Supplementary Data

The end of court-ordered desegregation and U.S. children's health: Quasi-experimental evidence

Guangyi Wang, Justin S. White, Rita Hamad

Table of Contents

Appendix S1. Supplemental materials

Outcome measurement

Child outcomes in the National Health Interview Survey (NHIS) were reported by an adult knowledgeable and responsible for the health of the child. General health status was asked using a five-item Likert scale, which we dichotomized as poor/fair/good vs very good/excellent. Body weight was captured using a *Z*-score of age- and sex-adjusted body mass index (BMI). In NHIS, child BMI was only available for children aged 12-17 years with a reported current height and weight, as values for younger children were deemed unreliable. Mental health was measured using the six-item Strengths and Difficulties Questionnaire (SDQ), which includes one impact item and five symptom items to screen behavioral and emotional problems for children aged 4-17 years.[1, 2] The impact question asks whether the child had difficulties with emotions during the past 6 months. It ranges from 0-3, with a higher value representing more difficulties. The five symptoms include whether the child was well-behaved, had many worries, was unhappy, had good attention, and got along with adults during the past 6 months (range 0-2). Based on standard cutoffs in the literature, we defined a child as having high risk of mental health problems if the total SDQ score was ≥6 using the five symptom items and/or if they had serious overall difficulties on the impact item.[3, 4] The SDQ score was available in 2002, 2005-2007, and 2010-2018. Asthma was measured as whether the child had an asthma attack/episode during the past 12 months, conditional on the child ever being told by a doctor that they had asthma. Children who never had asthma were coded as zero.

DiD assumptions

The validity of DiD analyses relies on several assumptions. The first assumption is that pre-post differences in outcomes would have been similar between children in the treatment group (i.e., released districts) and children in the control group (i.e., districts that were not released) without the intervention (i.e., Supreme Court decisions beginning in 1991). While this assumption cannot be directly tested, several widely used methods exist to implicitly test it, including observing or empirically assessing differences in outcomes in the treatment and control groups in the pre-intervention period (i.e., the "parallel trends" assumption). If pre-intervention outcome trends in the treatment and control groups are similar, DiD analyses may be satisfied. In this study, we used event-study analyses, which allow for estimation and visual presentation of the releases' impacts in both pre- and post-release periods. The coefficients on the pre-release periods present the differences in pre-release outcome trends between children in released districts and children in districts that were not released. Null results for the coefficients would suggest the parallel trends assumption may be satisfied. As shown in Figures 2-3, there were no significant differences in outcomes between children in released districts and those in districts that were not released during the pre-release period, suggesting parallel pre-intervention trends.

Second, DiD assumes non-differential compositional changes among the treatment and the control group over time that might contribute to observed differences. In this case, previous evidence has revealed that court orders did not affect the racial composition of students within school districts.[5] To test this in our sample, we ran traditional and BJS DiD regressions with individual characteristics as outcomes to empirically test for any differential pre/post changes in individual characteristics among children in released districts and unreleased districts. As shown in Supplemental Table S1, only BJS results showed some compositional changes in children's sex and their parents' education level. Further, we adjusted for those characteristics in our regression models to reduce the change of confounding. Nevertheless, our results may suffer from bias issue if there are any remaining changes of unobserved confounders, which is a limitation of all quasi-experimental studies.

Novel DiD methods

Several novel estimators have been proposed recently. Each uses slightly different methods to create more valid control groups for treated units (i.e., not-yet-treated units and/or never-treated units), and the various methods have been shown to generate roughly similar estimates.[6-8] In this study, we used the approach developed by Borusyak, Jaravel, and Spiess (BJS).[6, 9] We opted to use the BJS imputation method for several reasons. First, the BJS method is computationally faster with a larger dataset. In comparison, other methods like the dCdH method may require significantly more computing time, and for this analysis of restricted NHIS data we were operating within the confines of the Federal Statistical Research Data Center. Second, the imputation method is more intuitive and easier to conceptualize in the estimation process compared to other methods that utilize more complex weighting methods. Third, the BJS estimator is often more efficient and precise than other estimators, under the assumption that parallel trends hold for all groups and all periods.

The BJS method involves three steps. First, state and time fixed effects were fitted by a regression using a valid control group (i.e., not-yet-treated and never-treated observations only). Second, fitted values from this regression were used to impute the untreated potential outcomes and obtain an estimated treatment effect for each treated observation. Finally, the DiD estimator aggregates the weighted average of individual treatment effects.

We implemented the procedure using the community-contributed *did_imputation* package in Stata 17 (College Station, TX).[10] Notably, approximately half of the school districts in our sample were never released by the end of our study period (never-treated). Having a good number of never-treated observations as control helps alleviate concerns about the control group's representativeness when estimating effects for districts that were released later.

Justification of including fixed effects at the state level

We do not include fixed effects at the district level because of the small number of observations in each cell. However, recent work has indicated that using more aggregate level fixed effects is sufficient if treatment at the unit level is randomized within the more aggregated level geography.[11] As mentioned earlier, the timing of a district's release was effectively quasi-random, supported by historical information and empirical evidence from previous studies. Considering this, incorporating state-level fixed effects may be sufficient for this study. Nevertheless, this is a potential limitation, and future studies with larger samples could investigate whether incorporating district-level fixed effects substantively alters the results.

Missingness

About 30% of values for family income were missing, while less than 5% were missing for other variables. We therefore carried out additional secondary analysis to account for this missingness. Namely, NHIS provides five sets of imputed family income data that we used in a multiple imputation analysis.[12] While typically higher numbers of imputed data sets are preferred when 30% of values are missing, we were unable to generate additional imputed data sets because of computational limitations at the Federal Statistical Research Data Center where restricted NHIS data must be analyzed, given the size of the data set. Furthermore, while we applied multiple imputation to the traditional DiD model using the five sets of imputed family income data provided by NHIS, we could not use multiple imputation for the BJS DiD estimator as the regressions failed to converge.

Table S2 shows that estimated coefficients from the multiple imputation method were similar to the main results.

Power calculation

In this study, we clustered standard errors by school districts (n=352). According to Abadie et al. (2023), the correct level of clustering is determined by the treatment assignment mechanism, and failure to correctly cluster standard errors can lead to misleadingly small standard errors and narrow confidence intervals.[13] With 352 clusters and all covariates (i.e., individual and district characteristics, state and time FEs), the detectable Cohen's effect size index is approximately 0.09 using α of 0.05, power (1−β) of 0.8. Values < 0.15 are considered "medium".[14] Therefore, our current model is powered to detect medium effect sizes.

Multiple hypothesis testing using the Benjamini-Hochberg approach

Given the number of hypothesis tests in the subgroup analyses and event studies, we implemented multiple hypothesis testing for mental health in subgroup BJS DiD analyses (the only significant subgroup result) and all four outcomes in the event studies. We used the Benjamini-Hochberg approach, which controls the expected proportion of false positives among the significant results.[15] The main Results are reported in Table S3-11.

Placebo test

We conducted a placebo test by using children's height as the outcome since children's height is more likely to be affected by family-level risk factors such as malnutrition in infancy rather than school segregation. Results show no significant effects on Black and White children's height. The magnitude of estimated coefficients was very small and represented negligible percentage change from the baseline mean (Table S12).

Supplemental references

1. Goodman R. The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*. 1999;40(5):791-799.
2. Goodman R. The Strengths and Difficulties Questionnaire: a research note. *Journal of child psychology and psychiatry*. 1997;38(5):581-586.
3. Pastor PN, Reuben CA, Duran CR. Identifying Emotional and Behavioral Problems in Children Aged 4-17 Years: United States, 2001-2007. National Health Statistics Reports. Number 48. *National Center for Health Statistics*. 2012;
4. Ringeisen H, Aldworth J, Colpe LJ, Pringle B, Simile C. Estimating the prevalence of any impairing childhood mental disorder in the national health interview survey. *International journal of methods in psychiatric research*. 2015;24(4):266-274.
5. Reardon SF, Grewal ET, Kalogrides D, Greenberg E. B rown Fades: The End of Court-Ordered School Desegregation and the Resegregation of A merican Public Schools. *Journal of Policy Analysis and Management*. 2012;31(4):876-904.
6. Borusyak K, Jaravel X, Spiess J. Revisiting event study designs: Robust and efficient estimation. *arXiv preprint arXiv:210812419*. 2021;
7. De Chaisemartin C, D'Haultfoeuille X. *Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey*. 2022.
8. De Chaisemartin C, D'Haultfoeuille X. *Difference-in-differences estimators of intertemporal treatment effects*. 2022.
9. Borusyak K, Jaravel X, Spiess J. Revisiting event study designs: Robust and efficient estimation. *arXiv preprint arXiv:210812419*. 2022;
10. Borusyak K. DID_IMPUTATION: Stata module to perform treatment effect estimation and pre-trend testing in event studies. 2022;
11. Papke LE, Wooldridge JM. A simple, robust test for choosing the level of fixed effects in linear panel data models. *Empirical Economics*. 2022:1-19.
12. National Center for Health Statistics. Multiple imputation of family income and personal earnings in the National Health Interview Survey: methods and examples. *Hyattsville, MD: Centers for Disease Control and Prevention*. 2018;
13. Abadie A, Athey S, Imbens GW, Wooldridge JM. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*. 2023;138(1):1-35.
14. Cohen J. *Statistical power analysis for the behavioral sciences*. Academic press; 2013.
15. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*. 1995;57(1):289-300.

Figure S1. Sample selection flow chart

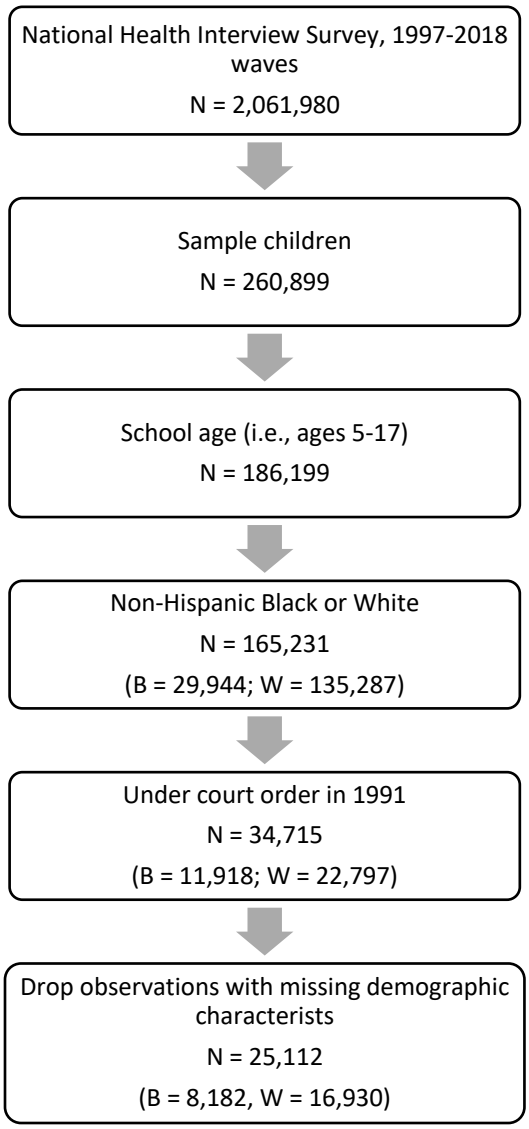
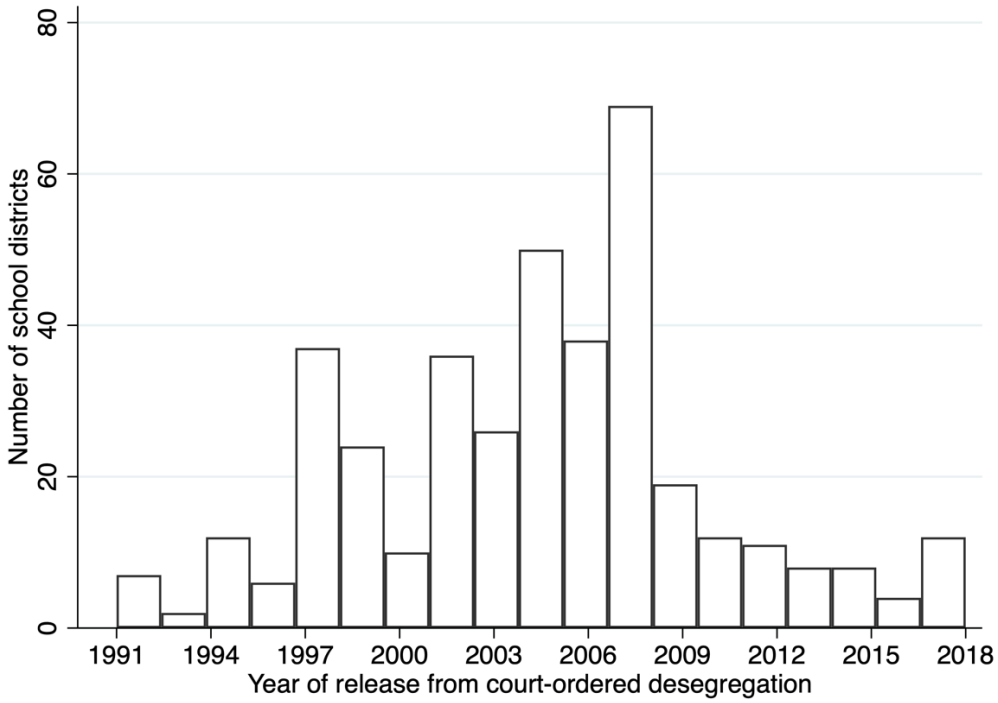


Figure S2. Number of school districts released from court-ordered desegregation, 1991-2018



Note: Data on school district court orders were compiled by Reardon et al. and ProPublica. Out of the 646 districts still under court-ordered desegregation in 1991, approximately 250 were not released by the end of our study period (i.e., 2018). Among the released districts, there was considerable variation in the time, with around 65 districts being released for more than 18 years (i.e., released before 2000).

Table S1. Test of compositional changes before and after the end of court-ordered desegregation

	Coefficient [95% CI]			
	Black children		White children	
Individual characteristics	Traditional DiD	BJS DiD	Traditional DiD	BJS DiD
Female	-0.011 [-0.039, 0.016]	-0.032* [-0.062, -0.001]	-0.007 [-0.024, 0.009]	-0.003 [-0.021, 0.015]
Age	-0.041 [-0.266, 0.183]	-0.092 [-0.335, 0.150]	-0.12 [-0.276, 0.037]	-0.148 [-0.334, 0.039]
At least one parent has high school degree	-0.01 [-0.032, 0.011]	-0.029* [-0.054, -0.004]	-0.016 [-0.051, 0.018]	-0.041** [-0.071, -0.010]
Parents married	-0.009 [-0.037, 0.020]	-0.013 [-0.047, 0.022]	0.003 [-0.020,0.026]	0.001 [-0.017, 0.019]
At least one parent employed	0.016 [-0.009, 0.042]	-0.006[-0.034,0.021]	0.003 [-0.014, 0.020]	-0.011 [-0.025, 0.003]
Family income ($US)				
<20,000	0.001 [-0.033, 0.036]	0.02 [-0.023, 0.062]	-0.001 [-0.029, 0.028]	0.033* [0.000, 0.065]
20,000-45,000	-0.007 [-0.038, 0.023]	-0.009 [-0.047, 0.030]	-0.007 [-0.030, 0.016]	-0.020 [-0.046, 0.007]
45,000-75,000	-0.007 [-0.033, 0.018]	-0.012 [-0.045, 0.020]	-0.012 [-0.034, 0.011]	-0.010 [-0.043, 0.024]
>75,000	0.013 [-0.017, 0.043]	0.001 [-0.028,0.030]	0.019 [-0.017, 0.055]	-0.004 [-0.037, 0.029]

Abbreviations: DiD, difference-in-differences.

Note: sample was drawn from the 1997-2018 waves of National Health Interview Survey and includes Black and White children who resided in districts that had been under the desegregation order in 1991. Regressions were run with individual characteristics as the dependent variables and controlled for fixed effects for year and state. Standard errors were clustered at the district level.

Table S2. Association of the end of court-ordered desegregation with school segregation and children' health, with and without imputation of family income

Outcome	Coefficient [95% CI]	
	Traditional DiD with complete case analysis	Traditional DiD with imputed family income
School segregation	0.015 [-0.006, 0.036]	0.015 [-0.006, 0.036]
Black children's health		
Very good/excellent health	-0.01 [-0.037, 0.017]	-0.011 [-0.038, 0.016]
Body mass index	-0.08 [-0.187, 0.028]	-0.069 [-0.177, 0.039]
Has mental health problems	-0.009 [-0.032, 0.014]	-0.008 [-0.031, 0.015]
Asthma episode	-0.008 [-0.027, 0.011]	-0.008 [-0.027, 0.011]
White children's health		
Very good/excellent health	0.020* [0.003, 0.037]	0.019* [0.002, 0.036]
Body mass index	-0.036 [-0.105, 0.033]	-0.034 [-0.103, 0.035]
Has mental health problems	-0.002 [-0.017, 0.014]	0.000 [-0.015 ,0.015]
Asthma episode	-0.003 [-0.012, 0.005]	-0.003 [-0.011, 0.006]

Abbreviations: DiD, difference-in-differences.

Note: sample was drawn from the 1997-2018 waves of National Health Interview Survey (NHIS) and includes Black and White children who resided in districts that had been under the desegregation order in 1991. School segregation was measured by the Black-White dissimilarity index within each district. Body mass index represent z-score adjusted by age and sex for children aged 12-17 years. Mental health problems represent whether the child's total Strengths and Difficulties Questionnaire (SDQ) score was ≥ 6. Asthma episode represents whether the child ever had an asthma attack/episode during the past 12 months. Estimates from the traditional DiD with imputed income models were derived by applying multiple imputation to the traditional DiD model using the five sets of imputed family income data provided by NHIS. All models were adjusted for 1991 school district characteristics, residential segregation, individual characteristics, and fixed effects for year and state. Standard errors are clustered at the district level.

Table S3. Multiple hypothesis testing for mental health results, by race, age, and sex

	Model p-value	Rank	FDR cutoff	BH significant
By race and age				
Black children aged 5-10 years	0.242	4	0.050	no
Black children aged 11-17 years	0.015	1	0.013	no
White children aged 5-10 years	0.120	3	0.038	no
White children aged 11-17 years	0.130	2	0.025	no
By race and sex				
Black female children	0.023	1	0.125	no
Black male children	0.455	4	0.500	no
White female children	0.272	2	0.250	no
White male children	0.281	3	0.375	no

Abbreviations: FDR, False discovery rates; BH: Benjamini-Hochberg
Note: Results represent the Benjamini-Hochberg approach to multiple hypothesis testing. "Model p-value" represents p-values from our regression models. If a model p-value is smaller than the FDR cutoff, the result passes multiple hypothesis testing and is considered statistically significant.

Table S4. Multiple hypothesis testing for event-study coefficients (very good/excellent health among Black children)

	Model p-value	Rank	FDR cutoff	BH significant
post 0	0.098	5	0.009	No
post 1	0.751	23	0.041	No
post 2	0.115	7	0.013	No
post 3	0.111	6	0.011	No
post 4	0.383	16	0.029	No
post 5	0.593	20	0.036	No
post 6	0.142	8	0.014	No
post 7	<0.001	1	0.002	Yes
post 8	0.717	22	0.039	No
post 9	0.564	18	0.032	No
post 10	0.943	26	0.046	No
post 11	0.178	9	0.016	No
post 12	0.384	17	0.030	No
post 13	0.208	11	0.020	No
post 14	0.317	14	0.025	No
post 15	0.7	21	0.038	No
post 16	223	28	0.050	No
post 17	0.194	10	0.018	No
post 18	<0.001	1	0.002	Yes
post 19	<0.001	1	0.002	Yes
pre 1	0.768	24	0.043	No
pre 2	0.016	4	0.007	No
pre 3	0.271	13	0.023	No
pre 4	0.38	15	0.027	No
pre 5	0.585	19	0.034	No
pre 6	0.962	27	0.048	No
pre 7	0.908	25	0.045	No
pre 8	0.254	12	0.021	No

Abbreviations: FDR, False discovery rates; BH: Benjamini-Hochberg

Note: Results represent the Benjamini-Hochberg approach to multiple hypothesis testing. "Model p-value" represents p-values from our regression models. If a model p-value is smaller than the FDR cutoff, the result passes multiple hypothesis testing and is considered statistically significant.

Table S5. Multiple hypothesis testing for event-study coefficients (BMI among Black children)

	Model p-value	Rank	FDR cutoff	BH significant
post 0	0.024	1	0.002	No
post 1	0.609	21	0.038	No
post 2	0.03	3	0.005	No
post 3	0.74	23	0.041	No
post 4	0.45	14	0.025	No
post 5	0.176	8	0.014	No
post 6	0.347	12	0.021	No
post 7	0.837	26	0.046	No
post 8	0.066	5	0.009	No
post 9	0.332	11	0.020	No
post 10	0.54	18	0.032	No
post 11	0.824	25	0.045	No
post 12	0.115	6	0.011	No
post 13	0.313	10	0.018	No
post 14	0.558	19	0.034	No
post 15	0.028	2	0.004	No
post 16	0.496	16	0.029	No
post 17	0.847	28	0.050	No
post 18	0.66	22	0.039	No
post 19	0.421	13	0.023	No
pre 1	0.229	9	0.016	No
pre 2	0.837	26	0.046	No
pre 3	0.491	15	0.027	No
pre 4	0.034	4	0.007	No
pre 5	0.783	24	0.043	No
pre 6	0.142	7	0.013	No
pre 7	0.517	17	0.030	No
pre 8	0.603	20	0.036	No

Abbreviations: FDR, False discovery rates; BH: Benjamini-Hochberg

Note: Results represent the Benjamini-Hochberg approach to multiple hypothesis testing. "Model p-value" represents p-values from our regression models. If a model p-value is smaller than the FDR cutoff, the result passes multiple hypothesis testing and is considered statistically significant.

Table S6. Multiple hypothesis testing for event-study coefficients (mental health among Black children)

	Model p-value	Rank	FDR cutoff	BH significant
post 0	<0.001	1	0.002	Yes
post 1	0.952	27	0.048	No
post 2	0.333	16	0.029	No
post 3	0.185	11	0.020	No
post 4	0.49	19	0.034	No
post 5	0.252	14	0.025	No
post 6	0.292	15	0.027	No
post 7	0.157	10	0.018	No
post 8	0.925	26	0.046	No
post 9	0.034	5	0.009	No
post 10	0.019	4	0.007	No
post 11	0.366	17	0.030	No
post 12	0.007	2	0.004	No
post 13	0.070	7	0.013	No
post 14	0.046	6	0.011	No
post 15	0.012	3	0.005	No
post 16	0.605	21	0.038	No
post 17	0.705	22	0.039	No
post 18	0.376	18	0.032	No
post 19	0.123	8	0.014	No
pre 1	0.525	20	0.036	No
pre 2	0.779	23	0.041	No
pre 3	0.131	9	0.016	No
pre 4	0.788	24	0.043	No
pre 5	0.223	12	0.021	No
pre 6	0.238	13	0.023	No
pre 7	0.994	28	0.050	No
pre 8	0.817	25	0.045	No

Abbreviations: FDR, False discovery rates; BH: Benjamini-Hochberg

Note: Results represent the Benjamini-Hochberg approach to multiple hypothesis testing. "Model p-value" represents p-values from our regression models. If a model p-value is smaller than the FDR cutoff, the result passes multiple hypothesis testing and is considered statistically significant.

Table S7. Multiple hypothesis testing for event-study coefficients (asthma episodes among Black children)

	Model p-value	Rank	FDR cutoff	BH significant
post 0	0.771	21	0.038	No
post 1	0.246	8	0.014	No
post 2	0.397	12	0.021	No
post 3	0.786	22	0.039	No
post 4	0.031	3	0.005	No
post 5	0.949	27	0.048	No
post 6	0.923	26	0.046	No
post 7	0.724	19	0.034	No
post 8	0.321	9	0.016	No
post 9	0.13	5	0.009	No
post 10	0.585	16	0.029	No
post 11	0.456	15	0.027	No
post 12	0.686	17	0.030	No
post 13	0.967	28	0.050	No
post 14	0.359	10	0.018	No
post 15	0.001	1	0.002	Yes
post 16	0.409	13	0.023	No
post 17	0.44	14	0.025	No
post 18	0.164	6	0.011	No
post 19	0.006	2	0.004	No
pre 1	0.077	4	0.007	No
pre 2	0.904	25	0.045	No
pre 3	0.737	20	0.036	No
pre 4	0.365	11	0.020	No
pre 5	0.197	7	0.013	No
pre 6	0.718	18	0.032	No
pre 7	0.802	23	0.041	No
pre 8	0.885	24	0.043	No

Abbreviations: FDR, False discovery rates; BH: Benjamini-Hochberg

Note: Results represent the Benjamini-Hochberg approach to multiple hypothesis testing. "Model p-value" represents p-values from our regression models. If a model p-value is smaller than the FDR cutoff, the result passes multiple hypothesis testing and is considered statistically significant.

Table S8. Multiple hypothesis testing for event-study coefficients (very good/excellent health among White children)

	Model p-value	Rank	FDR cutoff	BH significant
post 0	0.539	18	0.032	No
post 1	0.668	22	0.039	No
post 2	0.154	7	0.013	No
post 3	0.144	6	0.011	No
post 4	<0.001	1	0.002	Yes
post 5	0.044	2	0.004	No
post 6	0.964	28	0.050	No
post 7	0.914	27	0.048	No
post 8	0.363	11	0.020	No
post 9	0.247	8	0.014	No
post 10	0.120	5	0.009	No
post 11	0.309	9	0.016	No
post 12	0.524	17	0.030	No
post 13	0.754	25	0.045	No
post 14	0.586	20	0.036	No
post 15	0.364	12	0.021	No
post 16	0.621	21	0.038	No
post 17	0.108	4	0.007	No
post 18	0.373	13	0.023	No
post 19	0.457	14	0.025	No
pre 1	0.084	3	0.005	No
pre 2	0.542	19	0.034	No
pre 3	0.514	16	0.029	No
pre 4	0.489	15	0.027	No
pre 5	0.343	10	0.018	No
pre 6	0.752	24	0.043	No
pre 7	0.686	23	0.041	No
pre 8	0.787	26	0.046	No

Abbreviations: FDR, False discovery rates; BH: Benjamini-Hochberg

Note: Results represent the Benjamini-Hochberg approach to multiple hypothesis testing. "Model p-value" represents p-values from our regression models. If a model p-value is smaller than the FDR cutoff, the result passes multiple hypothesis testing and is considered statistically significant.

Table S9. Multiple hypothesis testing for event-study coefficients (BMI among White children)

	Model p-value	Rank	FDR cutoff	BH significant
post 0	0.019	2	0.0036	No
post 1	0.243	11	0.0196	No
post 2	0.463	17	0.0304	No
post 3	0.108	7	0.0125	No
post 4	0.084	5	0.0089	No
post 5	0.856	27	0.0482	No
post 6	0.333	14	0.0250	No
post 7	0.78	24	0.0429	No
post 8	0.797	25	0.0446	No
post 9	0.262	12	0.0214	No
post 10	0.601	18	0.0321	No
post 11	0.116	8	0.0143	No
post 12	0.831	26	0.0464	No
post 13	0.352	15	0.0268	No
post 14	0.776	23	0.0411	No
post 15	0.665	20	0.0357	No
post 16	0.234	10	0.0179	No
post 17	0.287	13	0.0232	No
post 18	0.107	6	0.0107	No
post 19	0.18	9	0.0161	No
pre 1	0.077	4	0.0071	No
pre 2	0.668	21	0.0375	No
pre 3	0.687	22	0.0393	No
pre 4	0.435	16	0.0286	No
pre 5	0.616	19	0.0339	No
pre 6	0.053	3	0.0054	No
pre 7	0.874	28	0.0500	No
pre 8	0.011	1	0.0018	No

Abbreviations: FDR, False discovery rates; BH: Benjamini-Hochberg

Note: Results represent the Benjamini-Hochberg approach to multiple hypothesis testing. "Model p-value" represents p-values from our regression models. If a model p-value is smaller than the FDR cutoff, the result passes multiple hypothesis testing and is considered statistically significant.

Table S10. Multiple hypothesis testing for event-study coefficients (mental health among White children)

	Model p-value	Rank	FDR cutoff	BH significant
post 0	0.209	14	0.025	No
post 1	0.034	8	0.014	No
post 2	0.019	6	0.011	No
post 3	0.032	7	0.013	No
post 4	0.018	5	0.009	No
post 5	0.281	17	0.030	No
post 6	0.62	21	0.038	No
post 7	0.471	18	0.032	No
post 8	0.059	10	0.018	No
post 9	0.009	3	0.005	No
post 10	0.968	26	0.046	No
post 11	0.063	11	0.020	No
post 12	0.817	22	0.039	No
post 13	0.161	12	0.021	No
post 14	0.013	4	0.007	No
post 15	<0.001	1	0.002	Yes
post 16	0.998	28	0.050	No
post 17	0.552	20	0.036	No
post 18	0.168	13	0.023	No
post 19	0.007	2	0.004	No
pre 1	0.058	9	0.016	No
pre 2	0.982	27	0.048	No
pre 3	0.248	15	0.027	No
pre 4	0.927	25	0.045	No
pre 5	0.265	16	0.029	No
pre 6	0.834	23	0.041	No
pre 7	0.502	19	0.034	No
pre 8	0.917	24	0.043	No

Abbreviations: FDR, False discovery rates; BH: Benjamini-Hochberg

Note: Results represent the Benjamini-Hochberg approach to multiple hypothesis testing. "Model p-value" represents p-values from our regression models. If a model p-value is smaller than the FDR cutoff, the result passes multiple hypothesis testing and is considered statistically significant.

Table S11. Multiple hypothesis testing for event-study coefficients (asthma episodes among White children)

	Model p-value	Rank	FDR cutoff	BH significant
post 0	0.905	28	0.050	No
post 1	0.832	21	0.038	No
post 2	0.505	15	0.027	No
post 3	0.903	26	0.046	No
post 4	0.322	14	0.025	No
post 5	0.904	27	0.048	No
post 6	0.09	6	0.011	No
post 7	0.001	1	0.002	Yes
post 8	0.218	11	0.020	No
post 9	0.213	10	0.018	No
post 10	0.797	20	0.036	No
post 11	0.863	23	0.041	No
post 12	0.722	18	0.032	No
post 13	0.244	12	0.021	No
post 14	0.003	2	0.004	Yes
post 15	0.895	24	0.043	No
post 16	0.016	5	0.009	No
post 17	0.732	19	0.034	No
post 18	0.858	22	0.039	No
post 19	0.17	8	0.014	No
pre 1	0.32	13	0.023	No
pre 2	0.007	3	0.005	No
pre 3	0.179	9	0.016	No
pre 4	0.167	7	0.013	No
pre 5	0.64	17	0.030	No
pre 6	0.505	15	0.027	No
pre 7	0.895	24	0.043	No
pre 8	0.007	4	0.007	No

Abbreviations: FDR, False discovery rates; BH: Benjamini-Hochberg

Note: Results represent the Benjamini-Hochberg approach to multiple hypothesis testing. "Model p-value" represents p-values from our regression models. If a model p-value is smaller than the FDR cutoff, the result passes multiple hypothesis testing and is considered statistically significant.

Table S12. Association of the end of court-ordered desegregation with school segregation and children's height

	Coefficient [95% CI]			
	Black children		White children	
Outcome	Traditional DiD	BJS DiD	Traditional DiD	BJS DiD
Height (in inches)	0.176 [-0.222, 0.574]	0.122 [-0.303, 0.548]	0.122 [-0.303, 0.548]	-0.159 [-0.480, 0.163]

Abbreviations: DiD, difference-in-differences; BJS: Borusyak, Jaravel, and Spiess.

Note: * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$. Sample was drawn from the 1997-2018 waves of National Health Interview Survey and includes Black and White children who resided in districts that had been under the desegregation order in 1991. School segregation was measured by the Black-White dissimilarity index within each district. Height was only collected for children aged 12-17. All models were adjusted for 1991 school district characteristics, residential segregation, individual characteristics, and fixed effects for year and state. Standard errors were clustered at the district level.