INSTITUTE for REPLICATION

No. 118
I4R DISCUSSION PAPER SERIES

Robustness Reproducibility of “Improving Workplace Climate in Large Corporations: A Clustered Randomized Intervention”

Alice Hallman

Magnus Johannesson

Essi Kujansuu

May 2024

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 118

Robustness Reproducibility of “Improving Workplace Climate in Large Corporations: A Clustered Randomized Intervention”

Alice Hallman[1], **Magnus Johannesson**[2], **Essi Kujansuu**[3]

[1]Uppsala University, Uppsala/Sweden
[2]Stockholm School of Economics, Stockholm/Sweden
[3]University of Innsbruck/Austria

MAY 2024

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research

Hohenzollernstraße 1-3
45128 Essen/Germany

www.i4replication.org

ISSN: 2752-1931

Robustness Reproducibility of "Improving Workplace Climate in Large Corporations: A Clustered Randomized Intervention"

Alice Hallman[1], Magnus Johannesson[2], Essi Kujansuu[3]

[1] Department of Economics, Uppsala University, Uppsala, Sweden

[2] Department of Economics, Stockholm School of Economics, Stockholm, Sweden

[3] Department of Economics, University of Innsbruck, Innsbruck, Austria

Abstract

Alan et al. (2023) carry out a field experiment where they randomly allocate 20 corporations in Turkey to a treatment group or a control group. White-collar employees at the headquarters of the corporations are invited to participate in a training program to improve the workplace environment. They report that the program reduces separation (workers quitting) and improves prosocial behavior, workplace quality and support networks. We test the robustness reproducibility of these results, focusing on the results reported in Table 8 of the original paper. We first successfully reproduce the results in Table 8 computationally based on the posted code and data, and we then carry out five robustness tests. We do not find robust support for an effect of the treatment on any of the four primary outcome variables (separation, prosocial behavior, workplace quality and support networks). The relative effect size of the robustness tests averaged across the primary hypotheses is 0.62, suggesting some inflation in the original effect sizes. The effects reported in the paper are driven by the additional employees added to the sample about one year after the initial baseline data collection and after the randomization of firms to treatment and control (and this sample is not balanced on observables across the treatment and control group). Not having access to the raw data limited the possible robustness tests.

1. Introduction

We carry out an evaluation of the robustness reproducibility of the paper by Alan et al. (2023). The paper reports the results of a field experiment involving employees at 20 corporations in Turkey. The 20 corporations were randomized to either a control (n=10 corporations) or a treatment group (n=10 corporations). The treatment corporations participated in a training program trying to improve the relational atmosphere at the workplace.

The authors collected registry data about separations (employees leaving the corporation) and survey data about prosocial behavior, workplace climate and support networks after the end of the implementation period of the program (these are the primary outcome measures, they also collected additional survey data). The authors stated that the final sample with registry data consisted of 4,239 employees, and out of these 3,083 employees gave informed consent to participate in the study and this is the maximum sample size in the analyses (they further stated that they had survey data for over 2,000 employees). The sample sizes used in their analyses were 3,076 for separation, 2,233 for prosocial behavior, 2,155 for workplace climate and between 137 and 163 departments for the analyses on support networks.

For their four primary outcomes, the authors conclude in the Introduction:

"We find that the program has a substantial effect on the likelihood of employee separation, mainly at the leadership level." (page 154)

"We also find that the program significantly increases prosociality and lessens antisocial tendencies in the workplace." (page 154)

"At the departmental level, the program significantly lowers the proportion of employees lacking support and makes intradepartment support networks denser and less segregated across cohorts." (page 155)

“We then show that the program successfully improves perceived workplace quality and relational atmosphere within departments.” (page 155)

We evaluate here the robustness reproducibility on these results, by carrying out 5 robustness tests. Our analysis is limited due to the unavailability of the raw data, which was not shared by the original authors; only the processed dataset was made available. This limits the robustness checks we can conduct. Another limitation is that there was also some ambiguity in interpreting the variables in the posted data. The authors did post a pre-analysis plan (PAP) at the AEA registry (AEARCTR-0007532), but the PAP lacks details about the construction and measurement of the outcome measures, and the analyses and tests (but it does list prosocial behavior, workplace climate and support networks as the primary outcome variables): the registry data outcome is not included in the PAP but was added later based on the following motivation in footnote 11: “In need of an objective outcome (after the feedback we received in various seminars), we decided to reach out to the companies and request employee separation information.”

The lack of details in the PAP implies that the PAP is not successful in constraining the researcher degrees of freedom in the analyses. This is a limitation of the study. We found posted analysis code for the results reported in the main text, but not for the results reported in the Online Appendix (the authors say there is code, but we could not find it in the folder). For the analysis code that was available, it delivered the reported results in Table 8 for the full sample, which is the analyses subjected to robustness tests in this study (we did not systematically evaluate the computational reproducibility of the other results reported in the main text).

Below we provide a plan for our analyses. We then report the results of the individual robustness tests and summarize the robustness results separately for the four main hypotheses and together for the paper using two indicators of robustness reproducibility: the statistical significance indicator and the relative effect size indicator. We end with some conclusions about our robustness tests.

2. Plan for our robustness analyses

According to the paper the four primary outcome measures in the study are (page 165 and Figure 2): separation, prosocial behavior, workplace climate, and support networks. Three of these are also listed as primary outcome variables in the PAP and the fourth (separation) was added later as mentioned above. The authors use several indicators for these variables (except for separation that is only one indicator) and in Table 8 they report results based on summary indices for the different measures (except for the support networks primary outcome variable). We interpret these results as the main findings of the study and base our robustness tests on the results for separation, prosocial behavior, and workplace climate in Table 8. Table 8 also includes results for an additional summary index in the last column for the outcome variable "leadership quality", but as that is defined as a secondary outcome measure in the PAP we do not include it among the main findings of the paper (and we therefore do not include robustness tests for this secondary outcome measure). The authors do not report results for a summary index of the support networks primary outcome variable, although they test for 6 different indicators of this outcome variable in Table 5. We therefore include a robustness test of a summary index measure of this outcome variable, although this result has no baseline result to be compared with in the paper. In total, we carry out 5 robustness tests detailed below.

As corporations rather than individuals were randomized to the treatment and control groups, the authors cluster standard errors at the corporation level (n=20). As 20 is a small number of clusters for reliable results with standard clustering, the authors also report results for Wild bootstrap clustering. We interpret these test results as the main test results. In all our robustness tests, we therefore use the Wild bootstrap clustering command in Stata and report the relevant regression coefficient and the Wild bootstrap p-value. As this Wild bootstrap routine in Stata does not report standard errors of the regression coefficients and the reported t-value is the one based on the "standard clustering command" (and does therefore not match the p-value) we do not report standard errors or t-values. Below we detail our robustness tests and the results.

In reporting the results of the individual robustness tests below, we interpret a robustness test with a p-value <0.05 and an effect in the same direction as "robust" and other results as "not robust".

This is in line with using this indicator, the statistical significance indicator, as one of our two summary indicators of robustness reproducibility (see more on this below in that section). Robust here implies the robustness of the conclusion in the original study that the study provides statistically significant or strong support of the tested hypotheses. The significance threshold used in the study is not mentioned in the PAP nor in the paper but the authors distinguish between significance at the 1%, 5%, and 10% in their regression tables. For the three main results in Table 8, which we assess in our robustness tests, two have a p-value below 0.05 in the Wild bootstrap test, whereas the third has a p-value below 0.10.[1] Although the third result (workplace climate) does not have an original p-value <0.05, we interpret the original study as reporting that the results from their study supports this hypothesis (see the quote from the original study on this in the Introduction above). We thus interpret the original authors as claiming that they find support for all their four primary hypotheses, and we test if these conclusions are robust.

3. Robustness tests of main results

The results of the robustness tests are reported in Table 1.

4.1 Robustness test 1

The original authors did not report results for a summary index for the support networks primary outcome variable; but only report results for 6 different indicators of support networks in Table 5 (but they reported results for summary indices of the other primary outcome variables with multiple indicators). In our first robustness test, we therefore provide a test of this hypothesis using a summary index of the 6 indicators of support networks. In constructing this summary index, we use the same methodology as the original authors use for their summary indices for some of the dependent variables in Table 8. We reversecode the two department density variables as they have hypothesized signs in the opposite direction of the other four support network indicators, and

[1] We think using significance at the 10% level should be avoided as it has low evidentiary value, and even the 5% level does not represent strong evidence; see Benjamin et al. (2018) that refer to p<0.05 as “suggestive evidence” and propose using p<0.005 for “statistically significant evidence”.

thereafter we z-standardize (deduct the mean and divide by the standard deviation) the six indicators and take the mean of these standardized indicators as the summary index. We control for the same variables in this regression analysis as the control variables used by the original authors in Table 5 in the original paper.

We find no significant evidence in support of this hypothesis (p=0.151). Note that this robustness test is different from our other robustness tests as we cannot compare it to a specific baseline result in the original paper (if this test had been reported in the original paper it would have provided a baseline result for additional robustness tests). This robustness test suggests that the original author's conclusions about this primary outcome measure is not robust.

4.2. Robustness test 2

Robustness tests 2-5 are all carried out on the three primary outcome variable results reported in Table 8 of the original paper.

In robustness test 2, we test the robustness of the results without the observations that were added after the randomization of firms. On page 160 the authors write: “After baseline data collection, we randomly assigned 10 corporations to treatment and 10 to control. Our initial plan was to implement the intervention in early 2020.”

The baseline data was collected in the fall of 2019 according to page 160 and the randomization of firms carried out after this, but in the fall of 2020, after the randomization of firms, an additional sample of employees was added to the study. On page 173, the original authors write:

“In the course of a single year, many changes took place in the firms, and when we decided to implement the program in fall 2020, we found that a large number of additional employees (some recently joined their firms) expressed their willingness to participate, both in treatment and control firms. Before the program rollout, we conducted a swift baseline for these new participants, a shorter version of our initial baseline. These new employees comprise 32% of our evaluation sample, and their distribution across treatment status is balanced (p-value = .59).”

In this robustness test we drop the observations added after the randomization of firms. We do this by using the variable "part_base" for participating in the baseline data collection and only include individuals coded as 1 on this variable in this robustness test. Some participants are only included in the second participation variable called "part" and we assume that these constitute the added participants; but there is some ambiguity about the definition of the variables in the posted data.

The original conclusion is not supported for any of the three primary outcome measures in this robustness test.

We furthermore compare the values of the individual-level control variables used in the Table 8 results in the original paper between the treatment and control group for the sample added after randomization of firms (defined as the individuals coded as 0 on the "part_base" variable). We report these balance tests in Table 2 using the same method as used by the original authors in their balance tests in Table 2 of the original paper; an OLS regression with sector fixed effects and clustering by firm. We report the Wild bootstrap p-values of these balance tests to be consistent with the use of Wild bootstrap p-values in the other tests in this report (the original authors did not report Wild bootstrap p-values for their balance tests in Table 2 in the original paper). We find a p-value<0.05 for 3 out of these 7 balance tests, suggesting that these observables are not balanced across the treatment and control groups. This suggests selection bias in this sample of employees added after the randomization of firms.

4.3 Robustness test 3

There is ambiguity about what control variables will be included in the analysis in the PAP and the motivation for the included control variables in the paper is also unclear (some of the measures collected at baseline such as Raven score and Eyes score are included whereas others like risk taking, competitiveness and cooperation are not). The following control variables included in the Table 8 regressions in the original paper are not explicitly mentioned in the PAP: age, male, married, number of children, department male share, tenure. We therefore carry out a robustness test dropping those control variables. This robustness test supports the original conclusion for the prosocial behavior outcome measure (p<0.05 and an effect in the original direction), but not for the other two primary outcome measures.

4.3 Robustness test 4

In robustness test 4, we further test the robustness of the included control variables and add the following variables measured at baseline to the control variables: risk taking, competitiveness, and cooperation. Given the attention in the paper to reporting separate results for the leaders and subordinates subsamples, we furthermore include a dummy variable for leaders/subordinates (leaders=1, subordinates=0). The original conclusion is not supported for any of the three primary outcome measures in this robustness test, but it should be noted that the sample size (individual observations) is also reduced substantially by more than 50% due to missing observations on risk taking, competitiveness, and cooperation. This seems to be due to the risk taking, competitiveness and cooperation variables only being measured at baseline, in the initial sample of the study, and not in the additional sample added after the randomization of firms (the additional sample that we excluded in Robustness test 2). This implies that Robustness tests 2 and 4 are both carried out excluding the observations of the sample that were added after randomization of firms (in deciding to conduct Robustness test 4, we were not aware that this robustness test would also imply excluding the sample of employees added after the randomization of firms, as we could not find information in the paper about risk taking, competitiveness and cooperation not being part of the baseline measures for the sample added after the randomization of firms).

4.5 Robustness test 5

In a final robustness test of the included control variables, we only add the leader/subordinates dummy, and thereby avoid losing observations compared to Robustness test 4. This robustness test supports the original conclusion for the separation and the prosocial behavior outcome measures (p<0.05 and an effect in the original direction), but not for the workplace climate outcome variable.

5. Robustness indicators

To summarize the results of our robustness tests, we report the results for two robustness indicators for results reported as statistically significant in the original study: the statistical significance indicator and the relative effect size indicator (these indicators were proposed by Dreber and Johannesson, 2023).

The statistical significance indicator is defined as the fraction of robustness tests that are significant at the 5% level with an effect in the same direction as the original. This is the indicator we used to interpret the results of each individual robustness test above. This indicator indicates how robust the conclusion is on whether the original hypothesis is supported or not. We estimate this for the robustness tests of the four primary outcome measures (with four robustness tests for three of these outcome measures, but only one for the support networks outcomes variable); but we also estimate it at the paper level, based on the average of the indicators for each outcome variable (we report this average both with and without the support networks variable, as we only carried out one robustness test of the results for this primary outcome variable).

The relative effect size indicator for one hypothesis is estimated as the average effect size of all the robustness tests of that hypothesis divided by the original effect size. This measure is reported for three of the primary outcome variables, but not for the support networks outcome variable as the original effect size is not reported for that result in the original paper. We aggregate the result on the paper level by taking the average of the relative effect size for each of the three primary outcome variables. The relative effect size indicator is an indicator of the systematic bias in original effect sizes; if the indicator is below 1, this suggests systematically overestimated effect sizes in the original paper. The results for the robustness reproducibility indicators are shown in Table 3.

For separation the statistical significance indicator is 0.25 and the relative effect size indicator is 0.76. This suggests that there is not robust support for these hypotheses, and that the effect sizes in the original study are somewhat inflated.

For prosocial behavior, the statistical significance indicator is 0.5 and the relative effect size indicator is 0.56. This suggests that the support for this hypothesis is not robust, although there is support for the hypothesis in two of the robustness tests, and that the effect size in the original study is inflated.

For workplace climate, the statistical significance indicator is 0 and the relative effect size indicator is 0.54. This suggests that there is no robust support for this hypothesis and that the effect size in the original study is inflated.

For support networks, the statistical significance indicator is 0 and the relative effect size indicator cannot be estimated. This suggests that there is not robust support for this hypothesis.

Aggregated on the paper level, the statistical significance indicator is 0.19 or 0.25 depending on if the result for support networks is included or not, and the relative effect size indicator is 0.62. This suggests low robustness reproducibility in terms of the support of the tested primary hypotheses, and that the effect sizes are inflated in the original study.

6. Concluding remarks

We have only conducted a limited number of robustness tests, implying our results should be interpreted cautiously. Not having access to the raw data limits the robustness tests that can be conducted. Our results should also be interpreted cautiously due to some ambiguity in interpreting the variables in the posted data. Overall, our results do not suggest robust support for an effect of the program on any of the four primary outcome variables (separation, prosocial behavior, workplace quality and support networks). The relative effect size of the robustness tests averaged across the primary hypotheses is 0.62, consistent with some inflation in the original effect sizes. The effects reported in the paper are driven by the additional employees added to the sample about one year after the initial baseline data collection and after the randomization of firms to treatment and control. This sample is not balanced on observables across the treatment and control group, suggesting selection bias in the added sample.

References

Alan S, Corekcioglu G, Sutter M. Improving workplace climate in large corporations: A clustered randomized intervention. Quarterly Journal of Economics 2023;138:151-203.

Benjamin D, et al. Redefine statistical significance. Nature Human Behaviour 2018:2:6-10.

Dreber A, Johannesson M. A framework for evaluating reproducibility and replicability in economics. SSRN working paper 2023.

Table 1. Results of robustness tests of the summary indices of outcomes. The original results reported in Table 8 of the original paper for separation (implementation), prosocial behavior, and workplace climate; the results for the summary index of support networks was not reported in the original paper but estimated by us for the six "support network" outcome variables reported in Table 5 in the original paper. The treatment coefficient and the wild-bootstrap p-value for the treatment coefficient reported in the table (with the p-value reported in parenthesis). NA=not applicable.

	Separation (implementation)	Prosocial behavior	Workplace climate	Support networks
Original results, Table 8, Treatment coefficient	**-0.022 (0.029) n=3,076**	**0.097 (0.002) n=2,233**	**0.198 (0.098) n=2,155**	NA
Robustness test 1: Summary index for support networks.	NA	NA	NA	-0.208 (0.151) n=137
Robustness test 2: Dropping employees added after the randomization of firms (page 173).	-0.0137 (0.3193) n=1,625	0.0238 (0.3614) n=894	0.0381 (0.7347) n=884	
Robustness test 3: Excluding the following control variables, as ambiguous in PAP if they should be controlled for: age, male, married, number of children, department male share, tenure.	-0.018 (0.138) n=3,076	0.081 (0.005) n=2,233	0.175 (0.089) n=2,155	
Robustness test 4: Adding the following control variables: leader/subordinates dummy variable (leader=1), risk taking, competitiveness, and cooperation.	-0.012 (0.410) n=1,472	0.016 (0.505) n=838	0.023 (0.840) n=831	
Robustness test 5: Adding leader/subordinates dummy variable (leader=1).	-0.023 (0.029) n=3,076	0.095 (0.001) n=2,233	0.195 (0.105) n=2,155	

Table 2 (not a robustness test). Balance tests of individual-level characteristics in the sample of employees added after the randomization of firms (the sample of employees excluded in Robustness test 2). All individual-level characteristics included as control variables in Table 8 in the original paper included in the balance tests. Tested in OLS regressions with sector fixed effects and clustering by firm (the same method used by the original authors in the balance tests in Table 2 of the original paper; with the exception that we report Wild bootstrap p-values as for the other tests with clustering by firm in this report).

Individual characteristic	N	Control Mean	Treatment Mean	Difference (T-C)	Wild bootstrap p-value
Raven Score	1453	-0.045	0.143	0.206	0.1502
Eyes Score	1453	0.036	0.350	0.324	0.0030
Age	1453	36.915	34.826	-1.720	0.0270
Male	1453	0.804	0.739	-0.074	0.3684
Married	1453	0.691	0.567	-0.134	0.0571
Kids	1453	1.027	0.731	-0.262	0.0290
Tenure (yearly)	1453	8.065	5.579	-0.813	0.5776

Table 3. Robustness reproducibility indicators for the four summary indices outcome measures and pooled for the paper (the average of the four summary indices outcome measures). For "support networks" we do not report results for the relative effect size as baseline results were not reported in the paper for this summary index. NA=not applicable.

Robustness reproducibility indicator	Separation (implementation)	Prosocial behavior	Workplace climate	Support networks	**Pooled for the paper**
Fraction significant at 5% level, in original direction	0.25	0.5	0	0	0.19 (0.25 if support networks not included as no baseline result).
Relative effect size	0.76	0.56	0.54	NA	0.62