INSTITUTE for REPLICATION

No. 17
I4R DISCUSSION PAPER SERIES

A Reply to Comment by Bonander et al. (2023)

Onur Altindag

Bilge Erten

Pinar Keskin

February 2023

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 17

A Reply to Comment by Bonander et al. (2023)

Onur Altindag[1], Bilge Erten[2], Pinar Keskin[3]

[1]*Bentley University, Waltham MA/USA*
[2]*Northeastern University, Boston MA/USA*
[3]*Wellesley College, Wellesley MA/USA*

FEBRUARY 2023

ISSN: 2752-1931

A Reply to Comment by Bonander et al. (2023)

Onur Altindag
Bentley University

Bilge Erten
Northeastern University

Pinar Keskin
Wellesley College

January 30, 2023

Abstract

In Altindag et al. (2022), we estimate the effects of an age-specific lockdown policy on mobility and mental health outcomes among adults aged 65 and older in Turkey using a regression discontinuity design. Bonander et al. (2023) successfully replicate all our main findings. They argue that the estimates for mobility outcomes are all robust to alternative sensitivity checks while some of the estimates for mental health—which were statistically significant around the 5-9 percent level—lose significance at the conventional level of 10 percent in the more conservative specifications. In this reply, we provide approximately 7,000 additional estimates that comprise a near universe of RD estimates for all our outcomes, each possible monthly bandwidth, and each possible combination of covariate adjustment, kernel selection, estimation methodology, standard error adjustment, and kernel weighting selection. This comprehensive analysis shows that our original results are robust to these choices. We show that Bonander et al. (2023) rely on a selection of very narrow bandwidths that produce highly sensitive and uninformative estimates due to overfitting. We also show that Bonander et al. (2023) report imprecise estimates, which are outliers in the distribution of all estimates that can be reported. We conclude that broader statistical tests are more informative for robustness checks.

1. Introduction

In Altindag et al. (2022), we estimate the effects of an age-specific lockdown policy on mobility and mental health outcomes among adults aged 65 and older in Turkey, using a regression discontinuity design. Bonander et al. (2023) successfully replicate our findings in their report. They also test the robustness of our results by using an alternative standard error adjustment, not including any control variables, and calculating the optimal bandwidth using a different methodology. They argue that the estimates for mobility outcomes are robust to these checks while those for mental health—which were statistically significant around the 5-9 percent level after multiple hypothesis testing adjustment—tend to lose significance at the conventional level of 10 percent when more noise was added following these modifications to the estimating specification.

In this response, we show that our mental health results are remarkably stable across more than 7,000 combinatorial choices of bandwidth, covariate adjustment, kernel selection, estimation methodology, standard error adjustment, and kernel weighting selection. We explain our sensitivity checks in greater detail in the next section by focusing on sensitivity to selection of (a) different bandwidths and estimation methodology, (b) standard error adjustment, and (c) covariate inclusion, which cover all dimensions of tests implemented by Bonander et al. (2023). We show in each case that Bonander et al. (2023) either report uninformative estimates due to overfitting or specifications that add sizeable noise to the estimated effects. More general tests of sensitivity show that our estimates are robust even under quite restrictive empirical specifications.

2. Sensitivity Analysis

2.a. Bandwidth Selection and Estimation Methodology

We start our response by addressing the bandwidth selection and estimation procedure. Bonander et al. (2023) argue that our results are sensitive to the bandwidth choice and the estimation strategy. To show their point, they use ***rdrobust***, which is a STATA package based on Calonico, Cattaneo and Titiunik's (CCT) several papers (Calonico et al. 2015, 2017). The package produces three estimates: (i) conventional RD estimates with conventional variance estimator, (ii) bias-corrected RD estimates with conventional variance estimator, and (iii) bias-corrected RD estimates with robust variance estimator. Bonander et al. (2023) only report (iii), which is the most conservative estimate that the package can produce. The CCT routine also uses an automated procedure to estimate the optimal bandwidth for estimation, which might vary substantially based on the outcome, covariates, functional form, and the type of kernel that is used for weighting the regression. Card et al. (2015), for example, fully omit the regularization term in the bandwidth selectors as it leads to narrow bandwidths with no change in asymptotic properties of the estimator. This translates into fully offsetting scaling factor for the regularization term to zero in the *rdrobust* package whereas the default option is one, which routinely selects a narrow bandwidth. Additional automated algorithm selection routines, such as Ludwig and Miller (2007) and Imbens and Kalyanaraman (2012), exist and may also provide significantly different estimation bandwidths. It is a standard exercise for any rigorous study to provide comparable RD estimates for a wide range of bandwidths and outcomes, as we did in the original study.

In this response, we provide additional RD estimates using the same CCT package for a combination of **all** possible primary outcomes (mobility and mental health), **any** bandwidth selection between 9 and 60 months, **all available** CCT estimation procedures (e.g., conventional, bias-corrected, robust & bias corrected), and kernel weights (e.g., uniform and triangular).

First, we incrementally expand Bonander et al.'s (2023) computer code to broader ranges of bandwidth options using the original specification.[1] We visualize the distribution for the CCT's most conservative RD estimates (bias-corrected and robust) for each outcome, bandwidth selection, and kernel-weight selection in Figure 1. The pink area depicts the bandwidth range that Bonander et al. (2023) show in their replication report. The white represents the results for the bandwidth reported in the original study. Regardless of the outcome, using only a very small number of monthly cohorts on each side of the discontinuity simply overfits the data and, as expected, produces highly unstable treatment effects. A one-month increase/decrease in the bandwidth changes the estimates significantly in this narrow range, where the kernel choice also makes a substantial difference, even for the first-stage outcomes and corresponding standard errors. It is important to note that the age-specific lockdowns were enforced by local security

[1] To construct the automated bias-corrected RD estimates, we use a bias bandwidth (b) that is twice of the main bandwidth (h) (rho = 0.5) which roughly matches the bias-bandwidth that the CCT automated package picks.

forces, and even a quick glance at the raw data can validate the first stage estimates that we present in the original study (Figure 2). Our conclusion therefore is that estimates within extremely narrow bandwidths are uninformative and lack statistical validity. The bandwidth selection outside this narrow range, however, provides very consistent estimates across any bandwidth and kernel weighting methodology that we could select and report, as Figure 1 clearly shows in the graphs for each outcome.

In Tables 1 and 2, we provide a numerical summary of this robustness exercise and compare it to the estimates reported in the original study and to the reported estimates in Bonander et al. (2023). In panel (1) of both tables, we show the average RD estimate for each of our outcomes by CCT estimation methodology and kernel choice for the bandwidth region of 17-60 months, which is the range that we also report in the original study following requests by the editors and the referees. Row (a) shows the average RD estimate, whereas row (b) shows the average clustered standard errors. These results are in line with Figure 1, showing statistically identical results independent of the estimation methodology, kernel choice or bandwidth choice compared to the original study. Bonander et al.'s (2023) estimates are sensitive to these parameter choices even within their reported estimates. We conclude that the robustness checks that Bonander et al. (2023) report are uninformative estimates that neither verify nor refute the robustness of our initial findings.

2.b. Robust Standard Errors

Bonander et al. (2023) also argue that some of the estimates become "insignificant" when they use robust standard errors for one specification that relies on a bandwidth of 45 with a uniform kernel. In Tables 1 and 2, we also report all the robust standard errors (HC1) and the corresponding average t-statistics that result from all possible bandwidth selections, estimation procedures, and kernel selections. As underlined by Bonander et al. (2023), robust standard errors are slightly larger on average than clustered standard errors, but they have no meaningful effect on the first or second-stage outcomes reported in the original study. For example, CCT's *rdrobust* package estimates for mental distress range from 0.262 to 0.286, with corresponding robust standard errors ranging from 0.109 to 0.132. We can confidently reject the null hypothesis of zero effect at any standard threshold level for most of these estimates. The same holds true for all the other outcomes. Given the standard and recommended approach in most RD studies in the empirical literature, we reported the clustered standard errors in the original study. Our conclusions remain unchanged compared to what we initially reported in the study.

2.c. Inclusion of Covariates

Finally, Bonander et al. (2023) remove all the covariates from our main specification and conclude that the results again become “insignificant” for mental health outcomes. First, the inclusion of covariates matters for not only producing precise estimates, but also for netting out potential confounders due to measurement error. For example, the inclusion of interviewer fixed effects is important to isolate potential reporting effects resulting from interviewer-related issues. This is a standard approach in any fieldwork-based empirical study. Moreover, one could also worry about the lockdown enforcement being slightly different across provinces due to the intensity of pandemic and existing police force; hence, adding province fixed effects isolates any such potential confounding effects. Controlling for basic demographic characteristics such as education and gender also improve precision, which is especially important in an empirical setting where statistical power is limited, such as producing an average RD treatment effect.

Nevertheless, to assess the sensitivity of our estimates with respect to covariate inclusion, we estimate RD effects based on all possible combinations of covariates and specifications that vary kernel weighting. There are 62 different covariate combinations that we could select from a list of 6, excluding the full set of covariates and no covariates at all. We report the results for the bandwidth of 45 months, though as shown before this choice also makes no difference to any of the findings.

Figure 3 shows the distribution of the most conservative RD estimates (robust and bias-corrected) reported in the *rdrobust* package. Almost all estimates for the mental health outcomes indicate a positive coefficient (deterioration of mental health), roughly evenly distributed around the estimates that we report in the original study (blue dashed line). The red dashed line shows the estimates reported with no covariates with uniform kernel by the replicators.

These findings indicate that the original study's point estimates are not outliers among all the possible coefficients that could be reported. Furthermore, for any of our outcomes, covariate adjustment does not have a strong impact on point estimates. The variation in mental health outcomes is greater, but this is to be expected given the second-stage nature of these outcomes. As a result, we draw the conclusion that covariate correction functions only to improve precision.

Figure 4 depicts the distribution of t-statistics for these estimates (in absolute values). Again, the t-statistics reported in the original study (blue dashed line) are not outliers in any covariate adjustment distribution that we could choose to report. The dashed red line, however, shows that the replicators' zero-covariates estimate falls in the left tail of the distribution. We conclude that removing all the covariates is not an informative robustness check given that point estimates are not sensitive to covariate adjustment, the substantial power requirement for an RD design, and the need to adjust for measurement error through survey or province- fixed-effects. It is important to note that in Altindag et al. (2022), we are upfront about the study's statistical power by reporting a wide range of estimates and acknowledging the uncertainty surrounding these estimates. Table 3 similarly shows the averages from these estimates.[2] The choice of estimation methodology in addition to covariate selection also makes no difference to our findings.

3. Concluding Remarks

In Altindag et al. (2022), we used a RD design to document the effects of an age-specific lockdown policy on mobility and mental health outcomes among adults aged 65 and older in Turkey. Bonander et al. (2023) successfully replicated all our original findings.

Based on only 16 alternative specifications, Bonander et al. (2023) argue that some of the estimates for mental health outcomes lose significance at conventional levels. In this paper, we provide approximately 7,000 additional estimates, which comprise a near universe of RD estimates for all our outcomes, based on each possible monthly bandwidth and each possible combination of covariate adjustment, estimation methodology, and kernel type. This comprehensive analysis shows that our original results are remarkably robust to these various adjustments.
We also show that Bonander et al. (2023) report point estimates that are outliers in the distribution of all estimates that could ostensibly have been reported. When narrow bandwidths are used, there

[2] Out of 868 estimations, *rdrobust* package is unable to estimate 6 specifications.

are clear patterns of overfitting. This generates uninformative estimates that substantially differ across specifications, even for the first-stage outcomes. Removing all covariates reduce statistical power, with no identification argument for doing so. As we show, covariate adjustment improves precision with little impact on point estimates. One lesson we can draw from this exercise is that a robustness check of results for a given study should rely on a comprehensive battery of statistical tests rather than a small number of alternative specifications.

Finally, the American Economic Association has long moved away from emphasizing coefficients to designate significance at 10, 5, or 1 percent levels. One of the reasons for doing so was the arbitrary designation of an estimate with a p-value of 0.09 as statistically significant while that with a p-value of 0.11 as being statistically insignificant. The aim was to reduce the incentives of researchers to undertake “p-hacking.” Looking at the full spectrum of estimates we could report, one can clearly see that neither the point estimates nor their t-statistics in the results of the paper are by no means outliers based on all possible choices of bandwidths, estimation methods, covariate combinations, and kernel choices.

We appreciate the effort put by Bonander et al. (2023) who thoroughly replicated our findings. All our results, along with any pertinent interpretations and discussions, in the original study remain unchanged.

References

Altindag, Onur, Bilge Erten, and Pinar Keskin. 2022 "Mental health costs of lockdowns: Evidence from age-specific curfews in Turkey." American Economic Journal: Applied Economics 14, 2: 320-43.

Bonander, Carl, Gabriella Chauca Strand, and Niklas Jakobsson. 2023. "Direct replication and additional sensitivity analyses for Altindag et al. (2022): A replication report from the Oslo Replication Games"

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2015. "Rdrobust: an R package for robust nonparametric inference in regression-discontinuity designs." R J. 7, 1, 38.

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell, and Rocio Titiunik. 2017. "rdrobust: Software for regression-discontinuity designs." The Stata Journal 17, 2: 372-404.

Card, David, David S. Lee, Zhuan Pei, and Andrea Weber. 2015. "Inference on Causal Effects in a Generalized Regression Kink Design." Econometrica 83 (6): 2453–83.

Imbens, Guido, and Karthik Kalyanaraman. 2012. "Optimal bandwidth choice for the regression discontinuity estimator." The Review of economic studies 79, 3: 933-959.

Ludwig, Jens, and Douglas L. Miller. 2007. "Does Head Start improve children's life chances? Evidence from a regression discontinuity design." The Quarterly journal of economics 122, 1: 159-208.

FIGURE 1: Bandwidth Sensitivity

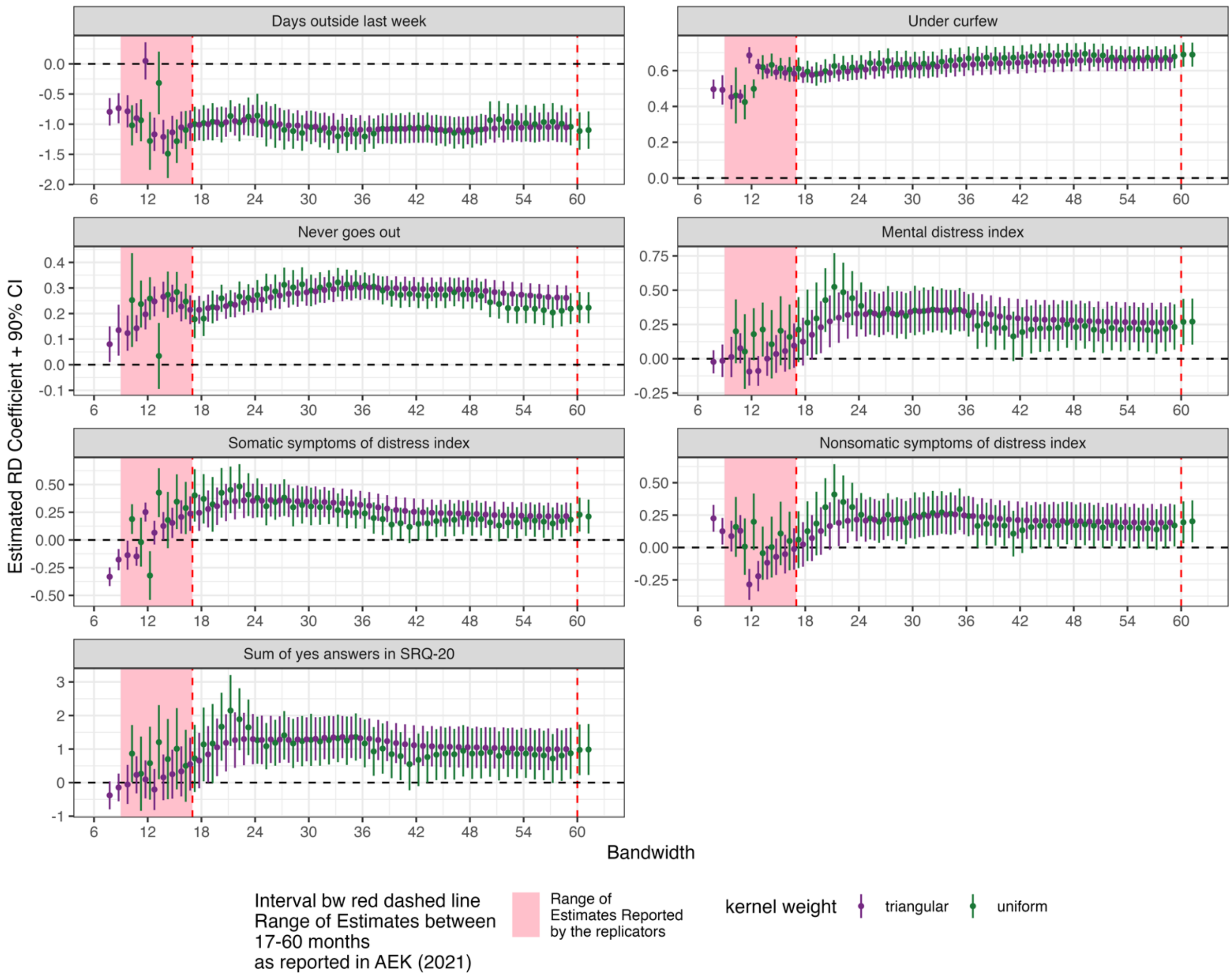


Figure 2. RD Treatment Effects on Mobility Outcomes from the Original Study

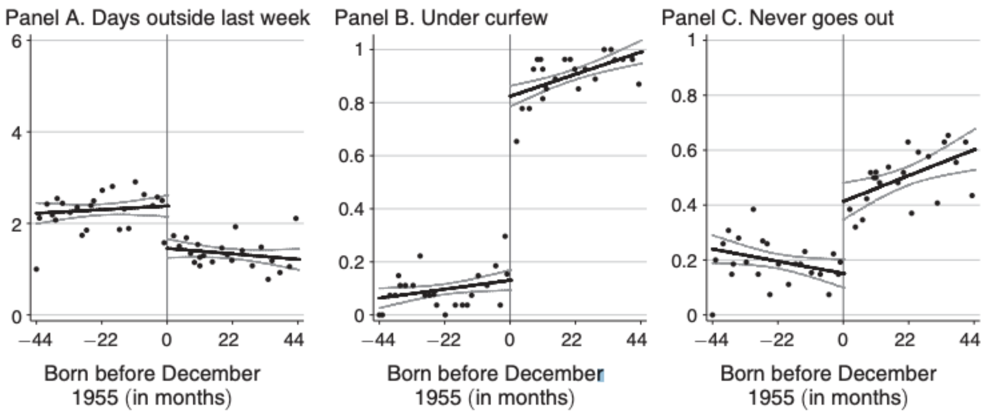
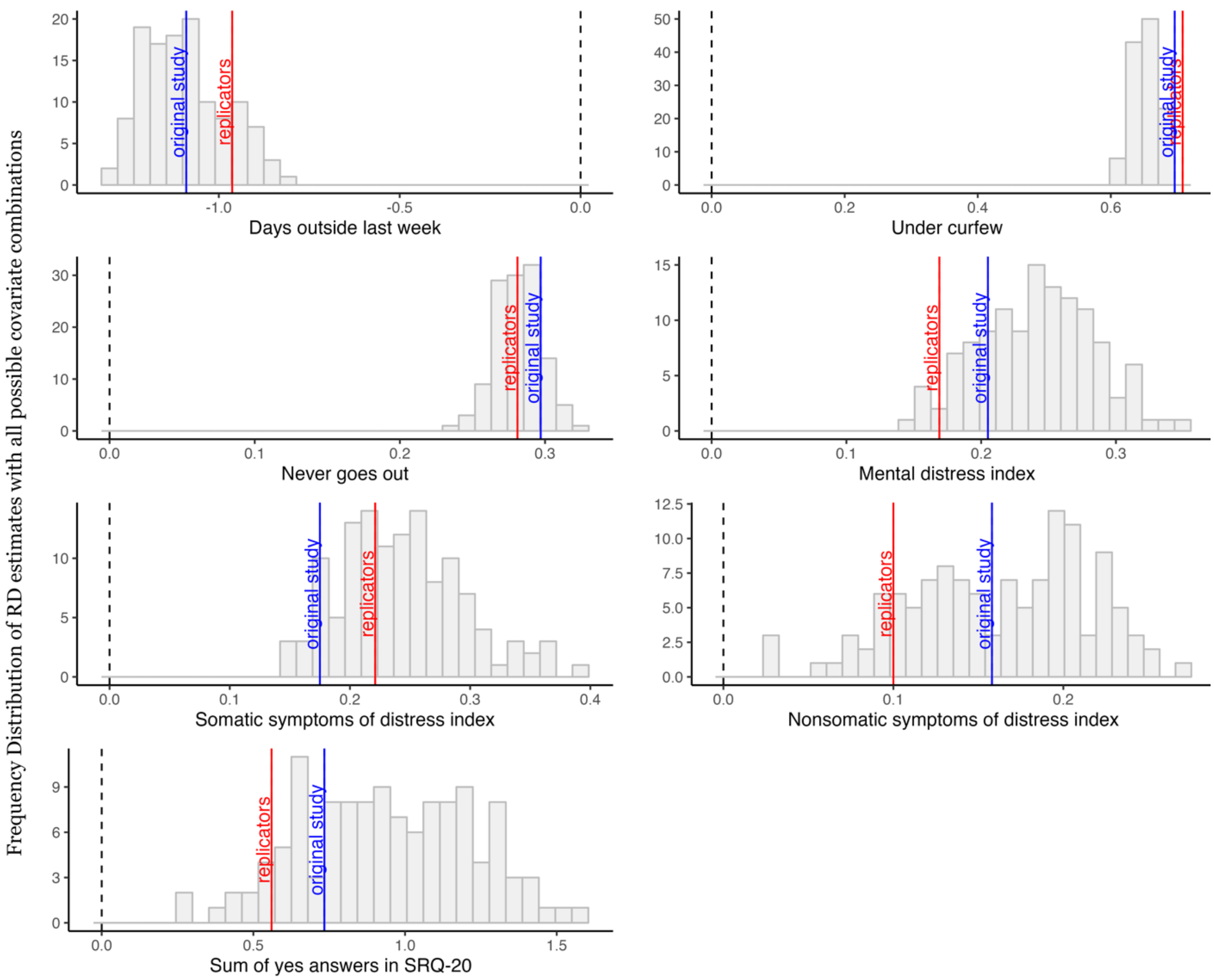
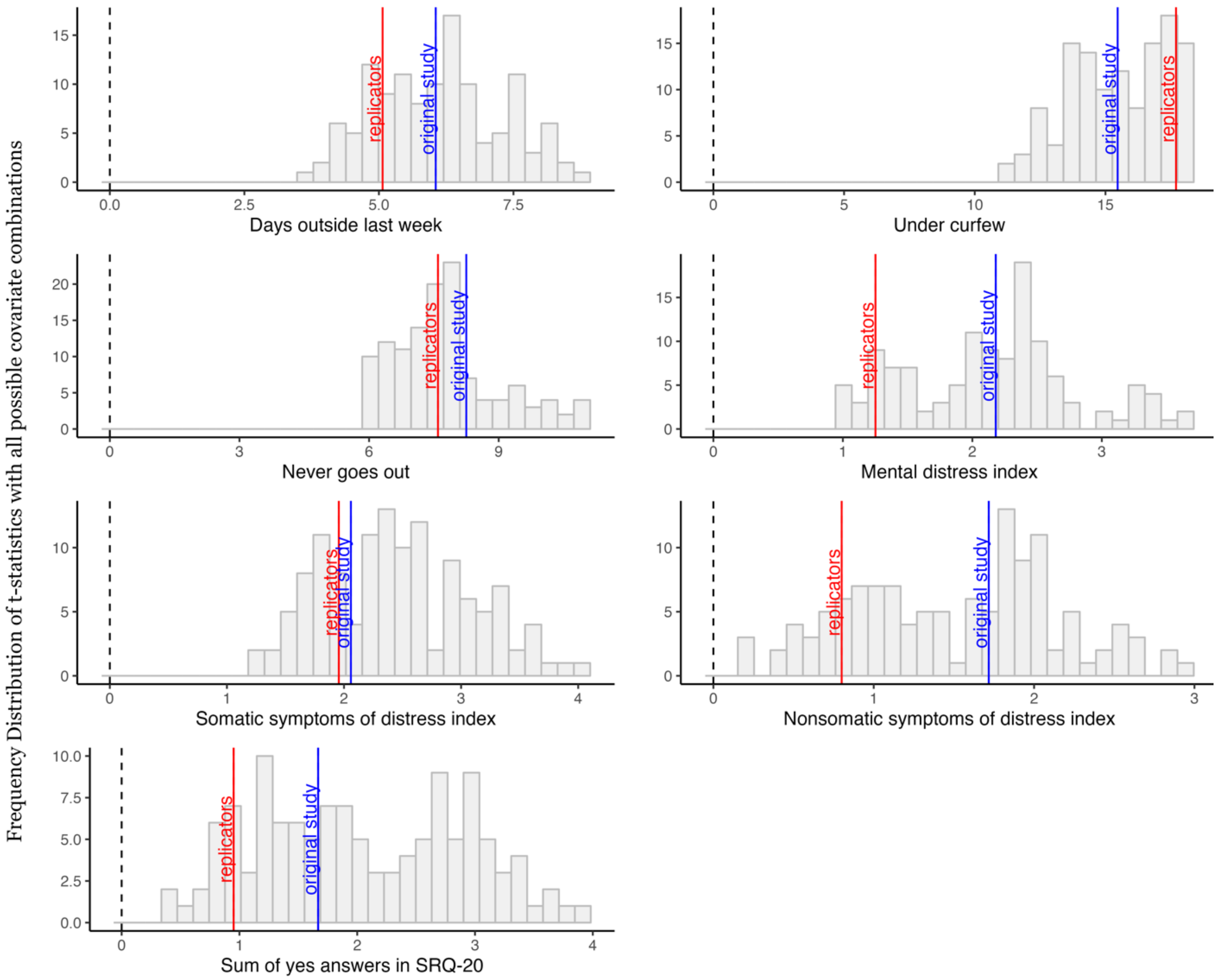


FIGURE 3: Sensitivity to the Inclusion of Covariates, Frequency Distribution of RD Estimates



Notes: *The figures show the frequency distribution of RD coefficient estimates for each outcome with all possible combinations of covariates for the bandwidth of 45 months using the rdrobust package. The blue line shows the coefficient estimate reported in Altindag et al. (2022). The red line shows the coefficient estimate without any covariates reported in Bonander et al. (2023).*

FIGURE 4: Sensitivity to the Inclusion of Covariates, Frequency Distribution of t-statistics



Notes: T*he figures show the frequency distribution of t-statistics for RD regressions using each outcome with all possible combinations of covariates for the bandwidth of 45 months using the rdrobust package. The blue line shows the coefficient estimate reported in Altindag et al. (2022). The red line shows the coefficient estimate without any covariates reported in Bonander et al. (2023).*

Table 1: Bandwith Sensitivity - First Stage

	All possible bandwidth selections						Bonander et al. (2022)			Original study
RD methodology	CCT conventional		CCT bias corrected		CCT bias corrected and robust		CCT bias corrected and robust		conventional (OLS)	conventional (OLS)
Kernel	Triangular	Uniform	Triangular	Uniform	Triangular	Uniform	Triangular	Uniform	Uniform	Uniform
Bandwidth	17-60	17-60	17-60	17-60	17-60	17-60	17.139	9.478	45	45
Outcome: Days outside last week										
(a) RD coefficient	-1.022	-1.014	-1.045	-1.049	-1.045	-1.049	-1.087	-1.019	-1.089	-1.089
(b) Clustered standard error	0.124	0.136	0.124	0.136	0.149	0.176	0.168	0.202		0.179
(c) Robust standard errors (HC1)	0.181	0.173	0.181	0.173	0.208	0.211			0.189	
t-stat (a)/(b)	-8.24	-7.46	-8.43	-7.71	-7.01	-5.96	-6.47	-5.045		-6.08
t-stat (a)/(c)	-5.65	-5.86	-5.77	-6.06	-5.02	-4.97			-5.76	
Total number of RD estimates	88	88	88	88	88	88	1	1	1	1
Outcome: Under curfew										
Bandwidth	*17-60*	*17-60*	*17-60*	*17-60*	*17-60*	*17-60*	*16.666*	*11.44*	*45*	*45*
(a) RD coefficient	0.647	0.679	0.629	0.659	0.629	0.659	0.588	0.531	0.708	0.708
(b) Clustered standard error	0.030	0.033	0.030	0.033	0.035	0.040	0.031	0.035		0.044
(c) Robust standard errors (HC1)	0.042	0.038	0.042	0.038	0.048	0.046			0.039	
t-stat (a)/(b)	21.57	20.58	20.97	19.97	17.97	16.48	18.97	15.17		16.09
t-stat (a)/(c)	15.40	17.87	14.98	17.34	13.10	14.33			18.15	
Total number of RD estimates	88	88	88	88	88	88	1	1	1	1
Outcome: Never goes out	*17-60*	*17-60*	*17-60*	*17-60*	*17-60*	*17-60*	*17.614*	*14.103*	*45*	*45*
(a) RD coefficient	0.282	0.277	0.274	0.264	0.274	0.264	0.233	0.297	0.297	0.297
(b) Clustered standard error	0.025	0.028	0.025	0.028	0.030	0.035	0.038	0.054		0.036
(c) Robust standard errors (HC1)	0.045	0.045	0.045	0.045	0.052	0.054			0.051	
t-stat (a)/(b)	11.28	9.89	10.96	9.43	9.13	7.54	6.13	5.50		8.25
t-stat (a)/(c)	6.27	6.16	6.09	5.87	5.27	4.89			5.82	
Total number of RD estimates	88	88	88	88	88	88	1	1	1	1

Table 2: Bandwith Sensitivity - Second Stage

	All possible bandwidth selections						Bonander et al. (2022)			Original study
RD methodology	CCT conventional		CCT bias corrected		CCT bias corrected and robust		CCT bias corrected and robust		conventional (OLS)	conventional (OLS)
Kernel	Triangular	Uniform	Triangular	Uniform	Triangular	Uniform	Triangular	Uniform	Uniform	Uniform
Outcome: Mental Distress										
Bandwidth	17-60	17-60	17-60	17-60	17-60	17-60	11.536	15.246	45	45
(a) RD coefficient	0.273	0.262	0.286	0.284	0.286	0.284	0.053	0.146	0.205	0.205
(b) Clustered standard error	0.077	0.085	0.077	0.085	0.092	0.107	0.074	0.142		0.094
(c) Robust standard errors (HC1)	0.113	0.109	0.113	0.109	0.13	0.132			0.119	
t-stat (a)/(b)	3.55	3.08	3.71	3.34	3.11	2.65	0.716	1.028		2.181
t-stat (a)/(c)	2.42	2.40	2.53	2.61	2.20	2.15			1.723	
Total number of RD estimates	88	88	88	88	88	88	1	1	1	1
Outcome: Somatic symptoms of distress index										
Bandwidth	17-60	17-60	17-60	17-60	17-60	17-60	17.687	16.009	45	45
(a) RD coefficient	0.266	0.227	0.281	0.243	0.281	0.243	0.199	0.375	0.0175	0.175
(b) Clustered standard error	0.069	0.078	0.069	0.078	0.082	0.097	0.097	0.144		0.085
(c) Robust standard errors (HC1)	0.114	0.111	0.114	0.111	0.13	0.133			0.012	
t-stat (a)/(b)	3.86	2.91	4.07	3.12	3.43	2.51	2.05	2.60		2.06
t-stat (a)/(c)	2.33	2.05	2.46	2.19	2.16	1.83			1.458	
Total number of RD estimates	88	88	88	88	88	88	1	1	1	1
Outcome: Nonsomatic symptoms of distress index										
Bandwidth	17-60	17-60	17-60	17-60	17-60	17-60	11.480	14.848	45	45
(a) RD coefficient	0.19	0.19	0.193	0.201	0.193	0.201	0.118	0.065	0.1580	0.1580
(b) Clustered standard error	0.078	0.082	0.078	0.082	0.093	0.104	0.074	0.143		0.092
(c) Robust standard errors (HC1)	0.106	0.103	0.106	0.103	0.122	0.124			0.114	
t-stat (a)/(b)	2.44	2.32	2.47	2.45	2.08	1.93	1.595	0.455		1.717
t-stat (a)/(c)	1.79	1.84	1.82	1.95	1.58	1.62			1.386	
Total number of RD estimates	88	88	88	88	88	88	1	1	1	1
Sum of yes answers in SRQ-20										
Bandwidth	17-60	17-60	17-60	17-60	17-60	17-60	11.15	16.151	45	45
(a) RD coefficient	1.038	0.955	1.115	1.07	1.115	1.07	-0.039	0.699	0.734	0.734
(b) Clustered standard error	0.341	0.383	0.341	0.383	0.404	0.473	0.352	0.636		0.44
(c) Robust standard errors (HC1)	0.542	0.518	0.542	0.518	0.623	0.627			0.563	
t-stat (a)/(b)	3.04	2.49	3.27	2.79	2.76	2.26	-0.111	1.099		1.668
t-stat (a)/(c)	1.92	1.84	2.06	2.07	1.79	1.71			1.30	
Total number of RD estimates	88	88	88	88	88	88	1	1	1	1

Table 3: Sensitivity to the Inclusion of Covariates

	All possible covariate combinations						Bonander et al. (2022) - No covariates	Original study - Full set of covariates
RD methodology	CCT conventional		CCT bias corrected		CCT bias corrected and robust		conventional (OLS)	conventional (OLS)
Kernel	Triangular	Uniform	Triangular	Uniform	Triangular	Uniform	Uniform	Uniform
Outcome: Days outside last week								
(a) RD coefficient	-1.041	-1.067	-1.079	-1.113	-1.079	-1.113	-0.963	-1.090
(b) Clustered standard error	0.156	0.154	0.156	0.154	0.179	0.187	0.19	0.180
t-stat (a)/(b)	-6.67	-6.93	-6.92	-7.23	-6.03	-5.95	-5.07	-6.06
Total number of RD estimates	62	61	62	61	62	61	1	1
Outcome: Under curfew								
(a) RD coefficient	0.660	0.699	0.637	0.671	0.637	0.671	0.708	0.696
(b) Clustered standard error	0.038	0.036	0.038	0.036	0.043	0.042	0.040	0.045
t-stat (a)/(b)	17.37	19.42	16.76	18.64	14.81	15.98	17.70	15.47
Total number of RD estimates	62	61	62	61	62	61	1	1
Outcome: Never goes out								
(a) RD coefficient	0.286	0.292	0.283	0.281	0.283	0.281	0.281	0.297
(b) Clustered standard error	0.029	0.031	0.029	0.031	0.034	0.040	0.037	0.036
t-stat (a)/(b)	9.86	9.42	9.76	9.06	8.32	7.03	7.59	8.25
Total number of RD estimates	62	62	62	62	62	62	1	1
Outcome: Mental Distress								
(a) RD coefficient	0.248	0.212	0.256	0.226	0.256	0.226	0.169	0.205
(b) Clustered standard error	0.098	0.098	0.098	0.098	0.114	0.121	0.135	0.094
t-stat (a)/(b)	2.53	2.16	2.61	2.31	2.25	1.87	1.25	2.18
Total number of RD estimates	62	60	62	60	62	60	1	1
Outcome: Somatic symptoms of distress index								
(a) RD coefficient	0.254	0.211	0.266	0.221	0.266	0.221	0.221	0.175
(b) Clustered standard error	0.083	0.087	0.083	0.087	0.097	0.106	0.113	0.085
t-stat (a)/(b)	3.06	2.43	3.20	2.54	2.74	2.08	1.96	2.06
Total number of RD estimates	62	62	62	62	62	62	1	1
Outcome: Nonsomatic symptoms of distress index								
(a) RD coefficient	0.168	0.155	0.166	0.157	0.166	0.157	0.10	0.158
(b) Clustered standard error	0.094	0.091	0.094	0.091	0.112	0.116	0.125	0.092
t-stat (a)/(b)	1.79	1.70	1.77	1.73	1.48	1.35	0.80	1.72
Total number of RD estimates	62	61	62	61	62	61	1	1
Sum of yes answers in SRQ-20								
(a) RD coefficient	0.943	0.842	0.984	0.904	0.984	0.904	0.560	0.734
(b) Clustered standard error	0.406	0.416	0.406	0.416	0.478	0.517	0.589	0.440
t-stat (a)/(b)	2.32	2.02	2.42	2.17	2.06	1.75	0.95	1.67
Total number of RD estimates	62	60	62	60	62	60	1	1