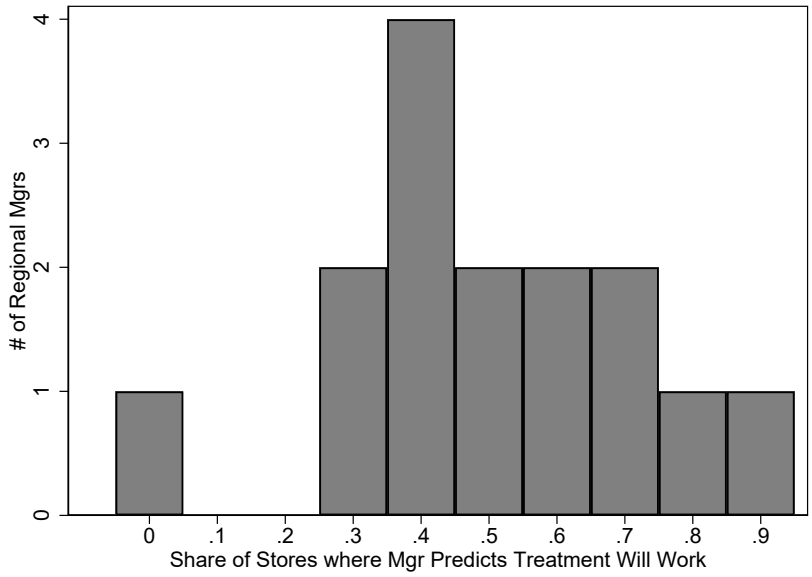
Web Appendix, "Is This Really Kneaded? Identifying and Eliminating Potentially Harmful Forms of Workplace Control", by Friebel, Heinz, Hoffman, Kretschmer, and Zubanov

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]

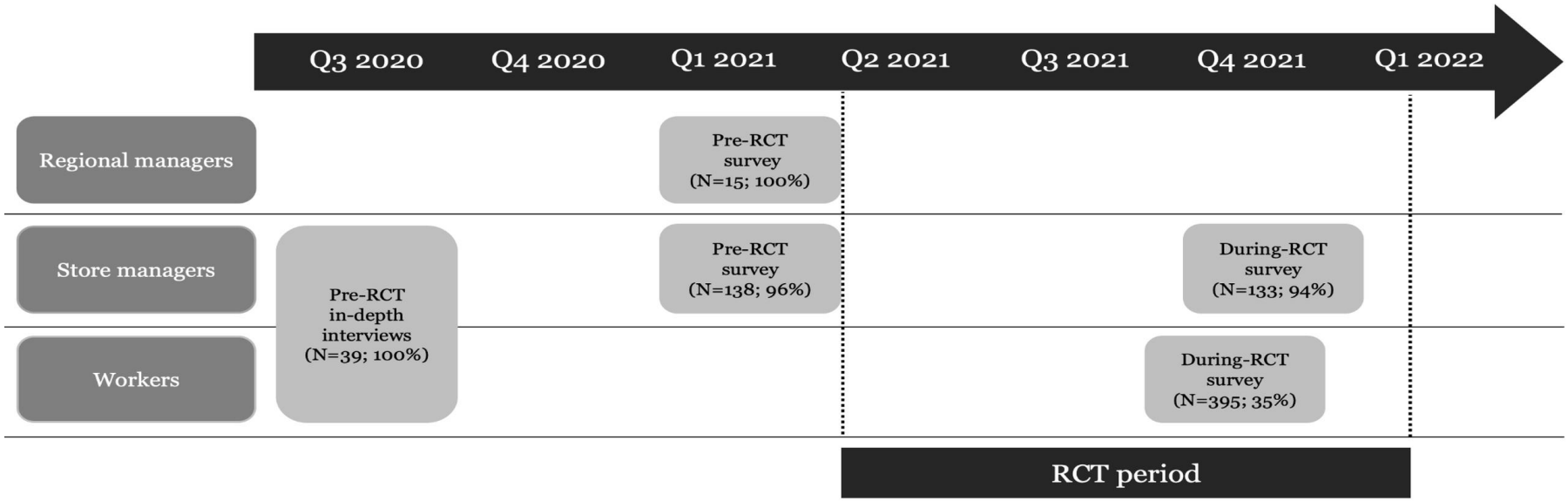
Figure A1: Variation in RM-Level Rates of Predicting that the Treatment Will Work



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.

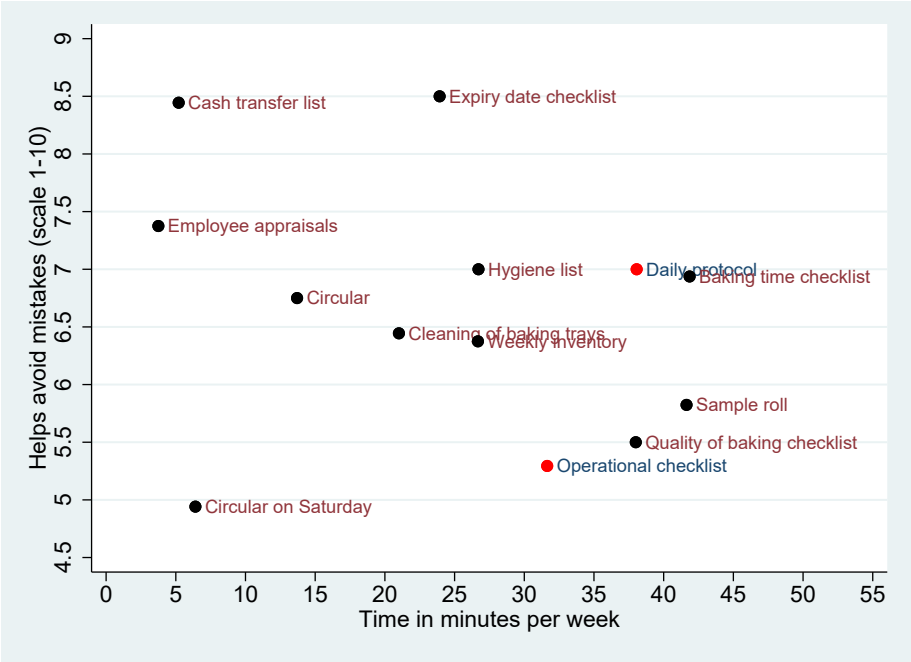
The *Pre-RCT survey of regional managers* collects regional manager predictions about whether the treatment will work for each store they supervise. This was a one question survey, and is discussed further in Appendix B.4.

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

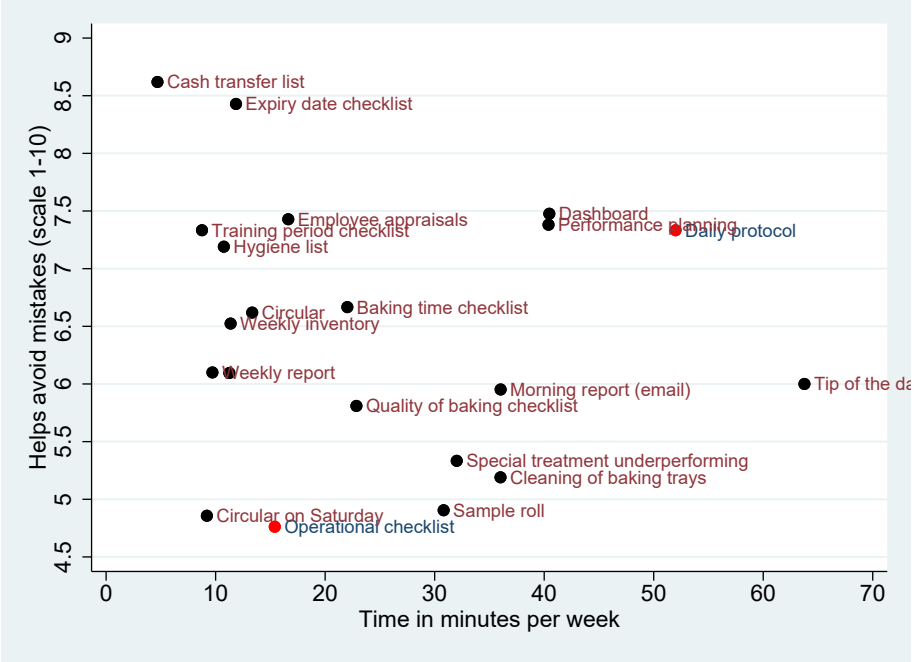
The *During-RCT survey of workers* measures workers' attitudes toward the firm. Results from this survey are presented in Panel A of Table 3. The response rate is calculated based on 395 responses from roughly 1100 workers contacted. Our paper only analyzes regular employees. This survey is discussed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



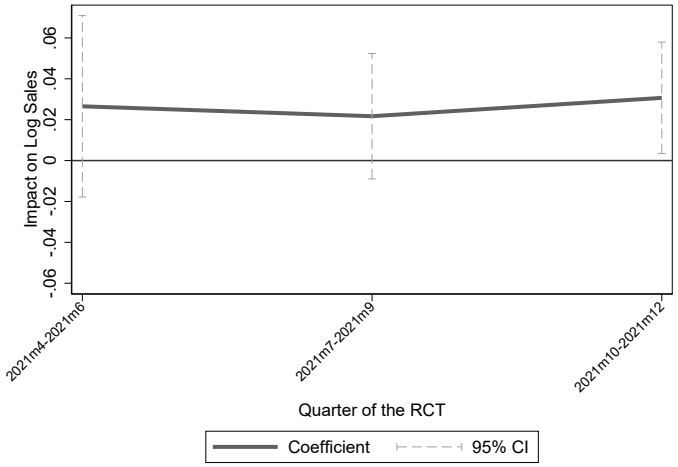
(a) Workers (N=18 workers)



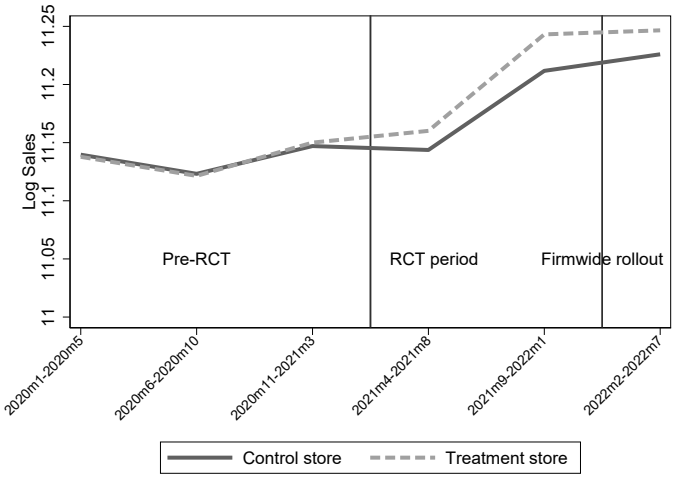
(b) Store Managers (N=21 store managers)

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.

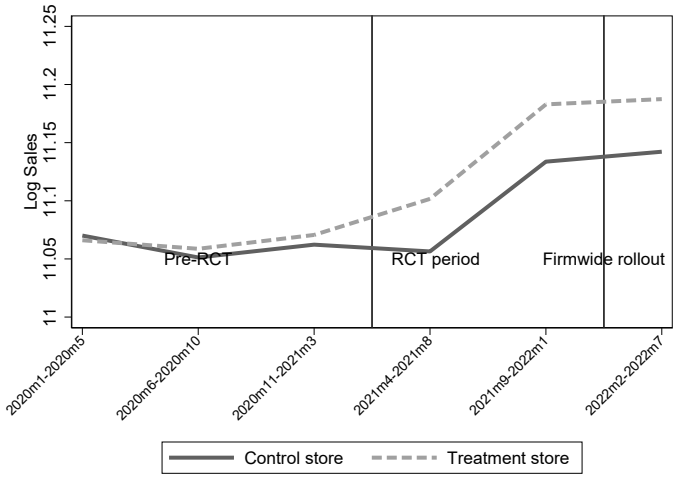
Figure A4: Further Figures Summarizing the Effect on Sales



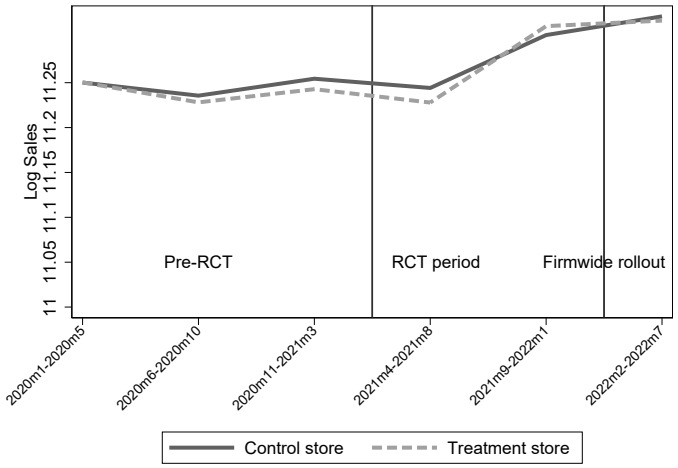
(a) Effects on Sales by Quarters of the RCT



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



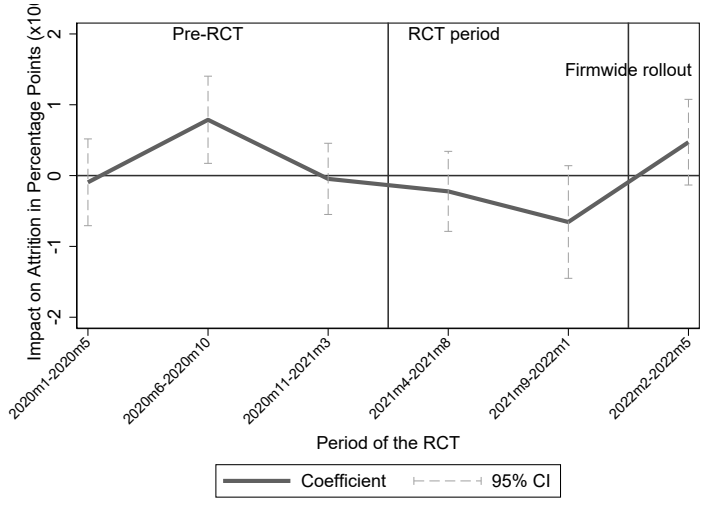
(c) Two Lines, Stores where RCT Predicted to Work



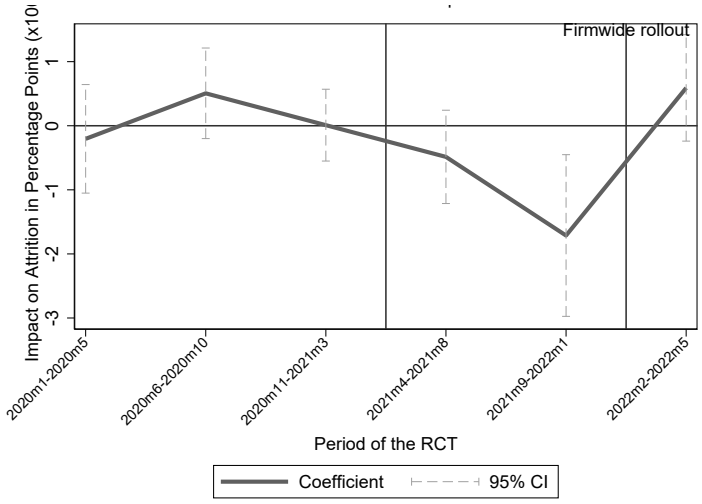
(d) Two Lines, Stores where RCT Predicted Not to Work

Notes: This figure provides robustness for our effects on sales. Panel (a) is similar to the "RCT Period" of Figure 6(a), but shows effects using quarter of the RCT instead of 5-month periods. Panels (b)-(d) compares treatment versus control stores with two separate lines. The control line plots the control store means, whereas the treatment store plots control means plus the treatment effect in each period. 95% confidence intervals based on conventional clustering by store.

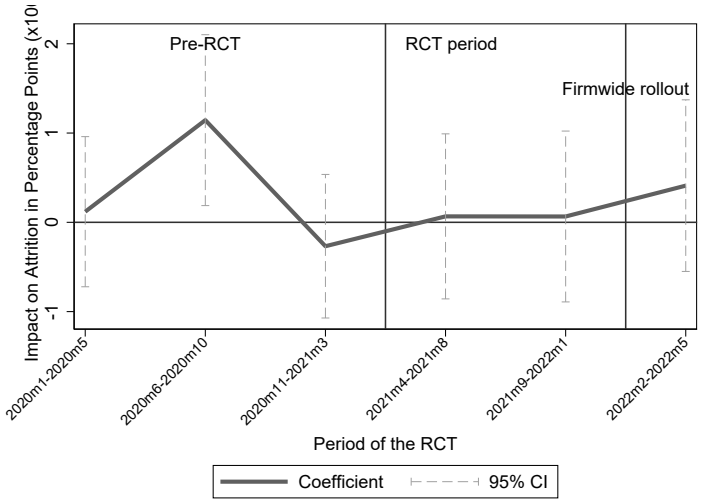
Figure A5: Differences Between Treatment and Control Stores Over Time in Trained Worker Attrition



(a) All Stores



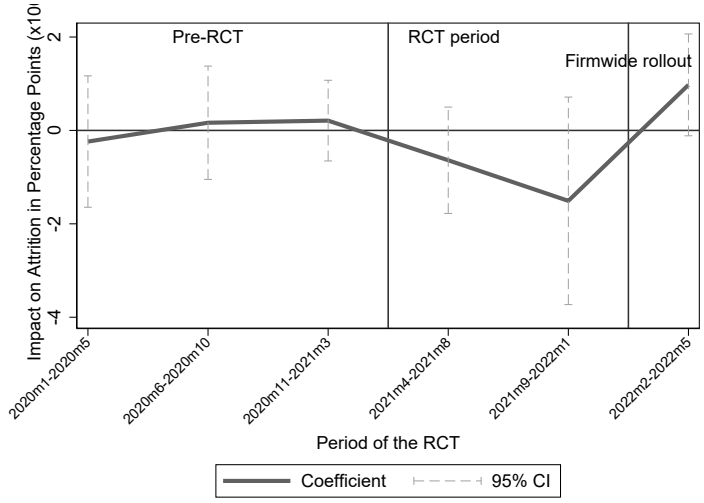
(b) Stores Where RCT Predicted to Work by Reg. Mgrs.



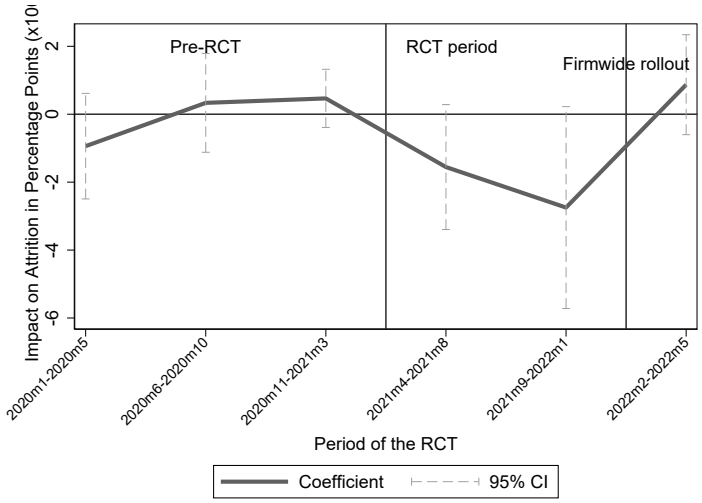
(c) Stores Where RCT Predicted Not to Work by Reg. Mgrs.

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.

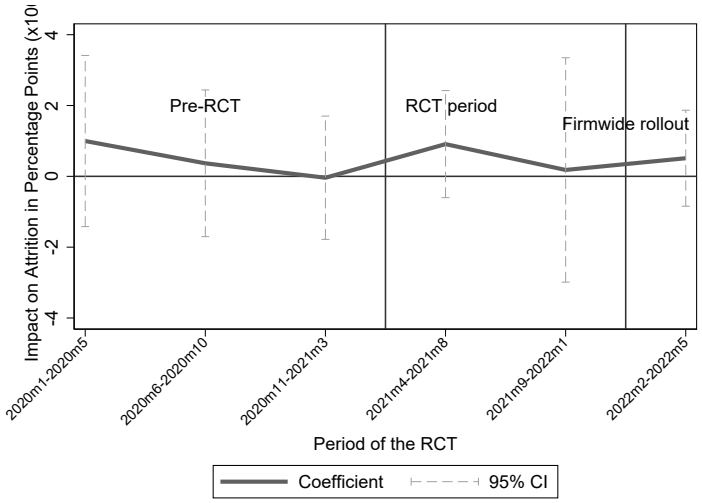
Figure A6: Differences Between Treatment and Control Stores Over Time in Store Manager Attrition



(a) All Stores



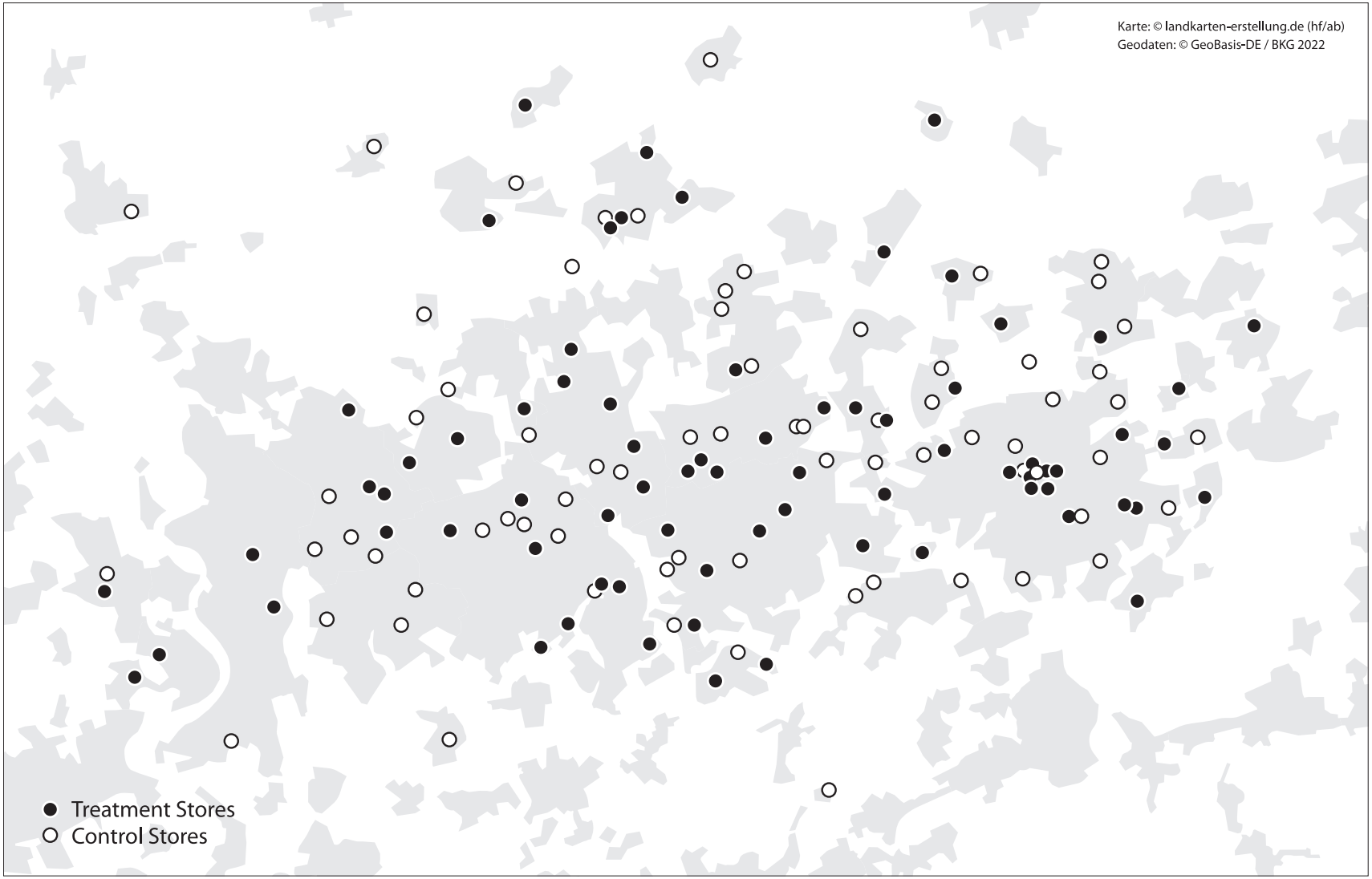
(b) Stores Where RCT Predicted to Work by Reg. Mgrs.



(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 Customers	Log Shrinkage	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 17 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 Sales Effects in Stores where RMs Predict Treatment to Work

	(1)	(2)	(3)	(4)	(5)	(6)
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
	(0.052)					(0.162)
Treat*(Pre-RCT mean head count)		-0.002				0.001
		(0.004)				(0.006)
Treat*(Pre-RCT mean tenure of workers)			-0.001			-0.007
			(0.005)			(0.006)
Treat*(Pre-RCT mystery shopping score)				-0.087**		-0.073**
				(0.037)		(0.036)
Treat*Pre-RCT Log (Shrinkage as % of Sales)					0.127	0.003
					(0.094)	(0.156)
1-sided p-val, Treat X Predict success	0.02	0.04	0.04	0	0.04	0.01
2-sided p-val, Treat X Predict success	0.04	0.08	0.07	0.01	0.09	0.03
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
	(0.394)					(0.434)
Treat*(Pre-RCT mean head count)		0.049				0.017
		(0.054)				(0.059)
Treat*(Pre-RCT mean tenure of workers)			-0.208**			-0.102
			(0.103)			(0.140)
Treat*(Pre-RCT mystery shopping score)				-0.275		-0.736
				(0.646)		(0.746)
Treat*Pre-RCT Log (Shrinkage as % of Sales)					-0.207	0.383
					(2.036)	(2.548)
1-sided p-val, Treat X Predict success	0.02	0.05	0.01	0.03	0.02	0.05
2-sided p-val, Treat X Predict success	0.05	0.09	0.03	0.05	0.04	0.10

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. 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: Regional Manager 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 reports the mean of the time that regional managers predict the treatment to work. We also report p-values of the mean equality tests across the quartiles, both unadjusted and Wesfall-Young (1993)-adjusted for multiple hypothesis testing. Standard errors are clustered by store. For trained worker attrition, most affected means the strongest negative effect on attrition. For example, column 1 shows that among stores with the least beneficial treatment effect on sales, the share of stores where the treatment is predicted to work is only 11%. However, among stores where the treatment is most beneficial for sales, the share predicted to work is 81%.

Table A11: 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 A12: 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 A11. In the predictions of regional manager 6, "focus store" refers to a small number of stores marked by top management as needing improvement.

Table A13: 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 A11.

Table A14: 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 A11.

Table A15: 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 A16: 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 A17: No Evidence that the Treatment Caused Regional Managers 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 regional managers 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 regional managers to their store in the *during-RCT store manager survey.** significant at 10%; ** significant at 5%; *** significant at 1%

Table A18: 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 Regional Managers		
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 Regional Managers		
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 A19: 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 A20: 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 A21: 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 RAs. 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 A15.

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 the variance. This suggests that the characteristics do not reflect a single underlying 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

[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.

may be fake. However, there is no reason to believe that our treatment would affect the 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.

B.11 Profit Calculation Details (Section 5.1)

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

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

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

[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.

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.

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.12 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.13 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 A20 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

In the wording, regional managers are asked how well the daily business would function with fewer checklists. A reader might be inclined to interpret this as a statement about the level of performance in a store as opposed to the treatment effect. However, in our preliminary conversations with regional managers, the idea of a treatment effect did not seem natural, and many regional managers seemed to struggle with the language of counterfactuals. The language we selected was designed to be an intuitive and natural way of capturing beliefs about where the treatment would have the largest effect.

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. *Review of Economic Studies*, **81**(2), 608–650.

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.

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.