worker Attrition & Sales						
Outcome:	Trust	Trained worker attrition	Trained worker attrition	Trust	Log Sales	Log Sales
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.429***	-0.567**	-0.489*	0.222	0.012	0.010
	(0.160)	(0.279)	(0.290)	(0.136)	(0.017)	(0.017)
Trust			-0.183			0.012
			(0.238)			(0.012)
Share of treatment effect mediated by trust			14% (19%)			21% (34%)
Observations	4,600	4,600	4,600	997	997	997
Stores	100	100	100	100	100	100
What is an obs?	Worker-mth	Worker-mth	Worker-mth	Store-mth	Store-mth	Store-mth
Panel B: Trained Employee Attrition as Mediator for Treatment Effect on Sales						
Outcome:	Trained worker attrition (x100)	Log Sales	Log Sales	Trained mgr. attrit. (x100)	Log Sales	Log Sales
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	-0.552	0.027*	0.027*	-0.746	0.027*	0.027*
	(0.382)	(0.015)	(0.015)	(0.474)	(0.015)	(0.015)
Trained worker attrition			0.062			
			(0.046)			
Trained mgr. attrition						-0.021
						(0.016)
Share treatment effect mediated by attrition			-1.3% (1.3%)			0.6% (0.6%)
Observations	1,431	1,431	1,431	1,431	1,431	1,431
Stores	145	145	145	145	145	145
What is an obs?	Store-mth	Store-mth	Store-mth	Store-mth	Store-mth	Store-mth

Notes: Standard errors clustered by store in parentheses. The Delta method is used to calculate a standard error for the share of the treatment effected mediated by a variable, implemented in Stata using seemingly unrelated regression. "Worker-mth" means a worker-month. In Panel A, observations are weighted by the number of survey responses per store, as stores vary substantially in the number of workers who do the worker survey. Employee trust is measured in the *During-RCT worker survey* and is discussed in the main text in Section 3. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A23: Differences Between Treatment and Control Stores During the Post-RCT Firmwide Rollout

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

Panel B: Worker Turnover	(1)	(2)	(3)	(4)	(5)
Sample of workers:	All	Untrained Workers	Trained workers	Trained Non-Mgrs	Trained Managers
Treatment	0.07	-0.33	0.47	0.32	0.98*
	(0.32)	(0.60)	(0.31)	(0.36)	(0.56)
Observations	5,095	2,530	2,565	2,066	499
Mean DV if Treat=0	1.647	2.765	0.430	0.525	0
Workers	1365	692	673	544	129

Notes: Standard errors clustered by store are in parentheses. This table is similar to Table 2, but instead of analyzing data from the RCT, it uses data from the post-RCT rollout. * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix B Additional Discussion and Information

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

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

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

Did store managers continue to use any sort of informal version of the checklists during the RCT despite being in the treatment group? We investigate this using a few questions

in our *During-RCT survey of store managers*, which was conducted toward the end of the RCT. For the operational checklist, there was no effort by store managers to replace it. In the *During-RCT survey of store managers*, not a single store manager reported continuing to use an informal version of the checklists. For the daily protocol, some store managers reported making abbreviated informal notes about information that was in the protocol to communicate across shifts. For example, some store managers sent notes saying they forgot to clean the oven or that someone was sick.

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

As discussed in the main text, RAs made weekly visits to ensure that the formal version of checklists was not used in stores. After 6 weeks, the firm asked if the RA could come only every couple of weeks, which we agreed to.

B.3 Store League Performance Ranking (Section 2)

The store league performance ranking is broadly inspired by the Bundesliga (Germany's pro soccer league) and is decided upon by the firm's top management. Top management reviews many performance indicators across stores during the last few weeks (for example, sales, shopping, waste, targets) and creates a subjective assessment for each store. There is no formula that is used to calculate the ranking. The rankings are not used to calculate any bonuses, but instead mark stores needing improvement. Store league performance ranking is only observed for right before the RCT, and is not observed during the RCT.

B.4 Pre-RCT Survey of Regional Managers (Section 2)

This subsection provides further discussion and details on eliciting regional manager (RM) predictions of whether the treatment will work in each store. This one question survey is the *Pre-RCT survey of regional managers*.

Further details on classification of predictions. Some RM predictions involve comments that workers may like the checklist removal, but that operations will suffer. These are classified as No predictions, as they are not clear and unambiguous predictions that the treatment will work.

Reliability between classifiers of the RM predictions, as discussed in footnote 21 in main text. The main classification of RM predictions was done by the native German-speaking coauthor who interviewed the RMs. In addition, we had a second German-speaking coauthor independently translate and classify the RM predictions. Comparing the two independent classifications of the two native German speakers, the only store with possible ambiguity in the classification is Store 113. For Store 113, the original German is "Starke Filialleitung, offen für alles und kann alles gut umsetzen, schon ein paar Jahre dabei." Our results are very similar if Store 113 is removed, as seen in Table A17.

Why weren't the interviews done using online surveys? Many economists collect survey data using online surveys. We opted to use phone instead of online surveys, as RMs are used to communicating by phone. Based on our interactions with the firm, we believe that our response rate would have been much lower using online surveys. We also thought that RMs would be most comfortable and candid giving responses by phone instead of in written form.

Why weren't the phone conversations recorded? As mentioned in the main text, Germany has very strong norms regarding recording of phone calls. This reflects the history of secret phone recordings and wiretapping under Communism and Nazism. It would have been extremely awkward to ask to record phone calls and this could have negatively affected RM participation. Instead, the coauthor conducting the calls made detailed notes by hand.

Why weren't incentives used for predictions? As discussed in the main text, no incentives were used for RM predictions because they are subjective, i.e., we did not precisely define what it means for the treatment to "work" such that one could check *ex post* to see if a prediction was correct. Even if it were possible to incentivize predictions, there are four advantages of not using incentives. First, not using incentives avoids "incentive effects" for RMs 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 RMs. Third, not using incentives seemed natural for higher-ranking RMs. Fourth, reviewing the literature, Haaland *et al.* (2023) argue that incentives are not needed to accurately elicit beliefs and discuss how incentives can sometimes worsen elicitation.

B.5 Randomization Procedure and Controlling for Stratification Variables in the Empirical Analysis (Sections 2-3)

As described in Section 2, we perform a stratified randomization using region, pre-RCT sales, pre-RCT head count, and pre-RCT store league performance ranking.[2] 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.

As described in the main text, stratification is done with three binary variables and a region variable having 9 values. There are 46 strata instead of 72 strata (i.e., 2x2x2x9) because not all combinations are present in the data (e.g., in some regions, there may be no store below or above mean in all three dimensions).

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. Table A2 shows results with strata dummies.

[2]Two of the 145 stores are missing pre-RCT store league performance ranking. They are placed in the strata with above-mean store league performance ranking.

B.6 Data Construction (Section 2)

B.6.1 Store-month panel dataset

We have data from accounting records on hourly sales going back to 2014. We observe total sales in each store in each hour, as well as hourly sales of different types of products, namely, snacks (such as sandwiches), drinks, and "bread" (including doughnuts, baked bread, cakes, and pretzels). In constructing the store-month panel, we exclude zero sales months, but we do not exclude months when stores have relatively low or high sales. To address outliers, we combined two strategies. First, we manually identified several store-months where we received indication that construction or renovations were occurring. Second, we excluded sales-months where stores experienced a change in sales that was below the 1st percentile or above the 99th percentile of change in log sales. Using these two strategies together, all our results are similar.

B.6.2 Employee-month panel dataset

Minijobbers. Our worker-month panel is based on regular workers at the firm. We exclude Germany's "minijobbers"—short-term employees capped at 12 hours per week and exempt from payroll taxes (Tazhitdinova, 2022)—from our worker-month panel. Minijobbers average only 7–8 hours (vs. $\approx$ 30 hours for regular staff), are hired temporarily, and naturally attrite. They represent just 8% of total hours in the RCT, and including them does not alter our findings: the treatment still leaves overall attrition unchanged, lowers skilled-worker attrition, and has no significant effect on minijobber attrition.

Total pay. Total monthly pay is given by 4.33*Weekly Pay + Monthly Bonus Pay.

Identifying employee store and employee movements across treatment arms. Employee store is provided using administrative data from the firm on employee affiliations. This dataset follows workers over time. In case where the same worker is affiliated with multiple stores in one month in the administrative data, we assign the worker to one store.

As seen in Table 5 in the main text, the sum of employees in treatment stores and employees in control stores is greater than the number of total employees, reflecting that some employees are affiliated with multiple stores over the RCT. To verify that such workers do not drive results, we repeated our attrition results excluding workers who are exposed to both treatment and control stores, and all our conclusions remain, with estimates slightly stronger, as seen below in Table B1.

B.7 Fidelity to Pre-registration (Section 2)

Outcomes. In our pre-registration, we stated that our primary outcome is sales, and our secondary outcomes are attrition, absenteeism, leadership styles, employee-manager interaction, and manager time use. We follow this closely. However, we do not analyze absenteeism or leadership styles. For absenteeism, we currently have been unable to obtain comprehensive employee absence data from the firm. For leadership style, we ultimately did not collect quantitative data. For employee-manager interaction, we focus on interactions between regional managers (RMs) and stores. Measuring interactions between store managers

Table B1: Robustness: Exclude Workers Who Work at Both Treatment and Control Stores

Panel A: Worker Turnover, All Stores	(1)	(2)	(3)	(4)	(5)
Sample of workers:	All	Untrained Workers	Trained workers	Trained Non-Mgrs	Trained Managers
Treatment	0.06	0.69*	-0.50*	-0.25	-1.36**
	(0.25)	(0.41)	(0.26)	(0.27)	(0.66)
Panel B: Worker Turnover, Stores Where RCT Predicted to Work					
Treatment	-0.55*	0.19	-1.18***	-0.73*	-2.66***
	(0.31)	(0.54)	(0.37)	(0.41)	(0.92)
Panel C: Worker Turnover, Stores Where RCT Predicted Not to Work					
Treatment	0.59	1.19*	0.14	0.17	0.50
	(0.40)	(0.62)	(0.36)	(0.35)	(0.96)
2-sided p-val: predicted to work vs. not	0.03	0.23	0.01	0.09	0.02
1-sided p-val: predicted to work vs. not	0.01	0.11	0.01	0.05	0.01

Notes: Standard errors clustered by store are in parentheses. This table is similar to Panel B of Table 2 and to Table 5. The difference from these tables is that we exclude workers who we observe working at both treatment and control stores during the RCT. * significant at 10%; ** significant at 5%; *** significant at 1%

and frontline employees proved difficult, as such communication is frequent, informal, and multi-modal. We analyze RM time use rather than store manager time use.

Heterogeneity. Our analysis in Section 4 focuses on our three main pre-registered heterogeneity dimensions: RM predictions, average worker tenure, and store head count. As additional / secondary dimensions, our pre-registration also listed worker reciprocity and sales director identity. For clarity of exposition and because worker reciprocity is not measured prior to the RCT, we omit these from Table A9. If included, interactions of sales director ID and the treatment variable are statistically insignificant.

B.8 Framing to Workers Explaining Checklist Removal (Sec. 2)

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

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

We acknowledge that part of the effects we estimate could be due to framing. However, we believe it is highly unlikely that a pure framing effect could lead to our RCT's quite sizable effects on sales and attrition that persist for 10 months. Prior work on framing in the field tends to estimate moderate effects that are fairly context-specific.[3] We view the

[3] Hossain & List (2012) find that whether an incentive is gain- or loss-framed matters for team, but not

framing of the RCT as complementary to the potential signaling of removing monitoring, i.e., the framing helps people understand the signaling. We also note that from a managerial standpoint, it is less policy-relevant to use neutral standpoint. To make policy changes comprehensive to workers, companies want to use positive framings, so using a positive framing is natural.

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

B.9 During-RCT Survey of Workers (Section 3)

Response rate and patterns. As noted in the main text, the survey response rate was around 35%—above the rates in many top-published papers and typical for employee surveys. For instance, Card *et al.* (2012) report a 20% rate among University of California employees; Hoffman & Burks (2020) report 28% in a trucker productivity survey and 25% in an exit survey; Biasi & Sarsons (2021) report 13% among teachers; and Cullen *et al.* (2023) report 13% in a survey of hiring managers. Due to privacy constraints, demographic questions were not included in our worker survey. However, Table A5 shows the treatment was uncorrelated with store-level response rates. Response rates were slightly higher in stores with more female or older workers; the gender pattern aligns with findings in Dutz *et al.* (2021).

Measuring commitment. Panel A of Table 3 reports results on commitment to the store. We also asked about commitment to the firm but find no treatment effect. In lower-skill retail jobs, it is common for workers to feel strongest attachment to their store and work team rather than to the broader firm. In interviews, store managers emphasized that "we" typically referred to the store, not the firm.

B.10 Use of Data on Google Reviews (Section 3)

Extracting reviews. We use data on Google reviews to better understand mechanisms for our effects. To extract Google reviews, we use a developer tool at `Apify.com`. Reviews are extracted using store addresses. While Google reviews observed in typical web searches have approximate dates (e.g., 2 years ago), `Apify.com` allows us to extract exact dates of reviews. The reviews were extracted in April 2025, and we restrict attention to reviews from January 2019 to January 2022, i.e., the 10 months of the RCT and the 27 months prior. After cleaning and data processing, we obtain 11,034 reviews posted from 2019m1-2022m1, with 4,084 reviews from the RCT period and 6,950 review from before the RCT. We drop ratings that are duplicate in the rater, rating, date, review text, and store. Google reviews contain a 1-5 star rating, as well as sometimes text (most critical for us given our focus on using the reviews to get at mechanisms).

Identifying qualitative characteristics. For each review, we had a German-speaking RA evaluate the reviews. The RA was not aware of whether the reviews came from control

individual performance. However, De Quidt *et al.* (2017) find no effect of framing on performance.

or treatment stores. Roughly half of the reviews (5,318 reviews) contain text in addition to a star rating. Having read a selection of reviews, we identified the following topics that were most frequently mentioned, positively or negatively: product characteristics (taste, look, smell), service quality (were sales personnel friendly and helpful?), store appearance (looks, ambience, hygiene level), speed of service, value for money, and product availability.

We instructed the RA to read through all reviews with text and indicate which of the above topics were mentioned in the text of each review. For example, the review "Gute Qualität bei den Waren. Freundliche Bedienung. Alles in allem zu Empfehlen!" ("Good quality goods. Friendly service. All in all recommended!") positively mentions product and service quality, and so the RA indicated this review as mentioning these two topics. Another review, "Personal ist leider teilweise unfreundlich, zumindest die älteren Mitarbeiter... Ware ist sehr oft leer, egal zu welcher Uhrzeit. Die Backwaren an sich sind sehr lecker" / "Unfortunately, some of the staff are unfriendly, at least the older employees... Shelves are very often empty, no matter what time of day. The baked goods themselves are very tasty", positively mentions product but negatively mentions service quality and product availability, which topics were accordingly indicated by the RA. 85% of reviews with text mentioned one or several of the above topics. The remaining 15% did not, the majority of which had a positive assessment, e.g., "Gut" / "Good", "OK", "Wie immer alles bestens"/ "As always, everything is fine", or simply positive emojis. We validated the RA's choices by reading a selection of reviews ourselves, finding a very high rate of correspondence.[4]

The qualitative characteristics are modestly correlated but appear distinct. In a principal component analysis, the first component—loading positively on all characteristics—explains only 28% of variance. This suggests the characteristics don't reflect one trait.

Sample restriction based on if reviews have text. Table A8 provides additional analyses related to our Google reviews data. Our main results on Google reviews shown in Panel B of Table 3 use data on all reviews, including reviews with no text. Panel A of Table A8 restricts exclusively to reviews with text and obtains results that are qualitatively extremely similar (and slightly stronger in terms of statistical significance). This indicates that our findings are not driven by assumptions regarding restricting to reviews with text.

Effects on quantitative score. Panel B of Table A8 performs analysis on how the treatment affects the quantitative score (1-5) in online reviews. As seen in column 1, the treatment effect is statistically insignificant. Online reviews are known to be left-skewed, with lots of high scores (Tadelis, 2016), so to dig further, we also look at rates of each level of score, i.e., the share of reviews that are 1s, 2s, 3s, 4s and 5s. We see no significant effects.

General issues with data on online reviews. As discussed in Tadelis (2016), there are various common concerns with using online review data. One issue is that some reviews may be fake. However, there is no reason to believe that our treatment would affect the

[4]These included reviews with text but not marked as positive or negative in any attribute; those mentioning positive shop appearance; and those mentioning positive speed of service. Without knowing whether the store was in the treatment or control group, we made a small number of corrections. We also cross-checked a subset of reviews using ChatGPT, which generally showed high agreement and helped identify several additional corrections. However, ChatGPT sometimes misclassified reviews based on superficial keyword matches—e.g., interpreting text like 'only suitable for a quick visit' as praise for fast service. Because of such errors, we use the human RA as our primary classifier.

likelihood of a store receiving fake reviews. Moreover, our partner firm has a traditional management culture and is unlikely to generate fake reviews. A second issue is selection: because many customers do not leave reviews, the treatment could affect whether customers choose to leave a review. However, as shown in column 7 of Panel B of Table 3, the treatment has no statistically significant effect on the number of Google reviews a store receives. This suggests that selection is likely limited in our context.[5]

Comments on negative aspects of stores. As noted in Section 3, negative comments are much less frequent than positive ones. Positive mentions dominate in all six categories—often by a factor of 3 to 10. As a result, we lack statistical power to analyze negative aspects of stores in Google reviews.

B.11 Selection Framework from Dal Bó *et al.* (2021) (Section 4)

Are RMs making good predictions by combining standard observable characteristics of stores, or are they making good predictions based on characteristics that are unobserved to the econometrician? Dal Bó *et al.* (2021) propose an estimation framework for separating treatment effect heterogeneity based on observables from that based on unobservables using prediction data. It is broadly similar in spirit to the simple interaction term model considered in Table A9. However, by making an assumption about the joint normality of several terms,[6] it proposes replacing terms involving RM Predictions with a generalized residual of the RM predictions using an inverse Mills ratio (e.g., the key interaction term of Treatment X Prediction is replaced by Treatment X Inverse Mills Ratio). Here is the two-step procedure:

1. Estimate a probit model where RM beliefs about a store, B_s, are regressed on various pre-RCT store characteristics, Z_s. Obtain the inverse Mills ratio $\Lambda(Z_s, B_s)$.

2. Estimate OLS models of treatment effect heterogeneity where the treatment dummy is interacted with $\Lambda(Z_s, B_s)$ (representing the *unobservable-to-the-econometrician* determinants of RM beliefs) and with Z_s (the *observable* determinants of RM beliefs).

As in Table 5 of Dal Bó *et al.* (2021), we perform the estimation using $Z_s = \varnothing$ and using a richer set of Z_s, thereby allowing us to learn about the relevance of observable characteristics in explaining treatment effect heterogeneity. In the second stage, we include our baseline controls for comparability with our main specifications, but conclusions are unchanged without them.

[5] While the estimate is statistically indistinguishable from zero, the coefficient is +0.4, meaning treatment stores receive 0.4 more reviews per month. Research on online reviews suggests that not leaving a review is more common after a negative experience (Tadelis, 2016). Thus, the true effect of the treatment on positive customer experiences could be understated if customers in treatment stores were equally or more likely to remain silent compared to those in control stores.

[6] In our setting, the joint normality assumption concerns (i) the unobserved component of the return to the treatment and (ii) the noise in the signal that RMs receive about this component. In the framework of Dal Bó *et al.* (2021), this assumption also applies to unobserved RM preferences, which are not explicitly modeled in our setting but could be incorporated without difficulty. An example of a violation would be bimodal noise in RM signals, such as strong signals in either direction about whether the treatment would work. While joint normality seems like a reasonable approximation, Table A9 shows that heterogeneity by RM predictions is extremely robust—even conditional on interactions between treatment and observable characteristics. The results in Table A9 do not rely on a joint normality assumption.

We refer to the estimation procedure as a selection model framework since it involves an inverse Mills ratio, as in Heckman-style selection models. Unlike a Heckit model, however, there is no sample selection: we observe full data for both stores where RMs predict the treatment will work (akin to "selected" stores) and stores where they do not ("non-selected" stores). The role of the inverse Mills ratio is not to solve an endogeneity issue, but to examine whether RM predictions exploit information on observed or unobserved characteristics.

Table B2 provides estimates, with results on sales in columns 1-2 and trained worker attrition in columns 3-4. In the odd columns, the selection model is estimated with $Z_s = \varnothing$, whereas in the even columns, the selection model is estimated using as Z_s the characteristics also analyzed in Table A9. For sales, the coefficient on Treatment X Inverse Mills Ratio is essentially unchanged going from column 1 to column 2, indicating that unobserved drivers of RM predictions are the key drivers of treatment heterogeneity, and that the observable characteristics listed play little role. For trained worker attrition, the interaction term declines by only 24% from column 3 to column 4, suggesting that information on observables explains only a limited portion of the heterogeneity associated with RM predictions.

Table B2: Effect Heterogeneity in Terms of Observables and Unobservables

Dep. var.:	Log Sales		Trained Worker Attrition (x100)	
	(1)	(2)	(3)	(4)
Treatment X Inverse Mills Ratio	0.034**	0.035**	-0.705**	-0.539
	(0.017)	(0.017)	(0.325)	(0.332)
Treat*(Pre-RCT Log Sales)		-0.117		
		(0.163)		
Treat*(Pre-RCT turnover rate)				0.156
				(0.434)
Treat*(Pre-RCT mean head count)		-0.000		0.036
		(0.006)		(0.057)
Treat*(Pre-RCT mean tenure of workers)		-0.007		-0.106
		(0.006)		(0.141)
Treat*(Pre-RCT mystery shopping score)		-0.053		-1.065
		(0.035)		(0.706)
Treat*Pre-RCT Log (Shrinkage as % of Sales)		0.013		0.211
		(0.156)		(2.549)
1-sided p-val, Treatment X Inverse Mills Ratio	0.02	0.02	0.02	0.05
2-sided p-val, Treatment X Inverse Mills Ratio	0.05	0.04	0.03	0.11

Notes: This table provides results from the "second stage" of the selection model framework of Dal Bó *et al.* (2021). For brevity, we omit showing first stage probit results because they are qualitatively similar to those in column 1 of Table 7. Analyses here are similar to those in Table A9, but we replace RM predictions with the Inverse Mills Ratio, which is a general residualized form of RM predictions. As in Table A9, each heterogeneity dimension is fully interacted with controls. Thus, we also include a treatment dummy and Inverse Mills Ratio as separate regressors. The Z_s variables used in the first stage are the ones listed in each column. The first-stage is intentionally parsimonious and includes only the listed Z_s variables; the second stage includes our baseline controls (pre-RCT store characteristics used in the stratified randomization and month-year dummies) for comparability with our main results. Month-year dummies can't be included in the first stage since RM predictions are done once per store, but our conclusions are unchanged (i) if stratification variables are included in the first stage probit or (ii) if baseline controls are excluded from first and second stages. Stars are based on two-sided p-values, with * significant at 10%; ** significant at 5%; *** significant at 1%

In all four columns, the one-sided p-value for Treatment X Inverse Mills Ratio being different from 0 is 0.05 or less. One-sided p-values are appropriate given the one-sided nature of the hypothesis (as Treatment X Inverse Mills Ratio is similar to Treatment X Prediction).

Overall, this exercise supports RM beliefs predicting treatment effect heterogeneity due to characteristics unobserved to the econometrician. This is fully congruent with other results in the paper, both the evidence in Table 7 that it is difficult to predict RM predictions, and the evidence in Section 4 and Table A9 of robust treatment heterogeneity by RM predictions.

B.12 Why Can't You Do Heterogenous Treatment Effects According to the Absence of Problems or Whether Workers Would Like Having Checklists? (Section 4)

Table 6 shows that many positive RM predictions mention (1) the absence of problems or (2) workers liking not having checklists. However, these statements should not be interpreted as systematic, comparable measurements of underlying store characteristics. Rather, they are qualitative rationales offered in the course of making a prediction, whose content is naturally shaped by the act of explaining why the treatment is expected to work. RMs were asked to explain whether they believed the treatment would work, not to report specifically on whether a store in fact lacked problems or whether workers would like not having checklists.[7]

B.13 Profit Calculation Details (Section 5.1)

Time cost of implementing RCT. The project team held two half-day meetings in the nine-person full group. Assigning each person a day rate of €1500 implies a total cost of €13.5k. There were also four half-day meetings of the small group; using 7 half-person days at a rate of €1200 (reflecting the lower seniority of participants) yields a cost of €4.2k. Data transmission required about 5 person-days at €800 per day, for a cost of €4k. Training the 15 regional managers and two top directors took about one workday, costing roughly €1k. In addition, RA time amounted to about €9k. Summing up, total time cost is roughly €31k.

Assumptions on cost of turnover. Blatter *et al.* (2012) use comprehensive survey data on Swiss firms to estimate hiring costs. As in Germany, there is a clear divide between skilled and unskilled workers in Switzerland. In Blatter *et al.* (2012), trained sales clerks have a mean hiring cost of 10.309 weeks of wages (Table 4). This corresponds to a cost of €4,320. Of our trained workers, about half the turnover events are from trained non-managers and half are from managers. Blatter *et al.* (2012) do not estimate turnover costs for managers, but one would imagine that costs for store managers would likely be substantially higher than for non-managers. Thus, we use a turnover cost of €6000 for trained workers.

According to Blatter *et al.* (2012), the interview time for a trained worker (excluding managers) is 5-7 times higher than that of untrained workers (Table 2 of their paper). If we assume that total hiring costs mirror the differences in interview time, hiring costs for

[7] An alternative design would have been to elicit detailed information about many aspects of each store from all RMs. This was not feasible given limited interview time with RMs, and because prompting RMs about specific characteristics could lead them and distort the belief-elicitation process.

untrained workers would be about 6 times lower than that for trained non-managers. This justifies a turnover cost of €4,320 ÷ 6 = €720 for untrained workers.

Ultimately, the exact assumptions on cost of turnover do not drive the qualitative findings related to profits, and the conclusions are highly robust. If we instead use a turnover cost for trained workers of €4,320 instead of €6,000 (i.e., we assume that store managers have the same turnover cost as trained non-managers), the turnover benefit is roughly €99k instead of €150k, and the estimated benefit to cost ratio is still 57:1. If there was no turnover benefit at all of the RCT (which would occur if trained and untrained workers had roughly the same cost of turnover), the estimated benefit to cost ratio is still 55:1.

Aggregation of benefits over all 145 stores. In calculating the benefits and the benefit to cost ratio of checklist removal, we aggregate over all 145 stores instead of merely the 74 stores that were treated. We do this because most of the costs of implementing the RCT are firm-wide and do not depend on the number of stores treated. That is, checklist removal would have broadly similar costs if implemented firmwide instead of only in treated stores. If instead one aggregated the benefits exclusively to the treated stores, then the benefit to cost ratio is still extremely large at roughly 30:1.

Estimating benefit to cost ratios conditional on RM predictions. To estimate benefit to cost ratios conditional on whether RMs predict the treatment to work or not work, we use the estimates separated by RM predictions. For sales effects, we use the treatment effects of 5.2% and -0.3% in Table 4. Attrition effects rely on Table 5.

Profit margin calculation. The firm's pre-RCT profit margin is $M_0 = \frac{\pi_0}{R_0}$, where π_0 is the pre-RCT profits and R_0 is pre-RCT revenues. For the post-RCT period, the 0 subscripts are replaced by 1. Post-RCT profits can be written as $\pi_1 = \pi_0 + \text{RCT benefits} - \text{RCT costs}$. Given that RCT costs are small relative to RCT gains, $\pi_1 \approx \pi_0 + \text{value added} \times \text{sales effect}$. The post-RCT profit margin is $M_1 \approx M_0 + .027 * \text{value added} = .01 + .027 * .56 \approx .025$, which is a more than doubling of the profit margin.

B.14 Time Use Channel (Section 5.4)

Section 5.4 investigates the time use channel using pre-RCT heterogeneity in time spent on the daily protocol. As a supplemental test, we also use a question from the *During-RCT store manager survey* that asked whether employees gained additional time due to checklist removal (yes or no). If we repeat the main analysis in Panel A of Table 2 but splitting the sample in two, we see no evidence that treatment effects on store outcomes vary by this reported time savings question. Because this variable is endogenous—reflecting store managers' perceptions during the RCT—this test should be interpreted with caution. Nonetheless, it supports the broader conclusion that the time use channel does not appear to be the main driver of the observed effects.

B.15 Mediation Analysis (Section 5.4)

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.

Beyond trained worker attrition and trained store manager attrition, we also examined whether the effect of sales was mediated by the increase in untrained worker attrition. Repeating the mediation analysis in Panel B of Table A22 but for untrained worker attrition, we see no evidence that it mediates the increase in sales.

Appendix C Materials in RCT and Firmwide Rollout

This section summarizes the materials and survey questions that were used in the RCT and firmwide rollout, and that are analyzed in the paper or Web Appendix. All have been translated from German.

C.1 *Pre-RCT Survey of Regional Managers*

C.1.1 Wording Used for the Regional Manager Predictions

I presented the pilot project in a regional manager meeting in February 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 shop XYZ indeed was a pilot shop: How well would the daily business work ("function") in the shop with fewer checklists?

C.1.2 Discussion of Wording

RMs are asked how well the daily business would function with fewer checklists. While this could be read as a statement about baseline performance rather than a treatment effect, our preliminary conversations with RMs suggested that the language of treatment effects and counterfactuals was not natural for many RMs. We therefore chose wording intended to intuitively elicit beliefs about where checklist removal would have the largest impact.

C.2 *Pre-RCT Survey of Store Managers*

- At what time do you or your employees usually fill out the daily protocol? `[INTERVIEWERS ASKED FOR A CONCRETE HOUR DURING THE DAY IF SOMEONE GAVE A RESPONSE LIKE`

AFTER LUNCH]

- How often do you or your employees usually fill out the daily protocol per day?
- How much time do you or your employees typically spend filling out the daily protocol each time?

C.3 *During-RCT Survey of Store Managers*

- Thinking about a typical work week over the past five months: On average, how often did your regional manager visit your store per week?
- When the regional manager visited: On average, how long did he or she stay in your store each visit?
- [ONLY ASKED FOR TREATMENT STORES] Did employees gain additional time due to the removal of the checklists?
- [ONLY ASKED FOR TREATMENT STORES] On a scale from 1 (very bad) to 7 (very good): How did you find [FIRM NAME]'s initiative to eliminate the operational checklist?
- [ONLY ASKED FOR TREATMENT STORES] On a scale from 1 (very bad) to 7 (very good): How did you find [FIRM NAME]'s initiative to eliminate the daily protocol?
- [ONLY ASKED FOR TREATMENT STORES] A few months ago, the daily protocol and the operational checklist were eliminated in your store. Let's start with the operational checklist. Did you do anything differently—such as introducing a new checklist or monitoring your employees more closely—in order to achieve the goals that were previously supported by the operational checklist? What about the elimination of the daily protocol: Did you or your employees do anything else (e.g., write notes or communicate via WhatsApp) to achieve the goals that were previously supported by the daily protocol?
- [ONLY ASKED FOR CONTROL STORES] A few months ago, as part of a pilot project, the daily protocol and the operational checklist were removed in some randomly selected stores. Did you hear anything about this pilot project?
- [ONLY ASKED FOR CONTROL STORES AND IF YES TO PREVIOUS QUESTION] Were you annoyed or disappointed that the daily protocol and the operational checklist were not removed in your store? Please answer on a scale from 1 (not annoyed) to 10 (very annoyed).
- [ONLY ASKED FOR CONTROL STORES] Did your employees notice that the daily protocol and the operational checklist were removed in other stores?
- [ONLY ASKED FOR CONTROL STORES AND IF YES TO PREVIOUS QUESTION] Were your employees annoyed or disappointed that the daily protocol and the operational checklist were not removed in your store? Please answer on a scale from 1 (not annoyed) to 10 (very annoyed).

C.4 *During-RCT Survey of Workers*

- Interpersonal relations and the culture at [FIRM NAME] are characterized by mutual trust between the head office and the employees in the stores.
- My [FIRM NAME] store has a great deal of personal meaning for me.
- The company [FIRM NAME] has a great deal of personal meaning for me.
- Thinking about the most recent colleague who was hired at your [FIRM NAME] store – do you agree that he or she was well trained and onboarded? (If you were the most recent new employee, please skip this question.)
- Please ask whether you disagree or agree with each of these statements on a scale from 1 to 7:
  - Whether baking processes are carried out correctly is regularly checked at [FIRM NAME].
  - The quality of the bread rolls is regularly checked at [FIRM NAME].
  - How we as employees interact with customers is regularly checked at [FIRM NAME].
  - Whether products are presented "correctly" is regularly checked at [FIRM NAME].
  - Whether current special promotions and guidelines are implemented correctly is regularly checked at [FIRM NAME].
  - [ONLY ASKED FOR TREATMENT STORES] The removal of the daily protocol a few months ago was a good decision.
  - [ONLY ASKED FOR TREATMENT STORES] The removal of the operational checklist a few months ago was a good decision.

C.5 Information on the RCT Provided to Store Managers and Employees

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

C.6 Examples of Older Versions of the Operational Checklist

Below are two examples of the operational checklist in the past. The first is from 2019/08. Instead of signing, workers indicate whether they did well, poorly, or average on different tasks. The second is from 2017/01. Workers would sign this at multiple points during the day.

INCREASING AVERAGE CUSTOMER SALES

Challenge August 2019

We are NAME OF THE COMPANY

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

Toolbox:

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

Customer list: Goal → Increase customer satisfaction!!!

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

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

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

C.7 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

Belloni, Alexandre, Chernozhukov, Victor, & Hansen, Christian. 2014. Inference on Treatment Effects After Selection Among High-dimensional Controls. *Rev. Econ Studies*, **81**(2).

Biasi, Barbara, & Sarsons, Heather. 2021. Flexible Wages, Bargaining, and the Gender Gap. *Quarterly Journal of Economics*, **137**(1), 215–266.

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

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

Card, David, Mas, Alexandre, Moretti, Enrico, & Saez, Emmanuel. 2012. Inequality at Work: The Effect of Peer Salaries on Job Satisfaction. *AER*, **102**(6), 2981–3003.

Cullen, Zoe, Dobbie, Will, & Hoffman, Mitchell. 2023. Increasing the Demand for Workers with a Criminal Record. *Quarterly Journal of Economics*, **138**(1), 103–150.

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 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.

Dutz, Deniz, Huitfeldt, Ingrid, Lacouture, Santiago, Mogstad, Magne, Torgovitsky, Alexander, & Van Dijk, Winnie. 2021. *Selection in surveys: Using randomized incentives to detect and account for nonresponse bias.* WP 29549. NBER.

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

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

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

Hoffman, Mitchell, & Burks, Stephen V. 2020. Worker Overconfidence: Field Evidence and Implications for Employee Turnover and Firm Profits. *Quantitative Economics*, **11**(1), 315–348.

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.

Jones, Damon, Molitor, David, & Reif, Julian. 2019. What do Workplace Wellness Programs do? Evidence from the Illinois Workplace Wellness Study. *QJE*, **134**(4), 1747–1791.

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.

Westfall, Peter, & Young, S. Stanley. 1993. *Resampling-based Multiple Testing.* Vol. 279.

Young, Alwyn. 2019. Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results. *QJE*, **134**(2), 557–598.