INSTITUTE for REPLICATION

No. 46
I4R DISCUSSION PAPER SERIES

Successful Replication of "The Long-Run Effects of Sports Club Vouchers for Primary School Children (2022)"

Felix Bacon

Abdel-Hamid Bello

Myriam Brown

Todd Morris

Laëtitia Renée

July 2023

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 46

Successful Replication of "The Long-Run Effects of Sports Club Vouchers for Primary School Children (2022)"

Felix Bacon[1], Abdel-Hamid Bello[1], Myriam Brown[1], Todd Morris[2], Laëtitia Renée[3]

[1]Université Laval, Québec/Canada
[2]HEC Montréal/Canada
[3]Université de Montréal/Canada

JULY 2023

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

Successful replication of "The Long-Run Effects of Sports Club Vouchers for Primary School Children (2022)"*

Felix Bacon† Abdel-Hamid Bello‡ Myriam Brown§ Todd Morris¶ Laëtitia Renée‖

July 7, 2023

Abstract

Marcus, Siedler and Ziebarth (2022 *American Economic Journal: Economic Policy*) examine the long-run health effects of a universal sports-club voucher program that was introduced in Saxony for primary school children in 2009. In 2018, the authors designed a survey that targeted the affected cohorts and nearby cohorts in Saxony and two neighboring states, and use a differences-in-differences identification strategy that exploits variation across states and cohorts in policy exposure. The authors document that treated individuals have knowledge of the program and recall receiving and redeeming the vouchers at higher rates, but find no effects on any health outcomes or behaviors. We successfully reproduce the main results of the paper exactly using data available in the paper's replication package and new Stata and R code. We also verify the robustness of the results using different outcomes, different control variables, different sample restrictions and different inference methods.

*This paper was completed as part of the 2023 Montreal Replication Games, organized by the Institute for Replication (`https://i4replication.org/`). For correspondence, contact Morris and Renée (`toddstuartmorris@gmail.com`; `laetitia.renee@umontreal.ca`).

†Université Laval

‡Université Laval

§Université Laval

¶HEC Montréal

‖Université de Montréal

1 Introduction

Marcus, Siedler and Ziebarth (2022) examine the long-run health effects of a universal sports-club voucher program that was introduced in Saxony for primary school children in 2009. In 2018, the authors designed a survey that targeted the affected cohorts and nearby cohorts in Saxony and two neighboring states, and use a differences-in-differences identification strategy that exploits variation across states and cohorts in policy exposure. The authors document that treated individuals have knowledge of the program and recall receiving and redeeming the vouchers at higher rates, but find no effects on any health outcome or behavior: membership of sports clubs; weekly hours of sport; and being overweight.

We assessed the reproducibility and robustness of these claims using the data provided by the authors in their replication package. Using the data provided by the authors and the sample restrictions described in the paper, we recoded the main analyses in Stata and R. We were able to reproduce the main estimates and standard errors exactly.

We assessed the robustness of results using different outcomes, different control variables, different sample restrictions and different inference methods. Overall, the main findings of the paper are robust: we find strong evidence of a positive effect on voucher knowledge, receipt and utilization, and little evidence of any long-run effects on physical activity or weight.

2 Reproducability

We started our analysis by attempting to reproduce the main results of MSZ 2022. We downloaded the AEA replication package for the paper. After some minor adjustments to file paths, we were able to reproduce the results in the paper using the authors' original Stata code. We then attempted to reproduce the main results using our own Stata and R code. We were successful with both statistical programs. As shown in Table 1, we were able to exactly match the authors' main estimates for all outcomes.[1]

We note that the data provided by the authors appears to have been cleaned before it was uploaded. Several variables seem to have been derived, such as treatment (which is defined

[1]Table 1 shows the estimates for the specification with state, cohort and municipality fixed effects, with standard errors clustered by municipality. We were also able to reproduce the results from the other specifications in MSZ 2022.

based on a respondent's state and year when they were in grade 3), an indicator for being overweight (defined as BMI>25), and a variable noting whether any outcome information is missing. Code was not provided for these derivations in the authors' replication package. To the best of our ability, we assessed whether these variables were correctly defined. This was possible for some outcomes (e.g., treatment), as we had access to information on a person's state and year when they were in grade 3, and we found no evidence that variables were incorrectly defined. However, we could not verify the derivations of other variables, such as the overweight dummy, as respondents' BMI was not provided in the replication package (nor was their height or weight).

We also discovered some minor inconsistencies between the code and the notes below certain figures and tables in the paper:

1. The notes below Figure 4 of MSZ 2022 incorrectly reports the treated cohorts as "third graders in school years 2009/10, 2010/11, and 2011/12" instead of 2008/09, 2009/10, and 2010/11. This appears to reflect a typo in the text rather than a coding error.

2. Similarly, the notes below Tables 4 and 5 of MSZ 2022 note that "all regressions include state and year fixed effects" when these regressions also included municipality fixed effects. The estimates are slightly different if municipality fixed effects are not included but the qualitative findings are similar. Again, this appears to reflect a typo in the text rather than a mistake in the code.

3 Replication

We test the robustness of the results to a direct replication by testing different outcomes, such as an obesity dummy (rather than an overweight dummy), different control variables, different sample restrictions and different inference methods. The original findings of the paper are broadly robust to these decisions.

3.1 Additional health outcomes

MSZ 2022 finds no effect on any health outcome: sports club membership, an overweight dummy and weekly hours of sport. However, it is possible that the vouchers affected hours of sport or Body Mass Index (BMI) at only certain parts of the distribution. Regarding hours of sport, we can construct indicator variables to assess whether hours are higher than a given threshold and

vary the threshold across the distribution. We cannot do the same for BMI, as the posted data does not contain respondents' BMI, but it does contain an overweight dummy (BMI>30).

Figure 1 shows that there is no evidence of any effect on weekly hours of sport at any part of the distribution. We also do not find any effect on the obesity dummy; the point estimate is close to zero and highly insignificant (Table 2). These results are consistent with the conclusions of MSZ 2022.

3.2 Different methods to account for sibling spillovers

MSZ 2022 discuss the possibility that spillovers between siblings could bias the estimates. They show in Table 5 of their paper that they obtain similar estimates if they omit individuals with a treated sibling or omit individuals with an older sibling. In Table 3, we show that the results are also robust to adding controls for having a sibling and its interaction with a Saxony dummy.

3.3 Different sample restrictions

We test the robustness of the results to the addition of the 2011/2012 cohort which was excluded from the analysis, and to the inclusion of all individuals regardless of the states they were living in during 3rd grade. Results are presented in Table 4. Other than those changes, we use the same three main specifications used in the original study (base DD, two-way FE, and two-way FE with municipality FE). The results are robust to the changes in the sample definition, with the exception of the effect on "Weekly hours of sport". Specifically, when we add the 2011/2012 cohort to the analysis, the effect of the voucher on hours of sport becomes significantly negative for the two first DD specifications. The effects on overweight and sports club membership remain insignificant. Overall, these results are consistent with MSZ 2022's conclusion that the voucher program did not have any positive long-term effects on physical activity or health.

3.4 Different inference methods

The authors show how the p-values of their estimates vary with different inference methods. Nonetheless, we consider further robustness checks in this domain. The main analysis in the paper clusters standard errors by municipality. While this captures the randomness in the sampling procedure used by the authors (which relied on municipalities to respond to an initial request), the variation in treatment occurs at the state level. In such difference-in-difference models it

is standard to cluster at the state level (Bertrand, Duflo and Mullainathan, 2004). While the authors show p-values with state-level clustering, recent papers highlight that hypothesis testing based on the $t(K-1)$ distribution, where K is the number of clusters, can be unreliable when the number of clusters is small (MacKinnon and Webb, 2018). Specifically, hypothesis tests will be over-rejected (MacKinnon and Webb, 2018). The wild cluster-bootstrap is one potential solution in this instance, but even this method may perform poorly when there is only one treated cluster (MacKinnon and Webb, 2018). In this instance, the authors recommend clustering by state and performing a subcluster bootstrap at the individual level. This can be implemented in Stata using the `boottest` command (Roodman et al., 2019) with the `bootcluster` option. In Table 5, we show how the p-values of the main estimates vary under different inference methods. In general, the p-values based on the wild subcluster bootstrap are similar to clustering by municipality, and none of the conclusions change regarding statistical significance.

Since the authors find no evidence of any health effects, they conduct a power analysis to assess what size effects would lie outside their confidence intervals (see Online Appendix Table B6 of MSZ 2022). Crucially, this analysis rests on the assumption that the inference method is correct. We revisit this power analysis using our alternative inference method (clustering by state and performing a subcluster bootstrap at the individual level). We present the upper limits of two-tailed 90% confidence intervals in Panel B of Table 5 for the three health outcomes and show how these compare to MSZ 2022.[2] In general, the estimated confidence intervals are wider, which suggests that the statistical power to detect significant health effects may be slightly smaller than reported in MSZ 2022.

The final set of robustness checks we consider are randomization inference methods. Specifically, we plot the distribution of regression coefficients and t-statistics when we vary which 3 of the 15 state-cohort cells are considered treated (with the other 12 considered controls). In reality, 3 consecutive cohorts in the same state are treated; there are 9 possible combinations if we restrict treatment assignment to have this structure (8 perturbations plus the correct treatment assignment). Using this approach, we see in Figure 2 that the estimated effects on voucher knowledge, receipt and utilization are extreme in the distribution of coefficients — larger than any of the 8 alternative assignments of treatment — while the estimated health effects are in

[2]These are referred to as 95% confidence intervals in Table B6 of MSZ 2022 (for a one-sided test). For overweightness, we present the lower limit of the confidence interval given the hypothesized negative effect of the voucher program, consistent with MSZ 2022.

the middle of the distribution.

A downside of this approach is that there are only 9 possible combinations of treatment assignment. We can obtain more combinations if we relax the assumption that the 3 treated cohorts have to be consecutive (we maintain the assumption that they are within the same state). Using this approach, there are 30 possible combinations of treatment assignment (29 perturbations plus the correct treatment assignment). Again, we see that the estimated effects on voucher knowledge, receipt and utilization are extreme in the distribution of coefficients (Figure 3) — larger than any of the alternative assignments of treatment — while the estimated health effects are in the middle of the distribution. This reinforces the conclusions of MSZ 2022: people can clearly remember the program and recall receiving and using the vouchers, but there appears to be no long-term effects on health outcomes or behaviors.

4 Conclusion

In this paper, we examine the reproducibility and robustness of Marcus, Siedler and Ziebarth's (2022) study of the long-term effects of a universal sports-club voucher program in Saxony that was introduced for primary school students in 2009. The authors document that treated individuals have knowledge of the program and recall receiving and redeeming the vouchers at higher rates, but find no effects on any health outcomes or behaviors.

Starting from the authors' AEA replication package, we were able to reproduce the main estimates with our own code using two different statistical programs (Stata and R). The data appears to have been cleaned before it was uploaded, but there was no do-file in the replication package showing such cleaning. We therefore checked that key variables were defined in a consistent manner (e.g., treatment status, which depends on birth cohort and state in third grade), and found no evidence of any mistakes.

We assessed the robustness of the results using different outcomes, different control variables, different sample restrictions and different inference methods. Overall, the main findings of the paper appear robust: we find strong evidence of a positive effect on voucher knowledge, receipt and utilization, and little evidence of any long-run effects on health or physical activity.

References

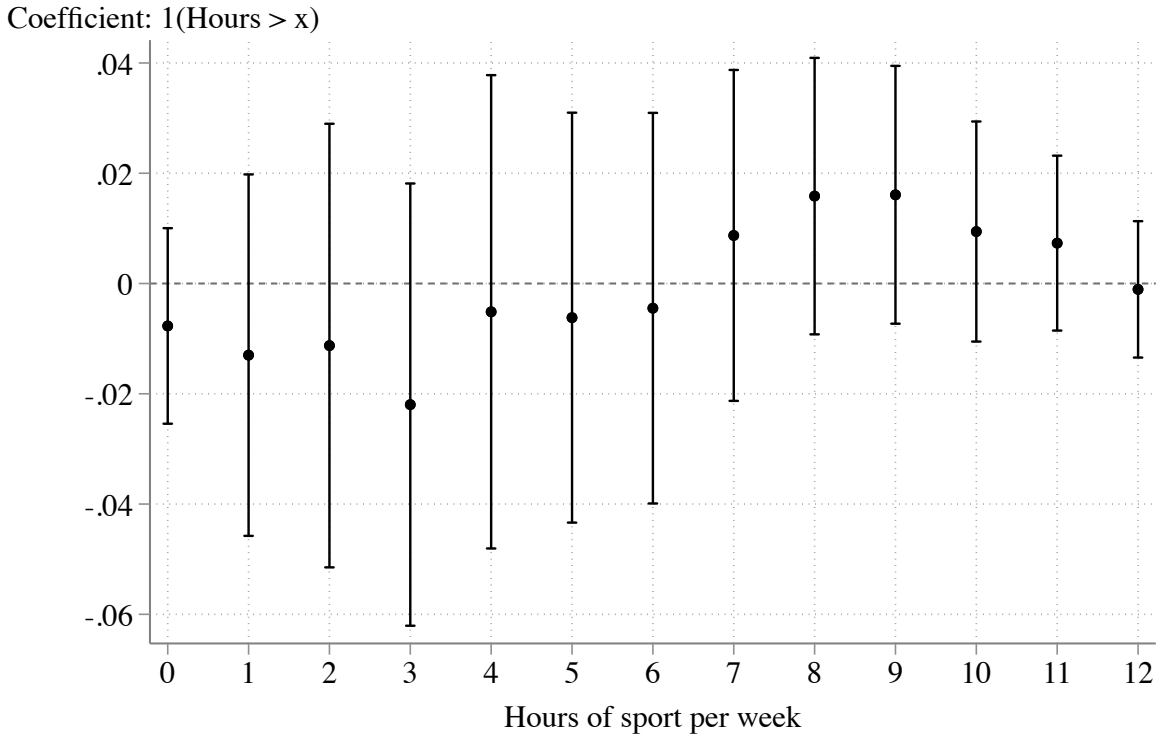
Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How much should we trust differences-in-differences estimates?" *Quarterly Journal of Economics*, 119(1): 249–275.

MacKinnon, James G, and Matthew D Webb. 2018. "The wild bootstrap for few (treated) clusters." *Econometrics Journal*, 21(2): 114–135.

Marcus, Jan, Thomas Siedler, and Nicolas R. Ziebarth. 2022. "The Long-Run Effects of Sports Club Vouchers for Primary School Children." *American Economic Journal: Economic Policy*, 14(3): 128–65.

Roodman, David, Morten Ørregaard Nielsen, James G MacKinnon, and Matthew D Webb. 2019. "Fast and wild: Bootstrap inference in Stata using boottest." *Stata Journal*, 19(1): 4–60.

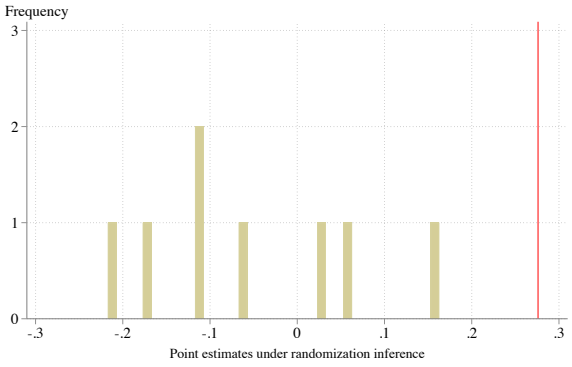
Figure 1: Effects on distribution of weekly hours of sport



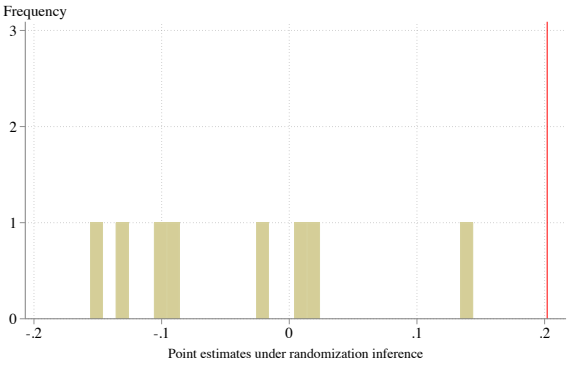
Notes: This figure shows point estimates and 95% confidence intervals for the effects on weekly hours of sport at different parts of the distribution. The dependent variable in each regression is an indicator variable for sport hours being higher than a given number. We vary the threshold from 0 to 12 hours and show the estimates. The estimates are based on the MSZ 2022's main specification with state, cohort and municipality fixed effects (with municipality-level clustering).

Figure 2: Distribution of coefficients under randomization inference (method with 8 permutations of treatment assignment)

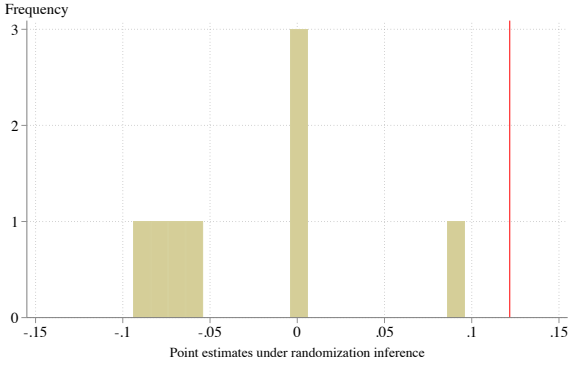
(a) Program known



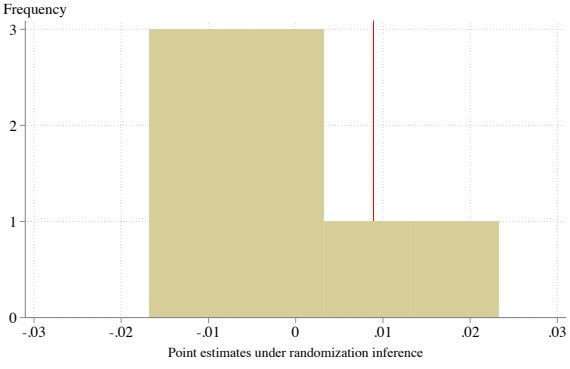
(b) Voucher received



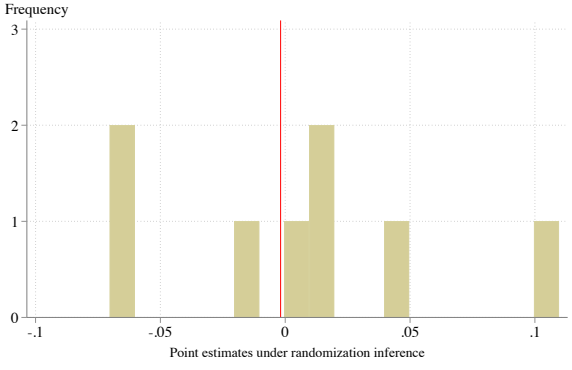
(c) Voucher redeemed



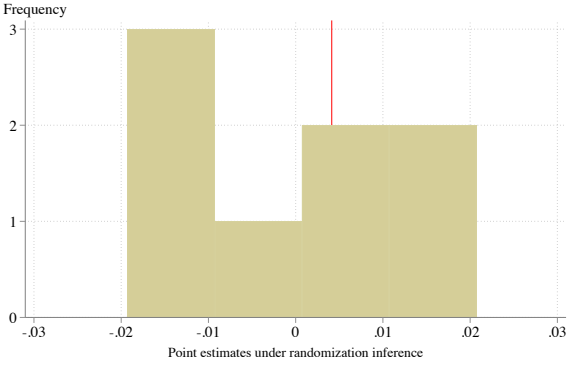
(d) Sports club member



(e) Weekly hours of sport



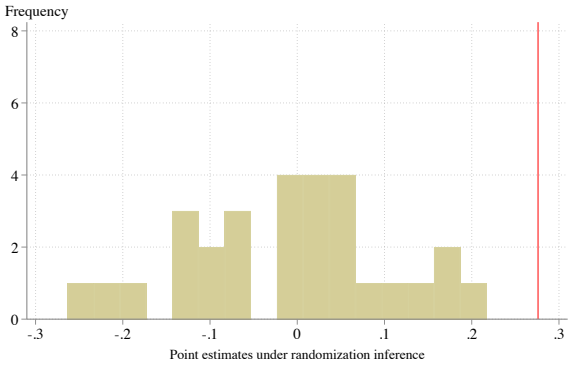
(f) Overweight



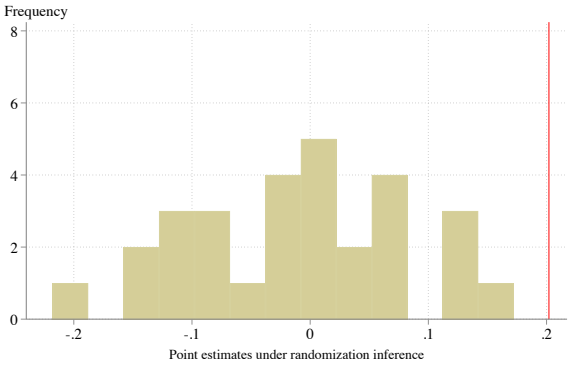
Notes: These figures show the distribution of placebo coefficients under a randomization inference procedure where we assign different state-cohort cells to treatment. This analysis maintains the assumption that three consecutive cohorts within the same state are treated, which gives 9 possible combinations of treatment assignment (8 permutations plus the correct assignment). The vertical bars show the distribution of the point estimates from the permutations and the vertical red line shows the point estimate under the correct assignment.

Figure 3: Distribution of coefficients under randomization inference (method with 29 permutations of treatment assignment)

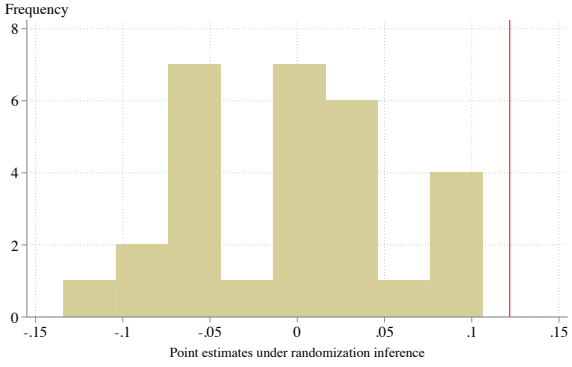
(a) Program known



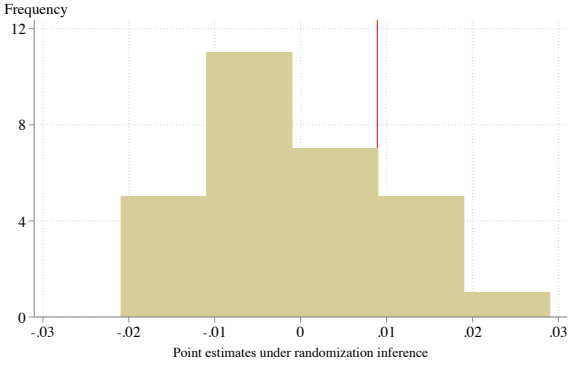
(b) Voucher received



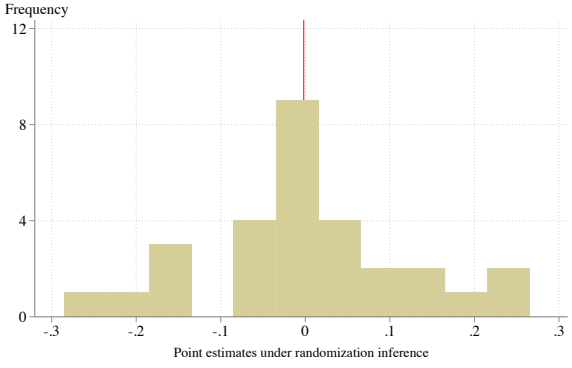
(c) Voucher redeemed



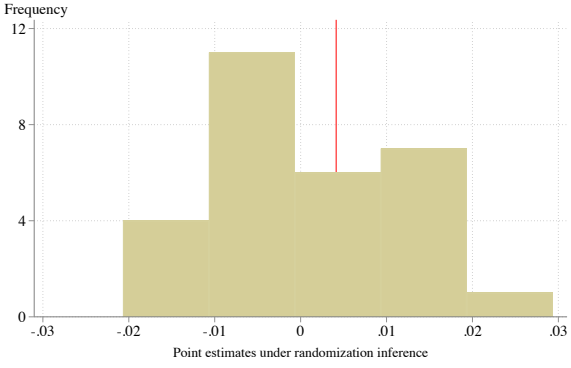
(d) Sports club member



(e) Weekly hours of sport



(f) Overweight



Notes: These figures show the distribution of placebo coefficients under a randomization inference procedure where we assign different state-cohort cells to treatment. This analysis maintains the assumption that three cohorts within the same state are treated (but relaxes the assumption that they are consecutive), which gives 30 possible combinations of treatment assignment (29 permutations plus the correct assignment). The vertical bars show the distribution of the point estimates from the permutations and the vertical red line shows the point estimate under the correct assignment.

Table 1: Reproduction of main estimates in MSZ 2022 using new Stata and R code

	(1) MSZ 2022	(2) Reproduction
Panel A: Awareness and take-up		
Program known	0.276	0.276
	(0.014)	(0.014)
p-value	0.000	0.000
Voucher received	0.202	0.202
	(0.011)	(0.011)
p-value	0.000	0.000
Voucher redeemed	0.122	0.122
	(0.006)	(0.006)
p-value	0.000	0.000
Panel B: Physical activity and overweight		
Member of sports club	0.009	0.009
	(0.019)	(0.019)
p-value	0.636	0.636
Weekly hours of sport	-0.002	-0.002
	(0.159)	(0.159)
p-value	0.991	0.991
Overweight	0.004	0.004
	(0.016)	(0.016)
p-value	0.795	0.795
Observations	13,334	13,334

Notes: This table shows the main estimates of Marcus, Siedler and Ziebarth (2022), which are contained in column 3 of Table 2 of their paper. These estimates include fixed effects for state, cohort and municipality. Standard errors in parentheses are clustered by municipality. Using new Stata and R code, we were able to reproduce these estimates exactly using both software packages. The reproduced estimates are shown in column (2).

Table 2: Estimates for obesity

	Obese
Long-term effect of voucher program	-0.0037
	(0.0087)
p-value	0.667
Observations	13,334

Notes: This table shows the estimated long-term effects of the voucher program on obesity. The estimates include fixed effects for state, cohort and municipality. Standard errors in parentheses are clustered by municipality.

Table 3: Estimates controlling for sibling participation

	(1) MSZ 2022	(2) Sibling controls
Panel A: Awareness and take-up		
Program known	0.276	0.279
	(0.014)	(0.015)
p-value	0.000	0.000
Voucher received	0.202	0.204
	(0.011)	(0.011)
p-value	0.000	0.000
Voucher redeemed	0.122	0.123
	(0.006)	(0.006)
p-value	0.000	0.000
Panel B: Physical activity and overweight		
Member of sports club	0.009	0.013
	(0.019)	(0.019)
p-value	0.636	0.510
Weekly hours of sport	-0.002	0.018
	(0.159)	(0.162)
p-value	0.991	0.913
Overweight	0.004	0.002
	(0.016)	(0.017)
p-value	0.795	0.899
Observations	13,334	13,334

Notes: This table shows the robustness of the estimates to the inclusion of controls for having a sibling and its interaction with living in a treated state. The estimates include fixed effects for state, cohort and municipality. Standard errors in parentheses are clustered by municipality.

Table 4: Estimates with different sample restrictions

Outcomes	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Program known	0.272	0.263	0.260	0.272	0.261	0.261	0.276	0.265	0.263
	(0.0143)	(0.0106)	(0.0103)	(0.0144)	(0.0107)	(0.0105)	(0.0145)	(0.0108)	(0.0107)
	0.00	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*
Voucher received	0.200	0.197	0.197	0.201	0.194	0.195	0.202	0.196	0.196
	(0.0107)	(0.0094)	(0.0091)	(0.0109)	(0.0101)	(0.0098)	(0.0111)	(0.0103)	(0.0100)
	0.00	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*
Voucher redeemed	0.122	0.119	0.119	0.122	0.118	0.118	0.122	0.118	0.118
	(0.0060)	(0.0055)	(0.0053)	(0.0060)	(0.0061)	(0.0057)	(0.0061)	(0.0063)	(0.0059)
	0.00	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*
Member of sports club	0.0040	-0.0162	-0.0237	0.0027	-0.0171	-0.0141	0.0089	-0.0137	-0.0112
	(0.0195)	(0.0150)	(0.0156)	(0.0194)	(0.0147)	(0.0142)	(0.0188)	(0.0146)	(0.0140)
	0.839	*0.283*	*0.133*	*0.889*	*0.249*	*0.324*	*0.636*	*0.349*	*0.425*
Weekly hours of sport	-0.068	-0.198	-0.256	-0.082	-0.196	-0.216	-0.002	-0.148	-0.185
	(0.161)	(0.119)	(0.115)	(0.159)	(0.117)	(0.117)	(0.159)	(0.119)	(0.118)
	0.671	*0.099*	*0.028*	*0.609*	*0.097*	*0.068*	*0.991*	*0.217*	*0.121*
Overweight	0.0050	0.0072	0.0098	0.0058	0.0068	0.0058	0.0041	0.0055	0.0040
	(0.0161)	(0.0105)	(0.0093)	(0.0161)	(0.0098)	(0.0088)	(0.0158)	(0.0099)	(0.0087)
	0.755	*0.493*	*0.298*	*0.721*	*0.492*	*0.513*	*0.795*	*0.581*	*0.652*
Observations	13,334	16,082	16,898	13,334	16,082	16,898	13,334	16,082	16,898
State FE	N	N	N	Y	Y	Y	N	N	N
Municipality FE	N	N	N	N	N	N	Y	Y	Y
Cohort FE	N	N	N	Y	Y	Y	Y	Y	Y
2011/2012 cohort	N	Y	Y	N	Y	Y	N	Y	Y
All states	N	N	Y	N	N	Y	N	N	Y
Shown in the paper	Y	N	N	Y	N	N	Y	Y	N

Notes: This table shows the robustness of the estimates to the addition of the 2011/2012 cohort which was excluded from the analysis, and to the inclusion of all individuals regardless of the states they were living in during 3rd grade. Standard errors in parentheses are clustered by municipality. P-values are shown in italics below the standard errors.

Table 5: Sensitivity of estimates to alternative inference methods

	(1) MSZ 2022	(2) Sub-cluster bootstrap
Panel A: p-values of main estimates		
Program known	0.000	0.000
Voucher received	0.000	0.000
Voucher redeemed	0.000	0.000
Member of sports club	0.636	0.469
Weekly hours of sport	0.991	0.991
Overweight	0.795	0.797
Panel B: Upper limits of 90% confidence interval		
Member of sports club	0.040	0.046
Weekly hours of sport	0.260	0.430
Overweight	-0.022^{LL}	-0.058^{LL}

Notes: This table examines the sensitivity of the estimates to alternative inference methods. Specifically, we compare the approach in Marcus, Siedler and Ziebarth (2022), which is to cluster by municipality, with a wild cluster-bootstrap approach that clusters by state. Specifically, we implement a subcluster bootstrap using the `boottest` command in Stata with subclustering at the individual level using the `bootcluster` option. Panel A shows that the p-values of the two approaches are similar, with no change to any of the conclusions regarding statistical significance. Panel B shows the upper limits of a two-sided 90% confidence interval based on the hypothesized sign of the effect. ^{LL} stands for lower limit (since the hypothesized sign of the effect on the overweight dummy is negative).