Supplemental Appendix to "Semiparametric Continuous Difference-in-Differences with Causal Forests and an Application to U.S. Charter Schools"

Feng Chen Douglas N. Harris Nicholas Lacoste

Contents

S.A Tests of Identifying Assumptions **2**
S.A.1 District-Specific Time Trends . . . 2

S.B AFGR Results **5**
S.B.1 Average Treatment Effect . . . 5
S.B.2 Distribution of Treatment Effects . . . 7
S.B.3 Effect Heterogeneity by District Traits . . . 9

S.C Additional Effect Heterogeneity **16**
S.C.1 Math vs. ELA Correlation . . . 16
S.C.2 Summary Statistics . . . 17
S.C.3 k-Means Matched Neighborhoods . . . 18
S.C.4 CACRT Event Studies . . . 19
S.C.5 Event Studies within Dose Groups . . . 22
S.C.6 Additional ELA Figures . . . 24
S.C.7 CGS-DiD (Linear Parametric) Heterogeneity . . . 26

S.D Proofs **27**
S.D.1 Proof of Proposition 1 . . . 27
S.D.2 Proof of Proposition 2 . . . 28
S.D.3 Proof of Proposition 3 . . . 29
S.D.4 Proof of Proposition 4 . . . 29
S.D.5 Derivation of Equation 14 . . . 30
S.D.6 Proof of Proposition S.1 . . . 31
S.D.7 Proof of Proposition S.2 . . . 32

S.E Technical Summary of GRF Estimation Procedure **32**

Table S.A1. Conditional Parallel Trends Pre-Test

	p-value, by random sample							
Outcome	(1)	(2)	(3)	(4)	(5)	Median	Reject at 5%	Treated
Graduation (AFGR)	0.33	0.67	0.28	0.50	0.56	0.50	0 of 5	171
Math	0.38	0.78	0.58	0.79	0.49	0.58	0 of 5	267
ELA	0.31	0.74	0.48	0.60	0.46	0.48	0 of 5	267

Notes: The table reports the integrated moments test of Callaway and Sant'Anna (2021) for the conditional parallel trends assumption holding in all pre-treatment time periods for all groups. For tractability, we conduct this exercise 5 times with 5 random subsamples of the data. In no draw for any outcome do we reject the null hypothesis that parallel trends holds conditional on covariates in our data.

S.A Tests of Identifying Assumptions

S.A.1 District-Specific Time Trends

One important possibility is the existence of district-specific time trends differentially affecting districts that adopt charter schools vs. districts that do not. The pre-trend placebo tests we estimate for the DiDCF suggest this may be present for graduation rates (and possibly ELA scores).[1] However, trend violations which correlate with binary treatment status, but not with overall dose, will not bias our estimates of heterogeneity. Specifically, they will not bias our "ranking" of districts most affected vs. least affected by charter entry nor will they bias our estimates of how the $CATT/CACRT$ varies with x.

To illustrate this, recall the DiDCF's partially linear specification of the $CATT$ from Equation 23:

$$\mathbb{E}[\Delta\widetilde{Y}_{i,g,t} \mid X_i = x] = \alpha_{g,t}(x) + \tau_{g,t}(x)\mathbb{E}[W_i \mid X_i = x] \tag{S.A1}$$

Going forward we drop the (g,t) subscripts to reduce clutter, noting that everything occurs within a (g,t) cell. Now, imagine we introduce an unobserved random variable within the set of treated districts, δ_i, into the data-generating process which creates differential trends that are uncorrelated with dose:

$$\Delta\widetilde{Y}_i = \alpha(X_i) + \tau(X_i)W_i + \delta_i + \varepsilon_i; \quad \text{Cov}(\delta_i, W_i \mid X_i) = 0; \quad \mathbb{E}[\varepsilon_i \mid W_i, X_i] = 0 \tag{S.A2}$$

[1] It is worth noting that the mechanics of the DiDCF estimator do not allow us to explicitly control for district-specific time trends (linear or otherwise) as is possible in the parametric model.

Proposition S.1. *Under Equation S.A2, the $CACRT$ bias equals:*

$$\text{plim } \hat{\tau}(x) - \tau(x) = \frac{\text{Cov}(\delta_i, W_i \mid X_i = x)}{\text{Var}(W_i \mid X_i = x)} = 0 \tag{S.A3}$$

The CATT bias equals:

$$\begin{aligned} \text{plim } \hat{\alpha}(x) - \alpha(x) &= \mathbb{E}[\delta_i \mid X_i = x] - \big(\text{plim}\, \hat{\tau}(x) - \tau(x)\big)\, \mathbb{E}[W_i \mid X_i = x] \\ &= \mathbb{E}[\delta_i \mid X_i = x] \end{aligned} \tag{S.A4}$$

For proof, see Section S.D.6.

Proposition S.1 states that under the existence of an unobservable covariate causing differential trends across treated districts relative to untreated districts, so long as the unobservable is otherwise uncorrelated with dose (and Conditional Strong Parallel Trends holds), the estimates of the $CACRT$ will be unaffected and all heterogeneity conclusions one may draw from them will remain correct. The unobservable *will*, however, bias the $CATT$ estimate by a scalar equal to the conditional mean of the unobservable.

Given that the $CATT$ will suffer bias under this type of contamination, under which conditions can we be confident that heterogeneity in the $CATT$ will remain valid?

Proposition S.2. *Define $\hat{\theta}_i = \theta_i + \eta_i$ as the estimated $CATT$ for district i, which is the sum of the true $CATT$ (θ_i) and the prediction error (η_i). The correlation between the true $CATT$ and the $CATT$ prediction can be stated as:*

$$\text{Corr}(\hat{\theta}, \theta) = \frac{1 + rs}{\sqrt{1 + 2rs + s^2}} \tag{S.A5}$$

Where $r = \text{Corr}(\theta, \eta)$ and $s = \frac{\text{sd}(\eta)}{\text{sd}(\theta)}$.

For proof, see Section S.D.7.

Proposition S.2 says that the $CATT$ estimates, despite being biased, preserve their ranking of units under two special cases. First, if $s = 0$ (i.e. the contamination is equal across treated units and is unrelated to X), then Corr $= 1$ and all $CATT$ heterogeneity results are valid. Second, if $r = 1$ (i.e. if the contamination is proportional to the $CATT$),

then Corr $= 1$ again for any s, because $(1+s)/\sqrt{(1+s)^2} = 1$. The case where $CATT$ rankings are not perfectly preserved is if both $s > 0$ and $r < 1$ (i.e. the contamination is not equal across units nor is it proportional to the true $CATT$). For example, in the case where the contamination is completely uncorrelated to the true $CATT$ (i.e. where $r = 0$), Corr $= 1/\sqrt{1+s^2}$ and the quality of the heterogeneity estimates degrades as the variance in η grows relative to the variance in θ. It also degrades rather slowly: e.g. if $s = 0.5$ (the standard deviation of the error is half as large as the standard deviation of the $CATT$), the correlation between true $CATT$s and estimated $CATT$s is still Corr ≈ 0.9.

In the case where the contamination takes on a district-specific linear time trend, then $\delta_i = \zeta_i(t - b_g)$, where ζ_i is the slope specific to district i. In this setting, $\text{Cov}(\delta_i, W_i \mid X_i) = (t-b_g)\cdot\text{Cov}(\zeta_i, W_i \mid X_i)$. In order for Proposition S.1 to hold, it must be that each district's linear trend is uncorrelated with its eventual dose in every period e: $\text{Cov}(\zeta_i, W_i \mid X_i = x) = 0$. We can check this by estimating each treated district's own pre-treatment trend, $\hat{\zeta}_i$ and regressing $\hat{\zeta}_i$ on $(1, W_i, X_i)$ to check if these slopes are conditionally correlated with the district's realized dose. We report these results in Table S.A2. We find no evidence that linear time trends in the pre-treatment years are correlated with dose, conditional on X_i. This provides support that under Conditional Strong Parallel Trends (Assumption 6), our estimates of $CACRT$ (i.e. $\tau(x)$) parameters remain unbiased even in the presence of linear trend divergence.

For the $CATT$ heterogeneity to remain valid under district-specific linear time trends, we require the linear trends to be either homogeneous across X_i neighborhoods or directly proportional to the true $CATT$. Because true $CATT$s are unknown, it is difficult to provide evidence for the latter condition. However, one can seek visual evidence for the former condition by revisiting the neighborhood event-study plots of Figure S.C2. Recall, this figure defined neighborhoods of districts based on covariate similarity using k-means clustering and estimated event-studies within each group. Evidence for homogeneous pre-trend divergence would appear as low variance in $CATT$ estimates during the pre-treatment periods. This is clearly evident in AFGR and ELA scores – the two outcomes where pre-trends seem to be diverging. We view this as strong evidence supporting the validity of the $CATT$ heterogeneity results for both potentially problematic outcomes.

Table S.A2. Correlation of District-Specific Linear Time Trends and Charter Intensity

Outcome	N	Dose coef.	Cl. SE	t	p	Partial r
afgr	5732	0.0128	0.0076	1.69	0.091	0.0699
math	7547	-0.0249	0.0427	-0.58	0.560	-0.0165
ela	7547	-0.0525	0.0380	-1.38	0.167	-0.0363

Notes: This table reports the results of a regression of district-specific linear trends on that district's charter intensity, conditional on X.

S.B AFGR Results

S.B.1 Average Treatment Effect

In this appendix section, we study the main results concerning the average freshman graduation rate (AFGR) outcome. We begin with the ATT as estimated by the CGS-DiD approach. Table S.B1 displays these results. Charter schools have a small, negative ATT of -0.001 when considering the top 10 VIF moderators (VIF scores come from the DiDCF), indicating that AFGR was about 0.1pp lower at the realized charter share.[2] The DiDCF estimates an $ATT = 0.004$ as seen in Figure S.B1. However, the $ACRT = 0.041$ indicates that at the average realized charter dose, the marginal effect of additional charters was positive (an additional 10pp of charter intensity would have increased AFGR by an additional 0.4pp). This indicates that districts which adopted charter schools tended to adopt too few for positive impacts on AFGR to be realized.

[2]We have a smaller number of treated units for AFGR than for test scores. Because the 2-stage estimator requires us to separately estimate CGS-DiD regressions per cohort-time cell, there are often too few treated units in an estimation cell to obtain reliable moderator coefficients when using all moderators. To gain back statistical power, we use only the top 10 VIF moderators.

Table S.B1. Average Treatment Effects of Charter-School Expansion on Graduation: CGS-DiD with Ten Moderators

	Graduation	
	(1)	(2)
ATT	0.006^{**}	-0.001
	(0.003)	(0.006)
ACRT	0.100^{***}	0.041
	(0.037)	(0.052)
Treated observations	5,732	5,732
Moderators	None	Top 10

Notes: This table reports the ATT and ACRT of charter-school enrollment share using a two-stage staggered CGS-DiD estimator for the AFGR outcome on the top 10 VIF moderators according to the DiDCF for AFGR. The first stage uses never-treated districts to estimate counterfactual outcome changes, while the second stage uses treated observations to estimate treatment effects. Column (1) excludes moderators, while Column (2) includes the top 10 moderators according to VIF (from the DiDCF in $e = 0$). The observation counts refer to treated observations. Standard errors are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table S.B2. Heterogeneity in the Effects of Charter-School Expansion on Graduation: CGS-DiD with Ten Moderators

	Graduation
Moderator	(1)
Baseline Treatment Effects	
Positive charter exposure	-0.004 (0.007)
Student Demographics	
White student share	-0.071 (0.047)
Black student share	-0.011 (0.049)
Hispanic student share	-0.006 (0.049)
Free or reduced-price lunch share	-0.086^{***} (0.027)
Special education share	0.194^{***} (0.071)
Market Setting	
Baseline outcome	0.006 (0.033)
Log enrollment	-0.006^{*} (0.004)

Continued on next page

Table S.B2 continued

Moderator	Graduation (1)
School Revenue and Spending	
Per-pupil spending ($1,000s)	0.004 (0.004)
Teacher salary ($1,000s)	−0.000 (0.000)
Student–teacher ratio	−0.001 (0.002)
Treated observations	5,732

Notes: The table reports heterogeneity estimates for the graduation rate from the staggered, continuous-dose CGS-DiD (Equation 15), estimated separately in each cohort g by calendar-year t cell. In the first stage, the outcome change $Y_t - Y_{g-1}$ of never-treated districts is regressed on the full set of moderators X to estimate the counterfactual trend. In the second stage, the residualized change of cohort-g districts is regressed on $(1, X_2, W, WX_2)$, where W is the charter enrollment share and X_2 contains the ten moderators with the highest causal-forest variable importance for ELA test scores at $e = 0$ (so they are not selected on graduation data) and their missing-value indicators; hence $CATT(w, X_2) = \alpha_0 + X_2'\alpha + w(\tau_0 + X_2'\tau)$. Reported coefficients average the cell estimates over post-treatment cells, weighted by each cell's number of treated observations. Moderators are centered at their weighted means: the first row reports α_0, the effect of positive charter exposure ($D = 1\{W > 0\}$) at the mean moderator vector, and each moderator row reports α_k, the coefficient on $D \times X_k$; the dose coefficients τ_0 and τ_k are estimated but not reported. Moderators are measured in the first observed year for each district, except the baseline outcome, Y_{g-2}, which is centered within each cell; missing values are set to the sample mean and flagged by missing-value indicators, which are included in the regressions but not reported. Dollar amounts are in thousands. Both stages are weighted by the graduation-rate denominator. Standard errors (in parentheses) are clustered by district and account for the first-stage estimation. $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

S.B.2 Distribution of Treatment Effects

The distribution of $CATT$ estimates per geographic district $\times$ event-time for AFGR can be found in Figure S.B1 for the DiDCF. Our estimates suggest there exists notable heterogeneity in $CATT$s, with the 5th and 95th percentiles of DiDCF effects ranging from approximately [-0.06, 0.07]. This indicates there exist districts which experienced as much as a 6pp *decrease* in AFGR at their observed charter intensity, while there are simultaneously districts which experienced as much as a 7pp *increase* in AFGR at their own observed charter share.

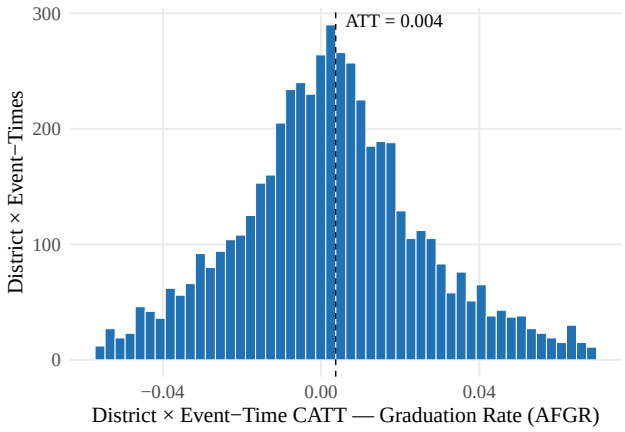


Figure S.B1. DiDCF CATT Distribution: Graduation Rate (AFGR)

Notes: Distribution of district × event-time CATT estimates from the DiDCF applied to graduation rates (AFGR). The vertical dashed line marks the ATT after fully aggregating across doses, event-times, and covariates. We drop the top and bottom 5% of CATT estimates from this figure.

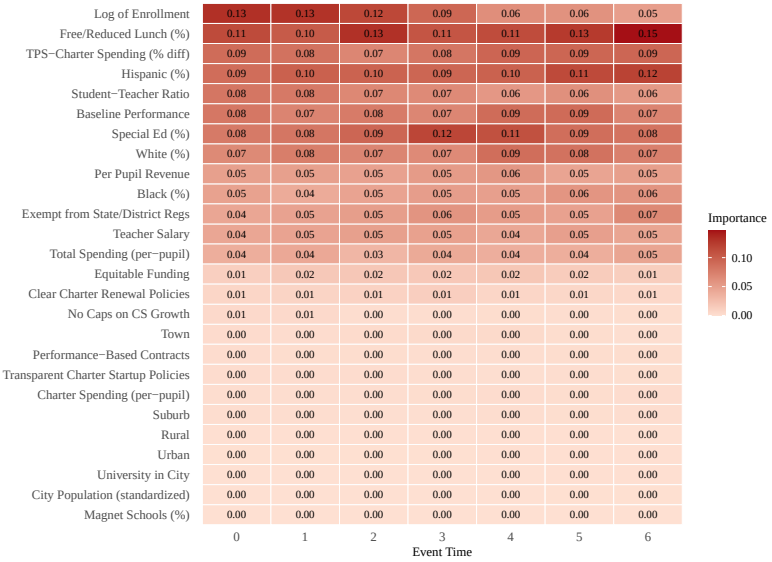


Figure S.B2. DiDCF Variable Importance Factors: Graduation Rate (AFGR)

Notes: Heatmap of variable importance factors (VIF) across all covariates and post-treatment event-times ($e = 0, \ldots, 6$) from the DiDCF applied to graduation rates. Cell values show the (depth-weighted) share of splits on each covariate at each event-time. Grey cells indicate covariates absent from the data.

Table S.B3. DiDCF Differences-in-Means: Graduation Rate (AFGR)

Variable	Mean (pos. CATT)	Mean (neg. CATT)	Difference	p-value
Log of Enrollment	8.69	8.75	-0.05	0.57
Free/Reduced Lunch (%)	0.24	0.22	0.02	0.18
TPS-Charter Spending (% diff)	0.23	0.19	0.05	0.13
Hispanic (%)	0.13	0.10	0.03	0.01
Student-Teacher Ratio	18.01	17.85	0.16	0.44
Baseline Performance	0.75	0.76	-0.01	0.56
Special Ed (%)	0.10	0.11	-0.01	0.00
White (%)	0.70	0.73	-0.03	0.13
Per Pupil Revenue	6294.47	6311.38	-16.91	0.90
Black (%)	0.12	0.12	0.00	0.86

Notes: For each of the ten moderators with the highest variable importance factor, the table reports the mean value among districts with positive CATT estimates (Mean^{+}), the mean among districts with negative CATT estimates (Mean^{-}), their difference, and the associated p-value of the difference. Rows sorted by variable importance factor.

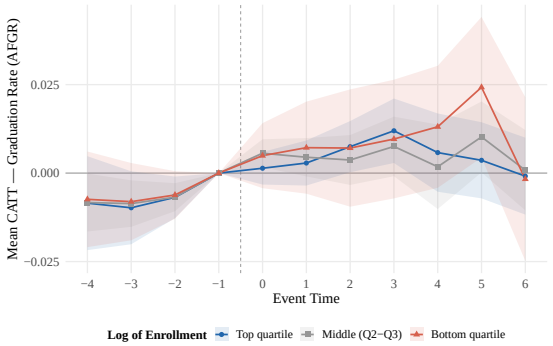
S.B.3 Effect Heterogeneity by District Traits

S.B.3.1 Trait Set #1: Demographics

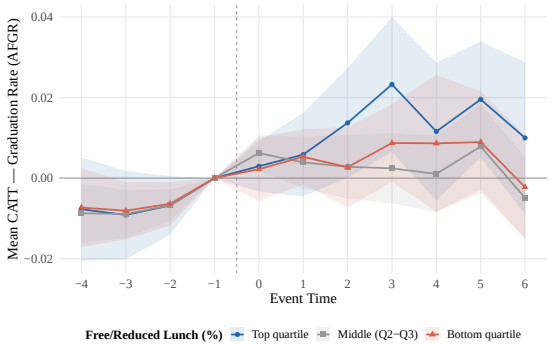
Like with test scores, we find that demographics are important explainers of charter school effects when considering graduation rates. Each of the six demographic variables score in the top 10 of VIF: Log Enrollment (1), Free-Reduced Lunch (FRL) % (2), Hispanic % (4), Special Ed % (7), White % (8), and Black % (10). Notably, while Log Enrollment is the most important variable for *short run* impacts according to DiDCF, it diminishes quickly in importance when considering long run effects in favor of FRL % and Hispanic %.

The demographic measures generally do not show evidence of linear relationships with the $CATT$ when considering AFGR. Districts with more FRL students seem to be positively impacted more often than not. Specifically, the event study suggests that districts in the top quartile of FRL share are impacted more. The k-means table reinforces this by showing that the two groups most positively impacted by charter schools are also the two groups with the largest FRL share. However, despite these signs that more low income students is associated with larger $CATT$ estimates, the effects seem to be non-linear beyond the top quartile, and the diff-in-means and BLP tables do not find notable linear relationships between FRL share and the $CATT$.

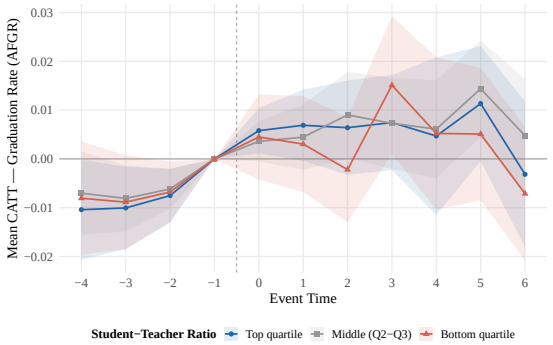
The ethnic variables (Black %, Hispanic %, and White %) appear to be highly non-linear in their relationships with the $CATT$. The diff-in-means and BLP for Hispanic %



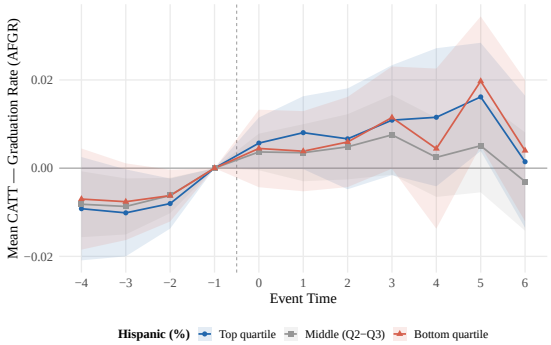
(a) Log Enrollment



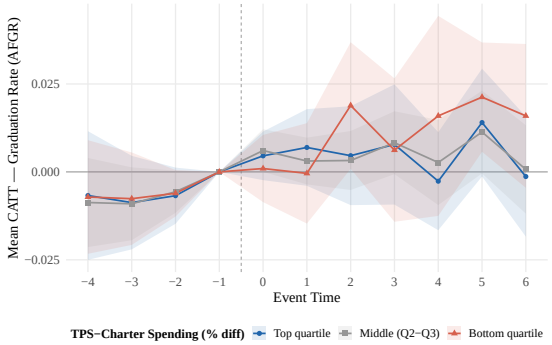
(b) Percent Free/Reduced Lunch



(c) Student-Teacher Ratio



(d) Percent Hispanic



(e) % Difference TPS vs. Charter Spending

Figure S.B3. DiDCF Quartile Event Studies (CATT): Graduation Rate (AFGR)

Notes: Each panel shows $\widehat{GATT}^{es}(e, \mathcal{A})$ estimates (Equation 25). These are group average treatment effects in each event-time, where the groups are districts in the top quartile (blue), middle two quartiles (grey), and bottom quartile (red) of the named moderator. Panels correspond to the five highest-ranked covariates by VIF at $e = 0$ for graduation rates.

Table S.B4. DiDCF Best Linear Predictor: Graduation Rate (AFGR)

	(1) No dose control			(2) Linear dose			(3) Dose-quintile FE		
	Est.	SE	p	Est.	SE	p	Est.	SE	p
(Intercept)	-0.03	0.02	0.21	-0.05	0.02	0.06	-0.04	0.02	0.13
Baseline Performance	-0.01	0.01	0.13	-0.01	0.01	0.25	-0.01	0.01	0.22
Black (%)	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.03
Hispanic (%)	0.05	0.01	0.00	0.05	0.01	0.00	0.05	0.01	0.00
Free/Reduced Lunch (%)	0.01	0.01	0.12	0.01	0.01	0.18	0.01	0.01	0.17
White (%)	0.03	0.01	0.00	0.03	0.01	0.00	0.03	0.01	0.00
Special Ed (%)	-0.03	0.02	0.08	-0.03	0.02	0.16	-0.03	0.02	0.14
Log of Enrollment	0.00	0.00	0.63	0.00	0.00	0.72	0.00	0.00	0.91
Per Pupil Revenue ($1,000s)	0.00	0.00	0.59	0.00	0.00	0.53	0.00	0.00	0.52
Student-Teacher Ratio	0.00	0.00	0.81	0.00	0.00	0.98	0.00	0.00	0.79
TPS-Charter Spending (% diff)	0.01	0.01	0.38	0.01	0.01	0.30	0.01	0.01	0.35
TPS-Charter Spending (% diff) (missing)	-0.01	0.00	0.21	-0.01	0.00	0.26	-0.01	0.00	0.21
Mean charter share				0.02	0.01	0.01			

Notes: Best linear predictor (BLP) regression of each district's mean post-treatment $\widehat{CATT}$ estimates on district moderators, applied to graduation rates. Each row reports the OLS coefficient of each covariate from regressing the mean post-treatment $\widehat{CATT}$ on the full covariate set. The "missing" indicator on the second instance of Charter Spending refers to a binary indicator for if the district was missing data for Charter Spending. Group (1) reports estimates without dose controls, Group (2) controls for dose linearly, Group (3) includes dose-quintile fixed effects.

suggest that more Hispanic students is associated with more positive effects, but the event study and k-means profiles suggest the relationship is much more interactive than the linear measures indicate. Specifically, the k-means table suggests that while the 2nd most positively impacted group of districts had a 78% Hispanic share, several other Hispanic-dominated district groupings fared much worse; the event study similarly shows that both the top *and* bottom quartiles of Hispanic % are impacted more positively than the middle 50%. The depth-3 policy tree sheds some light on Hispanic %'s interactivity. While it does not split on Hispanic % directly, it is clear that in order for a district to be both low Black % and low White %, the district is likely high Hispanic % in most cases. For these high Hispanic districts, charter school expansion is likely to be beneficial if the district is sufficiently low performing at baseline.

Black % seems particularly interactive when considering AFGR. The diff-in-means does not find a notable difference between positive vs. negative $CATT$ districts, and the k-means profiles suggest that while the best performing district group had a very high Black share, other districts with a high Black share did not fare as strongly. The depth-3 policy tree sheds some light on the interactivity for Black share. Specifically, high Black

Table S.B5. DiDCF Differences-in-Means (k-means matched groups): AFGR

	Covariate Grouping								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Log of Enrollment	8.92	9.40	10.07	8.80	10.57	9.51	7.57	7.05	8.14
Free/Reduced Lunch (%)	0.46	0.58	0.18	0.13	0.42	0.27	0.14	0.38	0.11
TPS-Charter Spending (% diff)	0.27	0.06	0.12	0.40	0.32	0.19	0.19	0.05	0.14
Hispanic (%)	0.05	0.78	0.10	0.04	0.13	0.31	0.02	0.13	0.03
Student-Teacher Ratio	16.08	18.50	17.18	20.11	17.36	22.18	15.71	15.84	18.44
Baseline Performance	0.59	0.63	0.70	0.82	0.52	0.75	0.87	0.74	0.80
Special Ed (%)	0.10	0.10	0.13	0.11	0.09	0.10	0.12	0.15	0.04
White (%)	0.40	0.15	0.69	0.90	0.22	0.49	0.93	0.72	0.91
Per Pupil Revenue	5969.98	5700.17	5733.96	6672.18	7864.86	5222.69	6596.92	8085.67	5422.08
Black (%)	0.46	0.04	0.18	0.03	0.62	0.08	0.02	0.04	0.04
Exempt from State/District Regs	0.72	0.71	0.64	0.49	0.54	0.76	0.50	0.70	0.71
Teacher Salary	54171.63	61972.37	56402.06	73471.05	77147.06	69418.35	55852.93	61532.26	56791.41
Total Spending (per-pupil)	5938.67	5945.22	5754.55	6730.30	7853.69	5361.01	6555.30	7671.86	5419.73
Equitable Funding	0.45	0.75	0.43	0.51	0.50	0.86	0.18	0.58	0.46
Clear Charter Renewal Policies	0.71	0.65	0.64	0.57	0.62	0.73	0.44	0.63	0.75
No Caps on CS Growth	0.78	0.74	0.89	0.80	0.75	0.78	0.80	0.79	0.69
Town	0.19	0.25	0.02	0.23	0.00	0.09	0.33	0.25	0.33
Performance-Based Contracts	0.59	0.58	0.62	0.56	0.56	0.32	0.49	0.51	0.50
Transparent Charter Startup Policies	0.57	0.51	0.57	0.59	0.51	0.51	0.39	0.56	0.43
Charter Spending (per-pupil)	10857.94	9553.39	9526.70	9098.00	12508.40	10032.24	11629.57	13351.44	9586.67
Suburb	0.16	0.22	0.40	0.46	0.16	0.51	0.14	0.14	0.35
Rural	0.29	0.03	0.13	0.09	0.00	0.07	0.45	0.57	0.22
Urban	0.36	0.50	0.44	0.22	0.84	0.33	0.07	0.04	0.10
University in City	0.01	-0.03	-0.03	0.02	-0.11	-0.01	0.00	0.02	0.00
City Population (standardized)	-0.00	-0.05	-0.06	0.00	0.15	0.02	0.00	0.04	0.00
Magnet Schools (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ATT	0.01	0.01	-0.00	-0.00	0.00	0.00	-0.00	-0.01	-0.01
ACRT	-0.04	-0.05	0.04	0.01	-0.01	-0.02	-0.10	-0.02	-0.03

Notes: Each panel shows $\widehat{GATT}^{es}(e, \mathcal{A})$ estimates (Equation 25). These are group average treatment effects in each event-time, where the groups are districts in each of the 10 k-means clusters. Panels correspond to the five highest-ranked covariates by VIF at $e = 0$ for graduation rates.

share districts should expand charter school presence if they are simultaneously high FRL (i.e. low income) or high Student-Teacher Ratio (i.e. low spending). Conversely, low Black share districts can also expand charter schools if they are simultaneously large and predominantly White %.

District size (in terms of Log Enrollment) seems to also be related to the $CATT$ in an interactive (and important) manner. The various tables and figures suggest different directions. The diff-in-means weakly suggests that smaller districts are more often better off. The BLP finds no relationship. The k-means table seems to suggest that bigger districts are more often positively impacted. The event studies show some evidence that smaller districts are impacted more. These conflicting results are only reconcilable by a highly interactive relationship, which both policy trees corroborate. Enrollment appears to interact with both spending variables (depth-2 tree) and ethnic demographics (depth-3 tree). In particular, larger districts are more often better off but only when the TPS-Charter

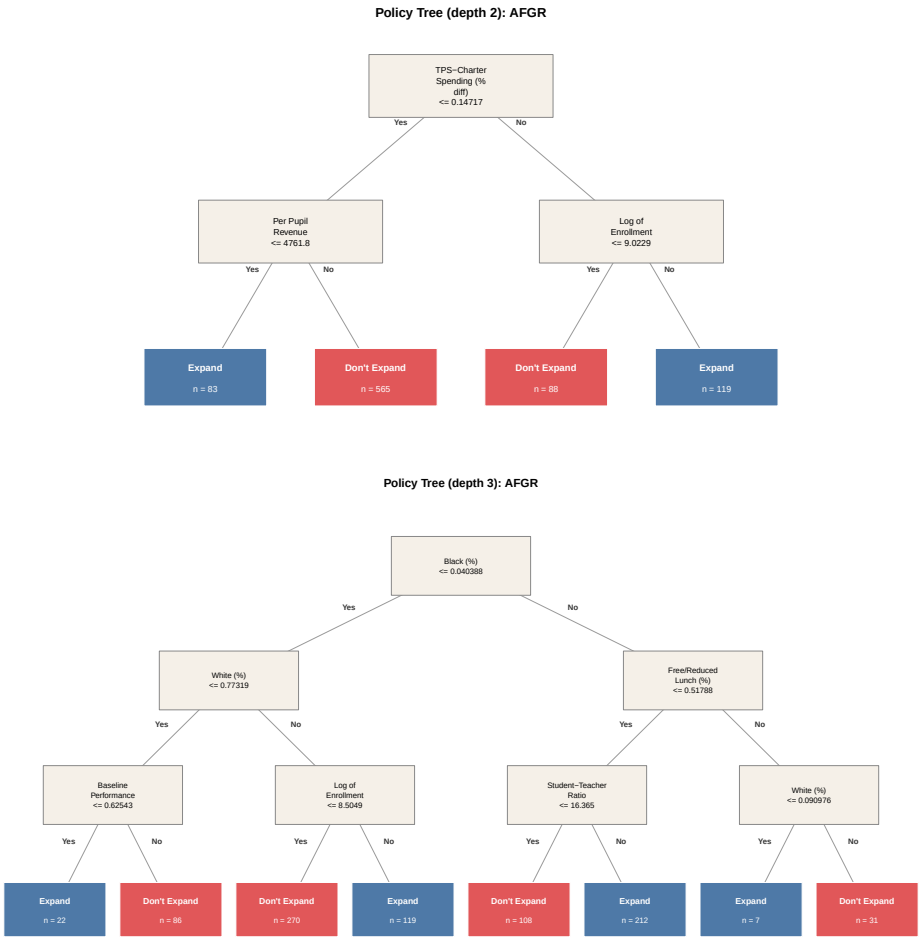


Figure S.B4. Optimal Policy Trees (Depth 2 and Depth 3): Graduation Rate (AFGR)

Notes: Depth-2 (top) and Depth-3 (bottom) optimal policy trees estimated via the Athey and Wager (2021) framework applied to DiDCF doubly robust scores for the graduation rate outcome (AFGR). Each internal node displays the key decision variable and threshold. The recommended policy action is indicated by the leaf colors: blue = expand charter market share; red = do not expand. We consider the ten highest-VIF moderators at $e = 0$ to construct the doubly-robust scores. The scores in question refer to those associated with the $CACRT$ such that the policy prescription refers to the binary decision to "marginally increase" charter intensity or not.

spending gap is large or if the district is predominantly White.

To summarize, these findings are similar to those of the two test score outcomes, though the relationships are more interactive and nonlinear. While it is more often the case that districts with more low-income and/or ethnic minority students (in larger districts) are more often positively impacted by charters, it is by no means guaranteed that a district with low-incomes or high non-White share will be positively impacted. Like

with math scores, the "demographically homogeneous" districts tend to be the ones more positively impacted.

S.B.3.2 Trait Set # 2: Market Context

Baseline AFGR ranks #6 in VIF score and remains important over time, indicating it is a consistent explainer of heterogeneity. No other market context variable carries notable information according to VIF. Perhaps because indicators like Urban vs. Suburban vs. Rural are better proxied by Log Enrollment according to DiDCF.

The relationship between a district's baseline AFGR and its $CATT$ appears to be more nonlinear and much weaker than with test scores. Our linear measures do not provide much evidence one way or another, but the BLP does weakly suggest that lower Baseline AFGR districts tend to have larger $CATT$ estimates, and the k-means table is fairly clear (though non linear in some groups) that the most positively impacted groups of districts are those with lower Baseline AFGR. Because Baseline AFGR was #6 in VIF, it does not get an event study, and the only instance of Baseline AFGR appearing in the policy trees was the final split of the depth-3 tree where districts with high White and/or Hispanic shares and low baseline AFGRs should expand charter schools.

To summarize: like with test scores, districts with lower Baseline AFGR are more often positively impacted. But this trend is not as clear as it was with test scores.

S.B.3.3 Trait Set #3: School Funding

School finance variables appear to drive heterogeneity in AFGR effects less than for test scores, but with some important interactions. The TPS-Charter Spending Difference (3), Student-Teacher Ratio (5), and Per-Pupil Revenue (9) each score in the top 10 for VIF.

These variables appear to have largely nonlinear relationships with the $CATT$. As was discussed in the **Demographics** subsection, the TPS-Charter Spending Gap appears to be interactive with both district Enrollment and Per-Pupil Revenue. Specifically, bigger spending gaps are associated with more positive $CATT$ estimates in larger districts, and smaller gaps are associated with positive effects in districts with low revenue per-student.

The other district finance variables do not seem to have linear or particularly meaningful relationships with the $CATT$. The k-means profiles do not exhibit clear trends for either variable, nor do the diff-in-means and BLP for Student-Teacher Ratio (Teacher Salary falls outside the top 10 by VIF and is not reported in those tables). Student-Teacher

Ratio does appear in the depth-3 policy tree, interacting with Black student share and FRL. Specifically, a high Student-Teacher ratio can be associated with positive $CATT$s in low-income, high-Black share districts.

S.B.3.4 Trait Set #4: State Charter Policy

Like with Math and ELA test scores, the policy setting variables generally score low on VIF. Districts where charter schools are automatically exempt from TPS regulations score #11, and districts with Equitable Funding agreements score #14. No other policy variable generates a VIF score greater than 0.01. Because these variables have low VIF scores, they do not appear in the BLP or quartile event studies, and (unsurprisingly) do not show up in either of the learned policy trees.

Our metrics have little to say about how $CATT$s vary with these moderators. They fall outside the top 10 by VIF, so they do not appear in the CGS-DiD, diff-in-means, or BLP tables. The k-means profiles suggest that these variables are likely interactive as (for example) the best and worst performing groups have a similar share of districts which are exempt from state/local regulations. Regardless, these interactions are likely weak explainers of $CATT$ heterogeneity given they do not appear in the policy trees and have low VIFs.

S.B.3.5 Trait Set #5: Diminishing Returns

We refer the reader to Section 4.3.6 for a review of the methodology for estimating the dose-response and marginal-response functions. Unlike with test scores where diminishing marginal return patterns were fairly clear, we do not find clear evidence of this for AFGR. We generally lack power to estimate precise $ACRT(w)$ parameters at any dose level for AFGR, and Figure S.B5 suggests there could be a U-shaped pattern in marginal effects. At low charter share (between 0-10%) there are diminishing-but-positive returns to additional charter schools. Marginal returns from 10-75% charter share are negative, before turning positive again at very high doses.

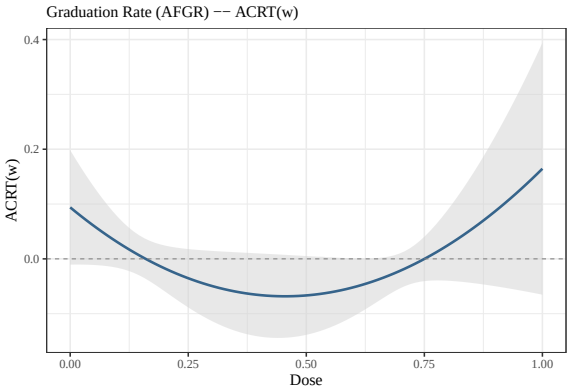


Figure S.B5. Diminishing Returns to additional Charter Intensity (DiDCF) – AFGR

Notes: This figure reports $ACRT(w)$ parameters (Callaway et al., 2025) across dose levels for the graduation rate outcome. 95% confidence intervals are displayed. This figure is interpreted as the average marginal effect of additional charter schools evaluated at dose w.

S.C Additional Effect Heterogeneity

S.C.1 Math vs. ELA Correlation

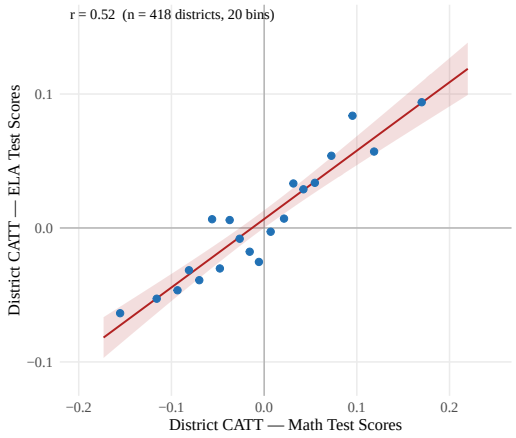


Figure S.C1. Math $\widehat{CATT}$ vs. ELA $\widehat{CATT}$ Correlation

Notes: This figure reports a binscatter plot of $CATT$ estimates for Math and ELA test scores within district. We also display the linear fit line and the correlation coefficient.

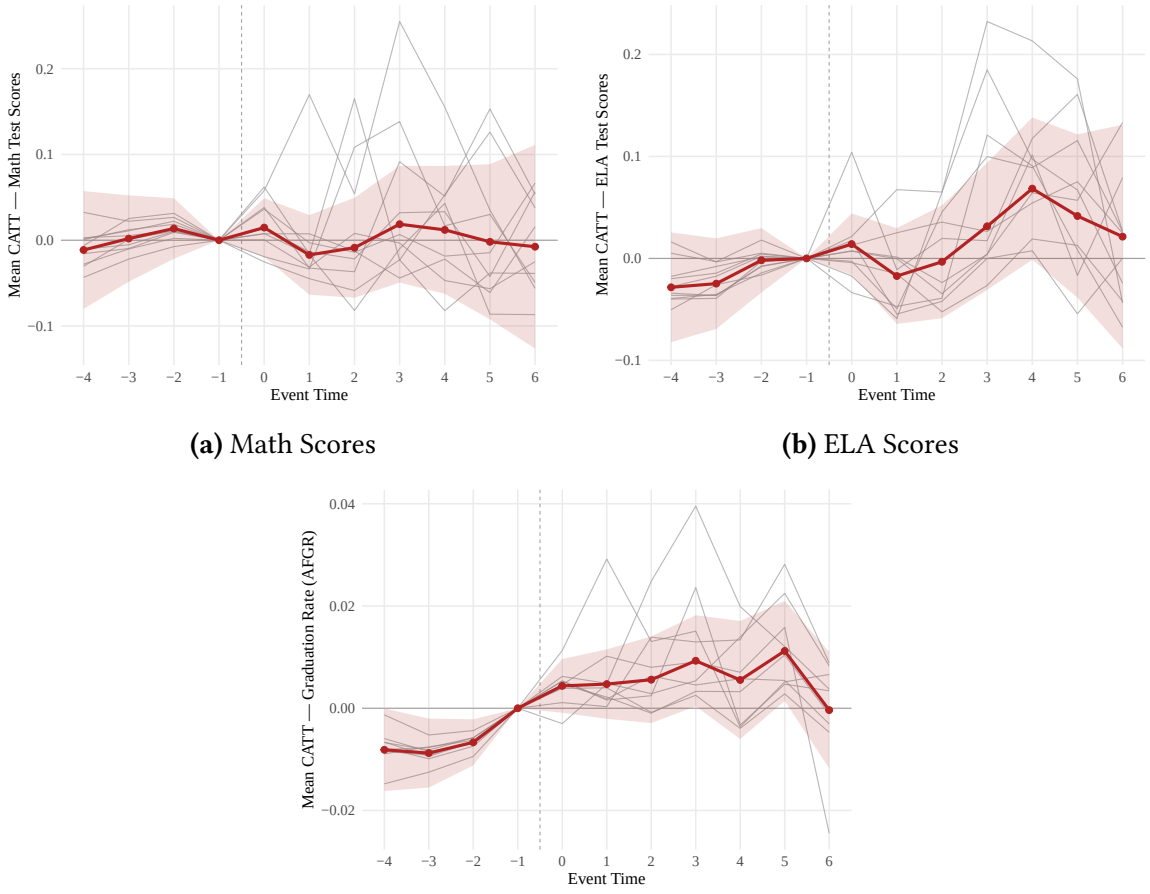
S.C.2 Summary Statistics

Table S.C1. Descriptive Statistics for the Test-Score Sample

Variable	Ever-Treated	Never-Treated	Difference	Range
Panel A: Outcomes				
Math test score	-0.253	0.278	-0.531^{***}	(-10.328, 7.533)
ELA test score	-0.277	0.285	-0.563^{***}	(-7.424, 6.256)
Panel B: Student Demographics				
White student share	0.408	0.690	-0.281^{***}	(0.000, 1.000)
Black student share	0.238	0.112	0.126^{***}	(0.000, 1.000)
Hispanic student share	0.285	0.136	0.149^{***}	(0.000, 1.000)
Free or reduced-price lunch share	0.557	0.405	0.152^{***}	(0.000, 1.000)
Special education share	0.125	0.129	-0.005^{***}	(0.000, 0.568)
Panel C: Market Setting				
Urban district	0.483	0.110	0.373^{***}	(0.000, 1.000)
Suburban district	0.400	0.432	-0.032^{***}	(0.000, 1.000)
Town district	0.050	0.173	-0.124^{***}	(0.000, 1.000)
Rural district	0.068	0.285	-0.217^{***}	(0.000, 1.000)
City population (standardized)	-0.019	-0.009	-0.009^{***}	(-1.873, 3.129)
District enrollment size (log)	10.165	8.390	1.775^{***}	(2.639, 13.441)
Number of magnet schools	16.620	0.466	16.154^{***}	(0.000, 266.000)
University presence (standardized)	-0.072	-0.002	-0.070^{***}	(-0.572, 6.346)
Panel D: School Revenue and Spending				
Per-pupil spending ($1,000s)	12.3	12.2	0.1^{***}	(8.3, 23.2)
Per-pupil revenue ($1,000s)	12.0	12.1	-0.1^{***}	(8.5, 22.8)
TPS–charter spending gap	0.124	0.003	0.121^{***}	(-0.657, 1.927)
Equitable operational funding	0.538	0.463	0.076^{***}	(0.000, 1.000)
Per-pupil charter-school spending ($1,000s)	8.3	0.3	8.0^{***}	(0.0, 29.8)
Teacher salary ($1,000s)	99.7	94.7	5.1^{***}	(63.8, 139.6)
Student–teacher ratio	16.504	15.529	0.974^{***}	(10.162, 21.815)
Panel E: State Charter-School Policies				
Regulatory exemptions	0.635	0.560	0.075^{***}	(0.000, 1.000)
No caps on charter-school growth	0.830	0.677	0.153^{***}	(0.000, 1.000)
Clear (non)renewal procedures	0.673	0.608	0.065^{***}	(0.000, 1.000)
Transparency	0.567	0.497	0.070^{***}	(0.000, 1.000)

Notes: Ever-treated districts have positive charter enrollment at some point in the sample; never-treated districts do not. The Difference column reports ever-treated minus never-treated means. Estimates are weighted by grade-level enrollment. $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

S.C.3 k-Means Matched Neighborhoods



(a) Math Scores

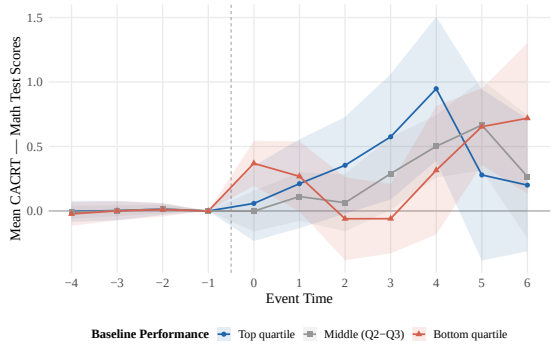
(b) ELA Scores

(c) Graduation Rate (AFGR)

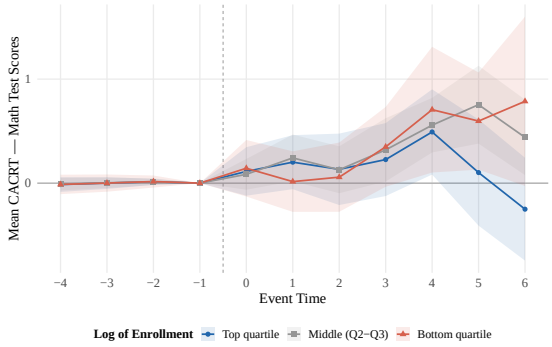
Figure S.C2. DiDCF K-Means Matched Event Studies (CATT): Math, ELA, and AFGR

Notes: Each panel shows $\widehat{GATT}^{es}(e, \mathcal{A})$ estimates (Equation 25). These are group average treatment effects in each event-time, where the groups are districts in each of the 10 k-means clusters. Each gray line is a single cluster, the red line is the overall event study line.

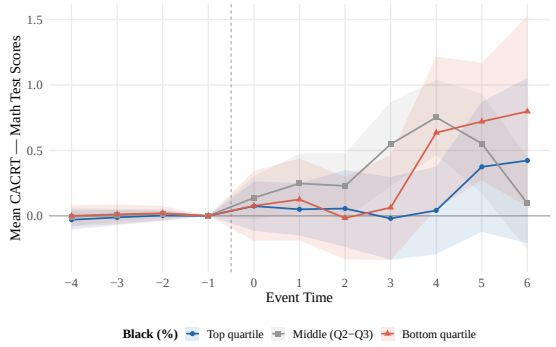
S.C.4 CACRT Event Studies



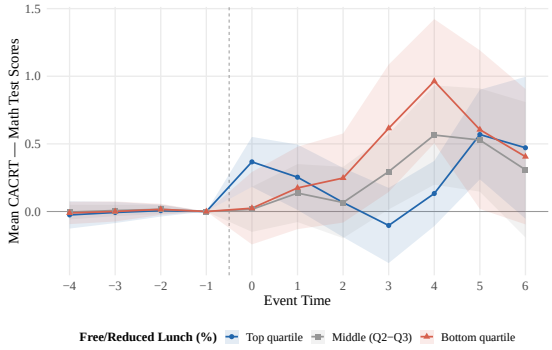
(a) Baseline Math Scores



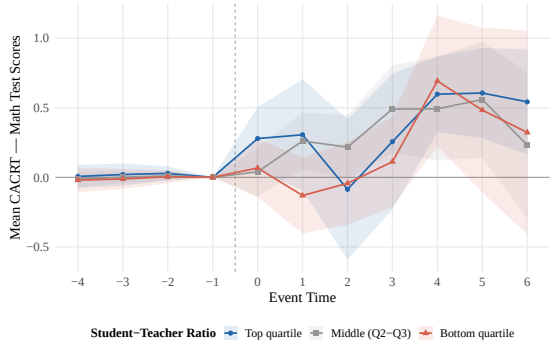
(b) Log Enrollment



(c) Black (%)



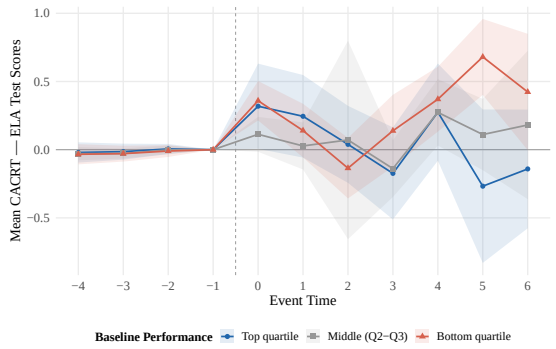
(d) Free/Reduced Price Lunch (%)



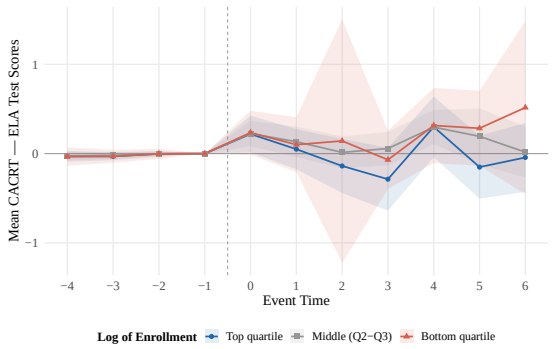
(e) Student-Teacher Ratio

Figure S.C3. DiDCF Quartile Event Studies (CACRT): Math Scores

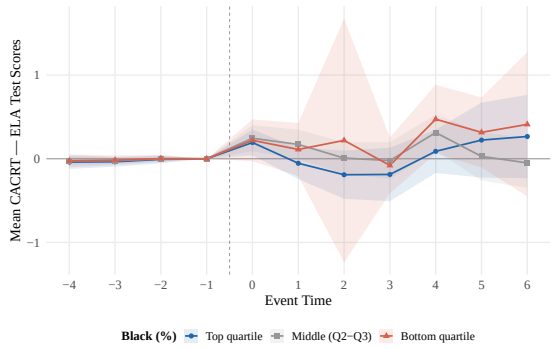
Notes: Each panel shows $\widehat{GACRT}^{es}(e, \mathcal{A})$ estimates (Equation 25). These are group average causal responses in each event-time, where the groups are districts in the top quartile (blue), middle two quartiles (grey), and bottom quartile (red) of the named moderator. Panels correspond to the five highest-ranked covariates by VIF at $e = 0$ for math scores.



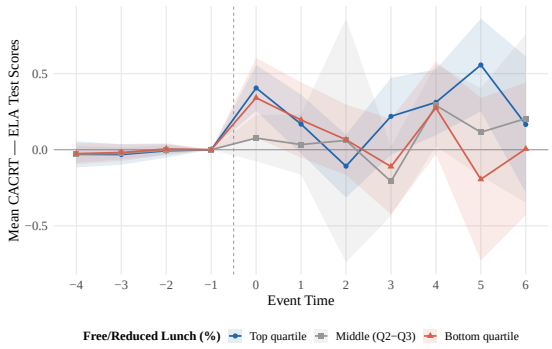
(a) Baseline ELA Scores



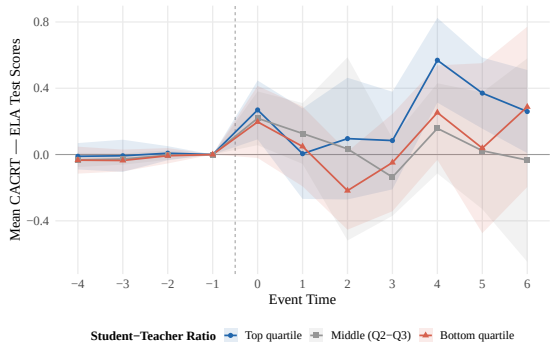
(b) Log Enrollment



(c) Black (%)



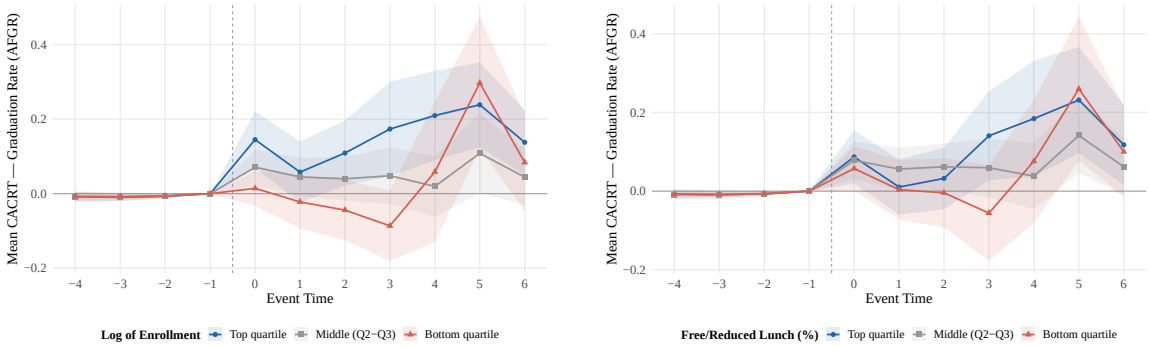
(d) Free/Reduced Price Lunch (%)



(e) Student-Teacher Ratio

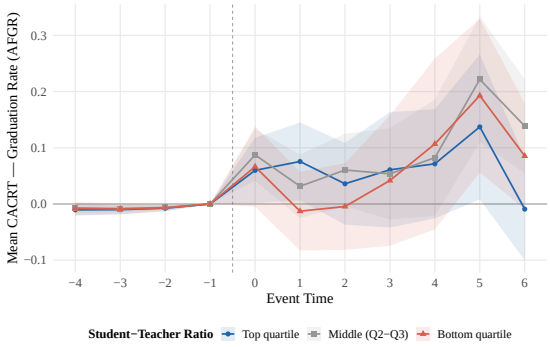
Figure S.C4. DiDCF Quartile Event Studies (CACRT): ELA Scores

Notes: Each panel shows $\widehat{GACRT}^{es}(e, \mathcal{A})$ estimates (Equation 25). These are group average causal responses in each event-time, where the groups are districts in the top quartile (blue), middle two quartiles (grey), and bottom quartile (red) of the named moderator. Panels correspond to the five highest-ranked covariates by VIF at $e = 0$ for ELA scores.

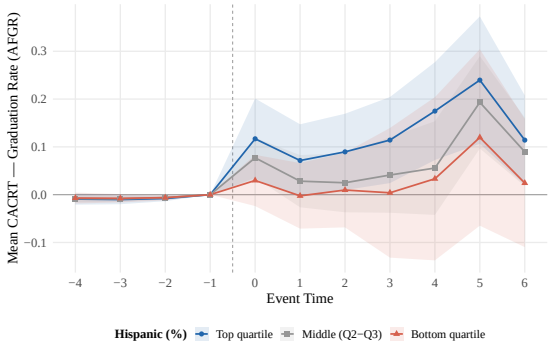


(a) Log Enrollment

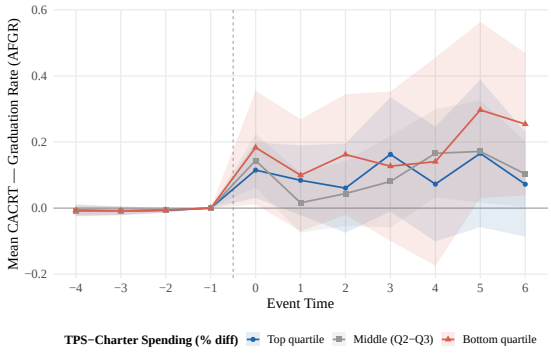
(b) Percent Free/Reduced Lunch



(c) Student-Teacher Ratio



(d) Percent Hispanic

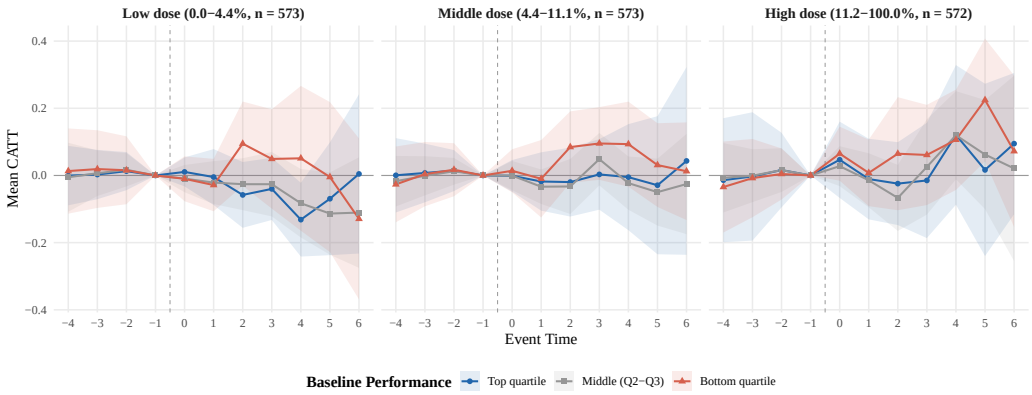


(e) % Difference TPS vs. Charter Spending

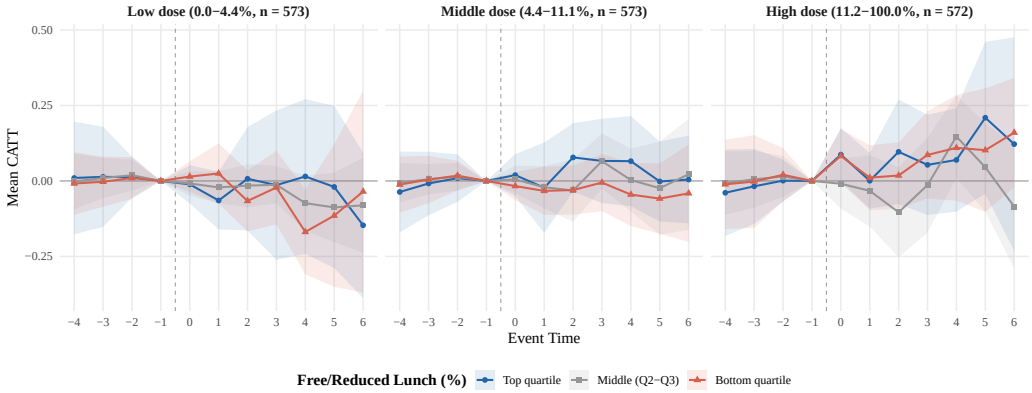
Figure S.C5. DiDCF Quartile Event Studies (CACRT): Graduation Rate (AFGR)

Notes: Each panel shows $\widehat{GACRT}^{es}(e, \mathcal{A})$ estimates (Equation 25). These are group average causal responses in each event-time, where the groups are districts in the top quartile (blue), middle two quartiles (grey), and bottom quartile (red) of the named moderator. Panels correspond to the five highest-ranked covariates by VIF at $e = 0$ for graduation rates.

S.C.5 Event Studies within Dose Groups



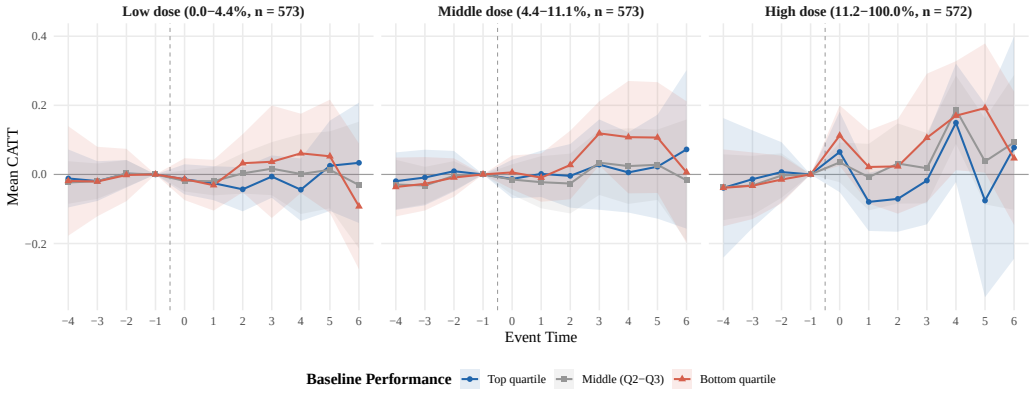
(a) Baseline Math Scores



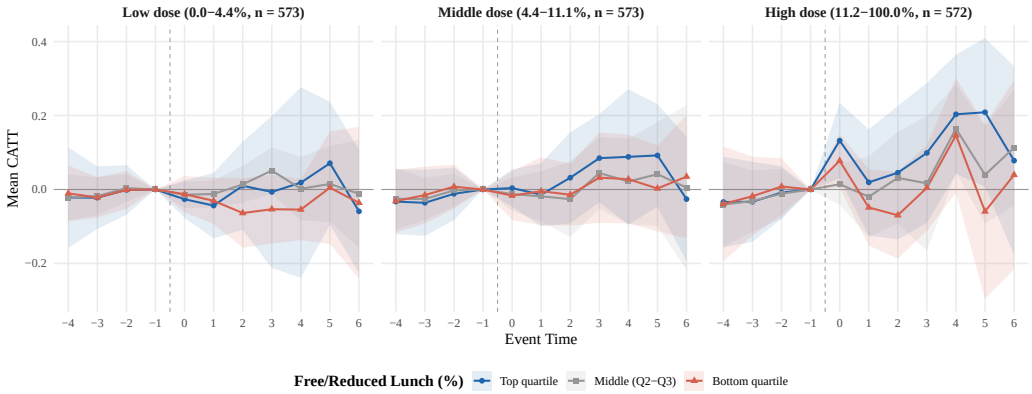
(b) Free/Reduced Price Lunch (%)

Figure S.C6. DiDCF Quartile Event Studies (CATT) by Dose Groups: Math Scores

Notes: Each panel shows $\widehat{GATT}^{es}(e, \mathcal{A})$ estimates (Equation 25) for Math test scores. These are group average treatment effects in each event-time, where the groups are districts in the top quartile (blue), middle two quartiles (grey), and bottom quartile (red) of the named moderator. Panels correspond to dose tertile group, e.g. "low dose" corresponds to the bottom 33% of districts by charter share. The top row is for the Baseline Math score and the bottom row is for the FRL share.



(a) Baseline ELA Scores

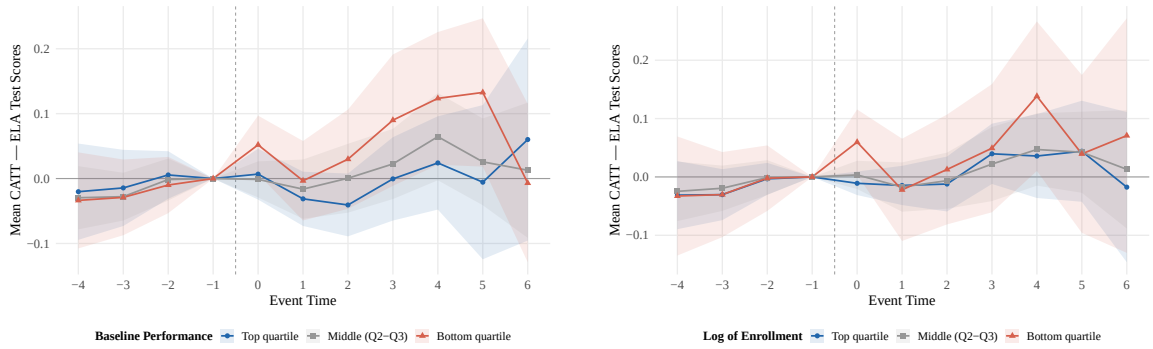


(b) Free/Reduced Price Lunch (%)

Figure S.C7. DiDCF Quartile Event Studies (CATT) by Dose Groups: ELA Scores

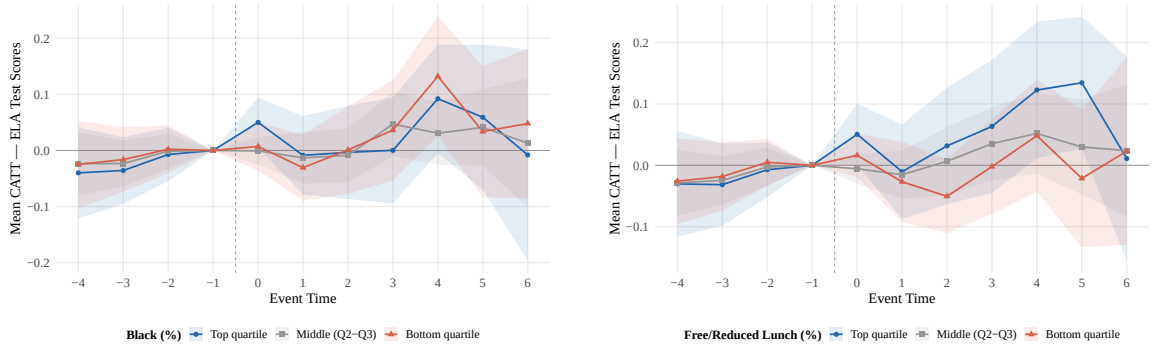
Notes: Each panel shows $\widehat{GATT}^{es}(e, \mathcal{A})$ estimates (Equation 25) for ELA test scores. These are group average treatment effects in each event-time, where the groups are districts in the top quartile (blue), middle two quartiles (grey), and bottom quartile (red) of the named moderator. Panels correspond to dose tertial group, e.g. "low dose" corresponds to the bottom 33% of districts by charter share. The top row is for the Baseline ELA score and the bottom row is for the FRL share.

S.C.6 Additional ELA Figures



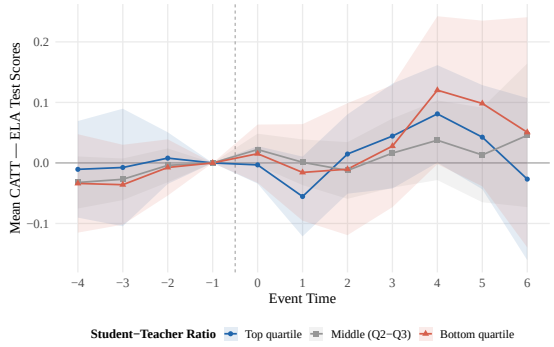
(a) Baseline ELA Scores

(b) Log Enrollment



(c) Black (%)

(d) Free/Reduced Price Lunch (%)



(e) Student-Teacher Ratio

Figure S.C8. DiDCF Quartile Event Studies (CATT): ELA Scores

Notes: Each panel shows $\widehat{GATT}^{es}(e, \mathcal{A})$ estimates (Equation 25). These are group average treatment effects in each event-time, where the groups are districts in the top quartile (blue), middle two quartiles (grey), and bottom quartile (red) of the named moderator. Panels correspond to the five highest-ranked covariates by VIF at $e = 0$ for ELA scores.

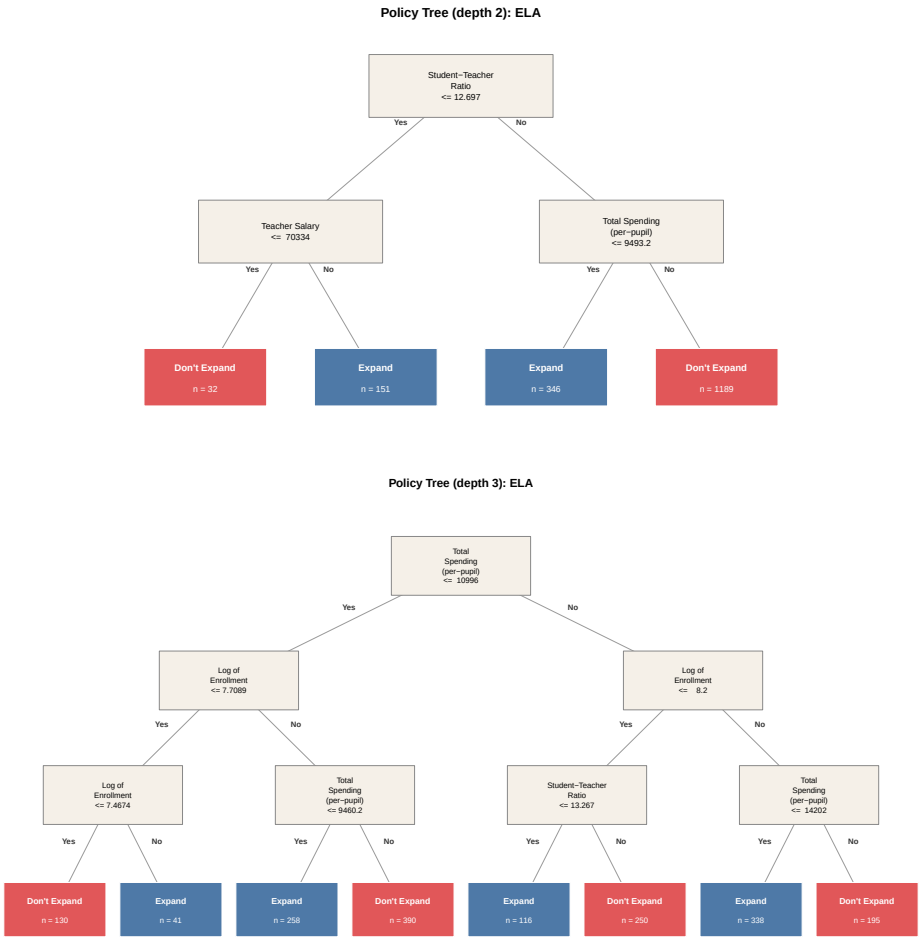


Figure S.C9. Optimal Policy Trees (Depth 2 and Depth 3): ELA Scores

Notes: Depth-2 (top) and Depth-3 (bottom) optimal policy trees estimated via the Athey and Wager (2021) framework applied to DiDCF doubly robust scores for the ELA test scores outcome (ELA). Each internal node displays the key decision variable and threshold. The recommended policy action is indicated by the leaf colors: blue = expand charter market share; red = do not expand. We consider the ten highest-VIF moderators at $e = 0$ to construct the doubly-robust scores. The scores in question refer to those associated with the $CACRT$ such that the policy prescription refers to the binary decision to "marginally increase" charter intensity or not.

S.C.7 CGS-DiD (Linear Parametric) Heterogeneity

Table S.C2. Heterogeneity in the Effects of Charter-School Expansion: CGS-DiD

	Math	ELA
Moderator	(1)	(2)
Baseline Treatment Effects		
Positive charter exposure	−0.045	−0.219***
	(0.097)	(0.084)
Student Demographics		
White student share	0.020	0.624**
	(0.300)	(0.284)
Black student share	−0.101	0.291
	(0.314)	(0.301)
Hispanic student share	0.043	0.327
	(0.297)	(0.297)
Free or reduced-price lunch share	0.071	0.173
	(0.118)	(0.134)
Special education share	−0.529	−0.744
	(0.601)	(0.559)
Market Setting		
Baseline outcome	−0.046**	−0.049*
	(0.021)	(0.026)
Urban district	−0.084	0.005
	(0.057)	(0.055)
Suburban district	−0.202***	−0.050
	(0.057)	(0.053)
Town district	−0.275***	−0.109**
	(0.057)	(0.053)
Log enrollment	0.051**	0.067***
	(0.022)	(0.022)
Magnet schools	0.013	−0.013
	(0.010)	(0.009)
School Revenue and Spending		
Per-pupil spending ($1,000s)	0.003	0.002
	(0.012)	(0.011)
Per-pupil revenue ($1,000s)	0.026*	0.006
	(0.014)	(0.014)
TPS–charter spending gap	0.032	0.182*
	(0.123)	(0.110)
Equitable operational funding	−0.089	−0.063
	(0.129)	(0.121)
Charter-school spending ($1,000s)	0.021**	0.019**
	(0.010)	(0.010)
Teacher salary ($1,000s)	−0.004**	−0.000
	(0.002)	(0.002)

Continued on next page

Table S.C2 continued

Moderator	Math (1)	ELA (2)
Student–teacher ratio	0.021^{**} (0.010)	-0.004 (0.009)
Other State Charter-School Policies		
Regulatory exemptions	-0.019 (0.102)	-0.083 (0.088)
No caps on charter-school growth	-0.163 (0.132)	0.109 (0.116)
Clear renewal and nonrenewal procedures	-0.047 (0.138)	-0.074 (0.126)
Transparency	0.252^{**} (0.128)	0.081 (0.115)
Performance-based contracts	0.251 (0.178)	0.165 (0.148)
Treated observations	7,547	7,547

Notes: The table reports heterogeneity estimates from the staggered, continuous-dose CGS-DiD (Equation 15), estimated separately in each cohort g by calendar-year t cell. In the first stage, the outcome change $Y_t - Y_{g-1}$ of never-treated district-grades is regressed on the moderators X to estimate the counterfactual trend. In the second stage, the residualized change of cohort-g district-grades is regressed on $(1, X, W, WX)$, where W is the charter enrollment share, so $CATT(w, X) = \alpha_0 + X'\alpha + w(\tau_0 + X'\tau)$. Reported coefficients average the cell estimates over post-treatment cells, weighted by each cell's number of treated observations. We report α_0, the effect of positive charter exposure ($D = 1\{W > 0\}$) at the mean moderator vector, and each row reports α_k, the coefficient on $D \times X_k$; the dose coefficients τ_0 and τ_k are estimated but not reported. Moderators are measured in the first observed year for each district-grade, except the baseline outcome, Y_{g-2}, which is centered within each cell; missing values are set to the sample mean and flagged by missing-value indicators, which are included in the regressions but not reported. Dollar amounts are in thousands; urban, suburban, and town districts are relative to rural. Both stages are weighted by grade-level enrollment. Standard errors (in parentheses) are clustered by district and account for the first-stage estimation. $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

S.D Proofs

S.D.1 Proof of Proposition 1

Proof. Recall Equation 1:

$$CATT(g, t, w, x \mid g, w, x) \equiv \mathbb{E}[Y_{i,t}(g, w) - Y_{i,t}(0) \mid G_i = g,\ W_i = w,\ X_i = x]$$

Add and subtract the base period untreated outcome inside Equation 1:

$$\begin{aligned} CATT(g,t,w,x \mid g,w,x) = \mathbb{E}\big[Y_{i,t}(g,w) - Y_{i,b_g}(0) \mid G_i = g,\ W_i = w,\ X_i = x\big] \\ - \mathbb{E}\big[Y_{i,t}(0) - Y_{i,b_g}(0) \mid G_i = g,\ W_i = w,\ X_i = x\big] \end{aligned}$$

The second term is the counterfactual untreated trend of cohort g. Under Assumption 5, this is equal to the observed trend of never-treated districts with covariates x:

$$\begin{aligned} &\mathbb{E}\big[Y_{i,t}(0) - Y_{i,b_g}(0) \mid G_i = g,\ W_i = w,\ X_i = x\big] \\ &\quad = \mathbb{E}\big[Y_{i,t}(0) - Y_{i,b_g}(0) \mid G_i = \infty,\ W_i = 0,\ X_i = x\big] \end{aligned}$$

From here, we can replace potential outcomes with observed outcomes. For cohort-g districts, consistency gives $Y_{i,t} = Y_{i,t}(g,w)$ in the post-period, and because $b_g < g$, Assumption 3 ensures $Y_{i,b_g} = Y_{i,b_g}(g,w) = Y_{i,b_g}(0)$. For never-treated districts, $Y_{i,t} = Y_{i,t}(\infty,0) \equiv Y_{i,t}(0)$ in every period. Substituting:

$$\begin{aligned} CATT(g,t,w,x \mid g,w,x) = \mathbb{E}\big[Y_{i,t} - Y_{i,b_g} \mid G_i = g,\ W_i = w,\ X_i = x\big] \\ - \mathbb{E}\big[Y_{i,t} - Y_{i,b_g} \mid G_i = \infty,\ W_i = 0,\ X_i = x\big] \end{aligned}$$

which is the statement of Proposition 1.

S.D.2 Proof of Proposition 2

Proof. Recall Equation 2:

$$CACRT(g,t,w,x \mid g,w,x) \equiv \frac{\partial}{\partial l}\,\mathbb{E}[Y_{i,t}(g,l) \mid G_i = g,\ W_i = w,\ X_i = x]\Big|_{l=w}$$

Add and subtract the untreated base-period outcome $Y_{i,b_g}(0)$ inside the expectation. It does not depend on l:

$$\begin{aligned} \mathbb{E}[Y_{i,t}(g,l) \mid G_i = g,\ W_i = w,\ X_i = x] = \mathbb{E}\big[Y_{i,t}(g,l) - Y_{i,b_g}(0) \mid G_i = g,\ W_i = w,\ X_i = x\big] \\ + \mathbb{E}\big[Y_{i,b_g}(0) \mid G_i = g,\ W_i = w,\ X_i = x\big] \end{aligned}$$

The first term is the trend that dose group w would have followed under dose l. Under

Assumption 6(a), this is equal to the trend of the districts that actually received dose l:

$$\begin{aligned}&\mathbb{E}\big[Y_{i,t}(g,l) - Y_{i,b_g}(0) \mid G_i = g,\ W_i = w,\ X_i = x\big] \\ &\quad = \mathbb{E}\big[Y_{i,t}(g,l) - Y_{i,b_g}(0) \mid G_i = g,\ W_i = l,\ X_i = x\big]\end{aligned}$$

From here, we can replace potential outcomes with observed outcomes. For cohort-g districts with dose l, consistency gives $Y_{i,t} = Y_{i,t}(g,l)$ in the post-period, and Assumption 3 ensures $Y_{i,b_g} = Y_{i,b_g}(g,l) = Y_{i,b_g}(0)$. Substituting:

$$\begin{aligned}\mathbb{E}[Y_{i,t}(g,l) \mid G_i = g,\ W_i = w,\ X_i = x] = \mathbb{E}\big[Y_{i,t} - Y_{i,b_g} \mid G_i = g,\ W_i = l,\ X_i = x\big] \\ + \mathbb{E}\big[Y_{i,b_g}(0) \mid G_i = g,\ W_i = w,\ X_i = x\big]\end{aligned}$$

Differentiating with respect to l at $l = w$, the second term drops out and

$$CACRT(g,t,w,x \mid g,w,x) = \frac{\partial}{\partial l}\,\mathbb{E}\big[Y_{i,t} - Y_{i,b_g} \mid G_i = g,\ W_i = l,\ X_i = x\big]\Big|_{l=w}$$

which is the statement of Proposition 2. This is the conditional analogue of the second part of Theorem C.1 in Callaway et al. (2025).

S.D.3 Proof of Proposition 3

Proof. Add and subtract $Y_{i,b_g}(0)$ inside Equation 8. The first term, $\mathbb{E}[Y_{i,t}(g, \mathbf{W}_{i,g:t}) - Y_{i,b_g}(0) \mid G_i = g,\ W_{i,t} = w,\ X_i = x]$, equals $\mathbb{E}[Y_{i,t} - Y_{i,b_g} \mid G_i = g,\ W_{i,t} = w,\ X_i = x]$ because $Y_{i,t} = Y_{i,t}(G_i, \mathbf{W}_{i,G_i:t})$ for $t \geq G_i$ and $Y_{i,b_g} = Y_{i,b_g}(0)$ by Assumption 3. The second term, $\mathbb{E}[Y_{i,t}(0) - Y_{i,b_g}(0) \mid G_i = g,\ W_{i,t} = w,\ X_i = x]$, is the never-treated trend by Assumption 5, where $Y_{i,t} = Y_{i,t}(0)$ for all t.

S.D.4 Proof of Proposition 4

Proof. Denote $\mathbf{h} = \mathbf{w}_{g:t-1}$ and drop $G_i = g$ and $X_i = x$ from the notation for simplicity. The base-period outcome $Y_{i,b_g}(0)$ does not depend on the current dose l (Assumption 3), so subtracting it does not change the derivative in Equation 11:

$$CACRT^{path}(g,t,\mathbf{w}_{g:t},x \mid g,\mathbf{w}_{g:t},x) = \frac{\partial}{\partial l}\,\mathbb{E}[Y_{i,t}(g,(\mathbf{h},l)) - Y_{i,b_g}(0) \mid \mathbf{W}_{i,g:t} = (\mathbf{h},w_t)]\Big|_{l=w_t}$$

By Assumption 7(a), this expected change equals the expected change of districts that actually followed $(\mathbf{h}, l)$, for whom it is observed:

$$\mathbb{E}[Y_{i,t}(g,(\mathbf{h},l)) - Y_{i,b_g}(0) \mid \mathbf{W}_{i,g:t} = (\mathbf{h}, w_t)] = \mathbb{E}[Y_{i,t} - Y_{i,b_g} \mid \mathbf{W}_{i,g:t-1} = \mathbf{h},\ W_{i,t} = l]$$

Differentiating at $l = w_t$ gives Equation 13.

S.D.5 Derivation of Equation 14

Proof. Drop $G_i = g$ and $X_i = x$ for simplicity. Denote $\Delta Y_{i,g,t} = Y_{i,t} - Y_{i,b_g}$, and assume the expectations below are differentiable in the dose, with derivatives that can be taken inside expectations over histories. By Proposition 3, the never-treated term does not depend on l, so $\partial\, CATT^{path}(g,t,l,x \mid g,l,x)/\partial l = \partial\, \mathbb{E}[\Delta Y_{i,g,t} \mid W_{i,t} = l]/\partial l$. By iterated expectations (first line), then by the chain rule (line 2):

$$\begin{aligned}
\frac{\partial}{\partial l}\,\mathbb{E}[\Delta Y_{i,g,t} \mid W_{i,t} = l] &= \frac{d}{dl}\,\mathbb{E}\Big[\mathbb{E}[\Delta Y_{i,g,t} \mid \mathbf{W}_{i,g:t-1},\, W_{i,t} = l] \;\Big|\; W_{i,t} = l\Big] \\
&= \mathbb{E}\left[\frac{\partial\, \mathbb{E}[\Delta Y_{i,g,t} \mid \mathbf{W}_{i,g:t-1},\, W_{i,t} = l]}{\partial l} \;\middle|\; W_{i,t} = l\right] \\
&\quad + \frac{\partial}{\partial l'}\,\mathbb{E}\Big[\mathbb{E}[\Delta Y_{i,g,t} \mid \mathbf{W}_{i,g:t-1},\, W_{i,t} = l] \;\Big|\; W_{i,t} = l'\Big]\Big|_{l'=l}
\end{aligned}$$

For the first term, Proposition 4 (first equality) and Equation 12 (second equality) give:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\partial\, \mathbb{E}[\Delta Y_{i,g,t} \mid \mathbf{W}_{i,g:t-1},\, W_{i,t} = l]}{\partial l} \;\middle|\; W_{i,t} = l\right] \\
&\quad = \mathbb{E}\big[CACRT^{path}(g,t,\mathbf{W}_{i,g:t},x \mid g,\mathbf{W}_{i,g:t},x) \;\big|\; W_{i,t} = l\big] \\
&\quad = CACRT^{path}(g,t,l,x \mid g,l,x)
\end{aligned}$$

The second term asks "if we keep the current dose fixed and vary only the dose history, how does the average $\Delta Y_{i,g,t}$ change?"

Consider a path $\mathbf{w}_{g:t}$ with history $\mathbf{w}_{g:t-1}$. Then

$$\begin{aligned}
\mathbb{E}[\Delta Y_{i,g,t} \mid \mathbf{W}_{i,g:t} = \mathbf{w}_{g:t}] &= \mathbb{E}[Y_{i,t}(g,\mathbf{w}_{g:t}) - Y_{i,b_g} \mid \mathbf{W}_{i,g:t} = \mathbf{w}_{g:t}] \\
&= \mathbb{E}[Y_{i,t}(g,\mathbf{w}_{g:t}) - Y_{i,b_g}(0) \mid \mathbf{W}_{i,g:t} = \mathbf{w}_{g:t}] \\
&= \mathbb{E}[Y_{i,t}(g,\mathbf{w}_{g:t}) - Y_{i,b_g}(0)]
\end{aligned}$$

The third line holds because Assumption 7(a) makes $\mathbb{E}[Y_{i,t}(g, \mathbf{w}_{g:t}) - Y_{i,b_g}(0) \mid \mathbf{W}_{i,g:t} = \mathbf{l}_{g:t}]$ the same for every $\mathbf{l}_{g:t}$, so averaging over the distribution of $\mathbf{W}_{i,g:t}$ returns line 3. Under Assumption 8, $\mathbb{E}[Y_{i,t}(g, \mathbf{w}_{g:t})]$ depends on $\mathbf{w}_{g:t}$ only through w_t, and $\mathbb{E}[Y_{i,b_g}(0)]$ does not depend on $\mathbf{w}_{g:t}$. Hence $\mathbb{E}[\Delta Y_{i,g,t} \mid \mathbf{W}_{i,g:t-1}, W_{i,t} = l]$ is the same value for every history $\mathbf{W}_{i,g:t-1}$, its average over the histories of districts with $W_{i,t} = l'$ does not depend on l', so the second term is zero.

S.D.6 Proof of Proposition S.1

Proof. From Equation S.A1, we know:

$$\operatorname{plim} \hat{\tau}(x) = \frac{\operatorname{Cov}(\Delta\widetilde{Y}_i, W_i \mid X_i)}{\operatorname{Var}(W_i \mid X_i)}$$

Plug Equation S.A2 into $\operatorname{Cov}(\Delta\widetilde{Y}_i, W_i \mid X_i)$. Noting that in order to move from the second line to the third, we notice that within x, $\alpha(x)$ and $\tau(x)$ are scalars. Therefore $\operatorname{Cov}\Big(\alpha(x), W_i \mid X_i\Big) = 0$ and $\operatorname{Cov}\Big(\tau(x)W_i, W_i \mid X_i\Big) = \tau(x)\operatorname{Cov}\Big(W_i, W_i \mid X_i\Big) = \tau(x)\operatorname{Var}(W_i \mid X_i)$. Also, $\operatorname{Cov}(\varepsilon, W_i \mid X_i) = 0$ because $\mathbb{E}[\varepsilon_i \mid W_i, X_i] = 0$. Going forward, we drop subscripts to avoid clutter:

$$\begin{aligned}
\operatorname{Cov}(\Delta\widetilde{Y}, W \mid X) &= \operatorname{Cov}\Big(\alpha(x) + \tau(x)W + \delta + \varepsilon, W \mid X\Big) \\
&= \operatorname{Cov}\Big(\alpha(x), W \mid X\Big) + \operatorname{Cov}\Big(\tau(x)W, W \mid X\Big) \\
&\quad + \operatorname{Cov}\Big(\delta, W \mid X\Big) + \operatorname{Cov}\Big(\varepsilon, W \mid X\Big) \\
&= \tau(x)\operatorname{Var}\Big(W \mid X\Big) + \operatorname{Cov}\Big(\delta, W \mid X\Big)
\end{aligned}$$

Divide this result by $\operatorname{Var}(W \mid X)$:

$$\begin{aligned}
\operatorname{plim} \hat{\tau}(x) &= \frac{\tau(x)\operatorname{Var}(W \mid X) + \operatorname{Cov}(\delta, W \mid X)}{\operatorname{Var}(W \mid X)} \\
&= \tau(x) + \underbrace{\frac{\operatorname{Cov}(\delta, W \mid X)}{\operatorname{Var}(W \mid X)}}_{=0} \\
&= \tau(x)
\end{aligned}$$

For the second half of Proposition S.1, note that the correct model would be on the

true DGP (i.e. the expectation of Equation S.A2). This value is $\mathbb{E}[\Delta\widetilde{Y} \mid X] = \mathbb{E}[\alpha(x) \mid X] + \mathbb{E}[\tau(x)W \mid X] + \mathbb{E}[\delta \mid X] + \mathbb{E}[\varepsilon \mid X] = \alpha(x) + \mathbb{E}[\tau(x)W \mid X] + \mathbb{E}[\delta \mid X]$. The bias between our DiDCF CATT and the true CATT is therefore:

$$\operatorname{plim} \hat{\alpha}(x) - \alpha(x) = \mathbb{E}[\delta \mid X] - \big(\operatorname{plim} \hat{\tau}(x) - \tau(x)\big)\mathbb{E}[W \mid X]$$

The first part of this proof showed that $\operatorname{plim} \hat{\tau}(x) - \tau(x) = 0$, so $\operatorname{plim} \hat{\alpha}(x) - \alpha(x) = \mathbb{E}[\delta \mid X]$, which is the second claim in Proposition S.1.

S.D.7 Proof of Proposition S.2

Proof. Let $\sigma_\theta^2 = \operatorname{Var}(\theta)$. Then $\operatorname{Cov}(\theta, \hat{\theta}) = \operatorname{Var}(\theta) + \operatorname{Cov}(\theta, \eta) = \sigma_\theta^2(1 + rs)$ and $\operatorname{Var}(\hat{\theta}) = \operatorname{Var}(\theta) + 2\operatorname{Cov}(\theta, \eta) + \operatorname{Var}(\eta) = \sigma_\theta^2(1 + 2rs + s^2)$. Divide: $\operatorname{Corr}(\theta, \hat{\theta}) = \sigma_\theta^2(1 + rs)/[\sigma_\theta^2\sqrt{1 + 2rs + s^2}]$, which is Equation S.A5.

S.E Technical Summary of GRF Estimation Procedure

There are two important concepts to note before explaining the GRF-training algorithm. First, GRF performs Neyman orthogonalization (i.e. local recentering) (Robinson, 1988; Chernozhukov et al., 2018; Nie and Wager, 2021) by differencing out the propensity score and conditional mean from W_i and Y_i, respectively. These propensity score functions are estimated via auxiliary regression trees with leave-one-out estimates. The purpose of this is to improve finite sample performance of the model, as GRF estimates can be biased in finite samples if the conditioning moderators in X_i are not fully partialled out because of the small sample size. Athey et al. (2019) show that this reduces finite sample bias in these situations. Second, trees are estimated based on the concept of "honesty" developed by Athey and Imbens (2016). The consistency and asymptotic normality of non-parametric estimators relies on cross-training: using a different sample for model estimation than the one used for model training. The reason is that using the same sample for training and estimation introduces dependence between the structure of the model and the estimates themselves. Therefore, an honest causal tree takes the full data sample and extracts a random sub-sample to grow the tree, then further divides this sample into a training sample and an estimation sample. The training sample is used to perform splitting and define the tree structure, the estimation sample is used to populate the leaves and

provide $CATE$ estimates.

A generalized random forest grows B causal trees indexed by $b = 1, ..., B$. Each tree is grown by the following steps:

1. From the full data sample, draw a random subsample without replacement.

2. Further divide the subsample into a training sample, $\mathcal{J}_1$, and an estimation sample, $\mathcal{J}_2$, which are non-overlapping partitions.

3. With training sample $\mathcal{J}_1$, perform greedy and recursive splitting by grid searching over the covariate space and selecting the single covariate, $j \in \{1, ..., k\}$, and value, c, that maximizes the squared difference in treatment effects across the resulting child nodes. The objective function is given as:

$$\max_{j,c} \quad |C_1||C_2|(\hat{\tau}_{C_1} - \hat{\tau}_{C_2})^2$$

   Where C_1 and C_2 denote the resulting child nodes such that $|C_1|$ and $|C_2|$ are the number of units in each resulting child node and $\hat{\tau}_{C_1}$ and $\hat{\tau}_{C_2}$ are the resulting ATEs within each child node. For computational tractability (i.e. in order to avoid estimating linear regressions for each potential split), Athey et al. (2019) show that this process reduces to splitting on a pseudo-outcome corresponding to a first-order gradient-based approximation of the subsample ATEs.

   (a) Denote a parent node, P, as any node that is currently being assessed for an optimal split. Compute $\hat{\alpha}_P$ and $\hat{\tau}_P$ by solving Equation 19, which can be alternatively expressed as:

$$\sum_{i=1}^{|P|}(1, W_i)'(Y_i - \hat{\alpha}_P - \hat{\tau}_P W_i) = 0$$

   (b) Also compute an estimate of the gradient of the expectation of $\psi_{\alpha,\tau}(Y_i, W_i) = (1, W_i)'(Y_i - \alpha - \tau W_i)$, denoted $A_P = \nabla\mathbb{E}[\psi_{\hat{\alpha}_P,\hat{\tau}_P}|X_i \in P]$:

$$A_P = \frac{1}{|\{i : X_i \in P\}|}\sum_{i:X_i \in P} \nabla\psi_{\hat{\alpha}_P,\hat{\tau}_P}(Y_i, W_i)$$

(c) Compute the pseudo-outcomes as:

$$\rho_i = -\xi' A_P^{-1} \psi_{\hat{\alpha}_P, \hat{\tau}_P}$$

Where ξ is simply a vector that picks out the τ coordinate from the $\big(\alpha, \tau\big)$ vector.

(d) Finally, use the pseudo-outcome in a standard classification and decision tree (CART), which just selects the split that maximizes the sum of the average influence functions between the resulting child nodes:

$$\widetilde{\Delta}(C_1, C_2) = \sum_{q=1}^{2} \frac{1}{|\{i : X_i \in C_q\}|} \Big(\sum_{\{i: X_i \in C_q\}} \rho_i \Big)^2$$

Intuitively, the influence function captures how much the treatment effect in the child node is likely to differ from the treatment effect estimate in the parent node. In the simplest case of a randomized experiment where the above process refers to OLS-based identification (without a constant, for simplicity), this first-order approximation looks like:

$$\hat{\tau}_{C_q} \approx \hat{\tau}_P + \frac{1}{\sum_{i \in C_q} W_i^2} \sum_{i \in C_q} W_i (Y_i - W_i \hat{\tau}_P)$$

(e) With the two child nodes, repeat steps (a) – (d) on each, treating each node as a new parent node until a full tree, $\mathcal{T}$, is trained.

4. Within tree $\mathcal{T}$, collect all observations in the estimation subsample $\mathcal{J}_2$ that fall in the same leaf as x, $L_b(x)$. Compute the GRF-weights as the share of total trees where unit i falls into the same leaf as x:

$$\omega_i(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{|L_b(x)|} \mathbb{1}\{X_i \in L_b(x)\}$$

In practice, we estimate each causal forest with 1,500 trees, with auxiliary forests estimated with 375 trees each. For each tree, we hold out half of the subsample for honest estimation, and require at least 20 observations in a resulting child node for a split to occur. We use the default rule-of-thumb for determining the number of variables considered

at each split point: if k is the number of covariates, select the minimum of $\sqrt{k} + 20$ or k. When training each tree, units are weighted by enrollment share for population representativeness, and we perform clustering at the district level. Variances of individualized $\left\{\hat{\alpha}(x), \hat{\tau}(x)\right\}$ are constructed using the bootstrap of little bags, which is accomplished by training sets of trees and comparing estimates within and across those sets (Athey et al., 2019).

References

Athey, S. and G. Imbens (2016, jul). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360. Publisher: Proceedings of the National Academy of Sciences.

Athey, S., J. Tibshirani, and S. Wager (2019, apr). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178. Publisher: Institute of Mathematical Statistics.

Athey, S. and S. Wager (2021). Policy learning with observational data. *Econometrica 89*(1), 133–161.

Callaway, B., A. Goodman-Bacon, and P. H. C. Sant'Anna (2025). Difference-in-differences with a continuous treatment. Working paper; conditionally accepted.

Callaway, B. and P. H. C. Sant'Anna (2021, dec). Difference-in-Differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Nie, X. and S. Wager (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika 108*(2), 299–319.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica 56*(4), 931–954. Publisher: [Wiley, Econometric Society].