The Effects of High-Stakes Teacher Evaluation on Office Disciplinary Referrals

David D. Liebowitz, Lorna Porter & Dylan Bragg

Background/Context: A rich theoretical and empirical debate presents contrasting views about the power of accountability in education.[1] In fact, much of the body of evidence on the causal effects of accountability policies on school outcomes finds either mixed or no effects.[2] In this paper, we study the effects of accountability pressures in the context of teacher responses to student behavioral infractions in the aftermath of high-stakes teacher evaluation reforms at the beginning of the 2010s. The overwhelming majority of U.S. states adopted high-stakes teacher evaluation policies between 2010 and 2016 with the goal of improving teachers' performance. In addition to incentives to improve instructional pedagogy, the introduction of high-stakes evaluation based on observations and test scores increased pressures to create respectful, undisrupted classroom learning environments. Unruly classrooms are easily observable for teachers' evaluators; more so than, for example, alignment of instruction to grade-level standards. Thus, one implicit goal of higher-stakes teacher evaluation policy is to encourage teachers to improve their classroom management practices to minimize disruptive behavior. However, teachers might also accomplish the goal of reducing disruptive behavior by imposing a lower floor of tolerance for misbehavior before removing a student from class.

Research Question: What is the causal effect of introducing high-stakes teacher evaluation policies on the rates at which teachers remove students from class for disciplinary reasons?

Sample/Setting: We study the effects of these teacher evaluation reforms in a sample of U.S. traditional public schools subject to state evaluation policies (see Figure 1). We form our measures of Office Disciplinary Referrals from counts of referrals at the grade-school-year level. Thus, our main analytic sample, includes 107,458 grade-school-year observations, nested in 20,135 school-year observations. These represent a total of 2,564 schools in 43 U.S. states from 2006 to 2018.

We draw our data from schools that use the School-Wide Information System data platform to assist in the implementation of a widely used behavior management system, Positive Behavioral Interventions and Supports (PBIS). Thus, our sample is clearly not random, and our results should be interpreted as generalizable only to schools with these characteristics. However, our sample represents a vast number of schools across most U.S. states. Further, the demographic characteristics of our sample broadly match national racial and family income enrollment patterns.

Program/Policy: We focus on the common accountability elements rather than on the intensity of accountability pressures across states and districts, which are largely endogenous. In almost all cases, teacher evaluation reforms entail adopting a common rubric for evaluating teachers' performance with multiple rating categories. All state reforms to teacher evaluation require that classroom observation of teaching practice be a part of a teacher's final rating, and in most cases these reforms establish a minimum frequency of classroom observations. In addition, many states

[1] Compare, for instance, Dee & Wyckoff (2015), Chiang (2009), Hanushek (2009), Jackson, Rockoff & Staiger (2014) and Macartney, McMillan & Petronijevic (2018) with Figlio (2006), Ladd & Lauen (2010), Rothstein (2015) and Strunk, Barret & Lincove (2017).

[2] See, among others, Brehm, Imberman & Lovenheim (2017), Chakrabarti (2014), Cullen, Koedel & Parsons (2019), Deming, Cohodes, Jennings, Jencks (2016), Eren (2019), Kraft, Brunner, Dougherty & Schwegman (2019), Macartney (2016), Özek (2012), Pope (2019), Reback, Rockoff & Schwartz (2014), Steinberg & Sartain (2015), Stecher et al. (2018). Deming and Figlio (2016) and Liebowitz (2019) summarize this nuanced literature.

require some or all teachers to be evaluated based on student-learning gains (Donaldson & Papay, 2015; Jacobs & Doherty, 2015; Steinberg & Donaldson, 2016; Winters & Cowen, 2013).

Analytic Approach: Our identification strategy relies on the differential timing across states of the introduction of teacher evaluation reform. We begin by estimating a non-parametric event study. We extend this approach into the full difference-in-differences framework where we impose a functional form that allows us to formally test the policy effects. We fit the following model:

$$ODR_{gjst} = \beta_1 EVAL_{st} + (\mathbf{X_{jt}})\theta + \Delta_g + \Gamma_j + \Pi_t + +\varepsilon_{st} \qquad (1)$$

In simplified form, this represents the per-500-student per-day rate of Office Disciplinary Referrals (ODR_{gjst}) for each grade-year observation in grade *g*, school *j*, state *s* and time *t*, regressed on the indicator $EVAL_{st}$ which takes the value of 1 if the observation is in a state that is in a year with a high-stakes evaluation system. β_1 is the causal parameter of interest. The two-way fixed effect model includes school- (gamma) and year- (pi) fixed effects and a vector (**X**) of school-level (*j*) background characteristics. We also relax the assumption of the standard difference-in-differences model of time-invariant treatment effects by adding a linear time trend. For the coefficient β_1 to be an unbiased estimand, we make three assumptions: (1) schools and grades in untreated states (and not-yet-treated states) provide a valid counterfactual for schools and grades in treated states; (2) there are no unobserved simultaneous shocks correlated with our outcomes and the introduction of high-stakes teacher evaluation reforms; and (3) the estimands for each grade and year are appropriately pooled to create the full sample Average Treatment Effect (ATE). We subject our main results to a battery of robustness checks to determine whether these assumptions hold.

To better understand the effects of accountability pressures, we examine differences between grade levels under greater and lesser accountability. We also examine the extent to which the implementation of effective disciplinary support strategies serves to moderate the effects of greater accountability. We pre-registered our analytic approach in the SREE Registry of Efficacy and Effectiveness Studies in Education (#1748).

Results: We find no causal effect of the implementation of high-stakes evaluation on rates of office disciplinary referrals. In Figure 1, we present the graphical results of our event study estimates, and in Table 1 we present the results of Equation (1). In our preferred estimates (Models II and V), we can confidently rule out ranges of effects greater than a decrease of 0.21 referrals or an increase of 0.04 referrals per-500 students, per day for classroom referrals and a decrease of 0.14 or an increase of 0.06 referrals per-500 students, per day for subjective-classroom referrals. These confidence intervals correspond to a 0.08 standard deviations (SD) decrease and a 0.02 SDs increase or a 0.08 SDs decrease and a 0.03 SDs increase for classroom and subjective referrals respectively. We find no evidence of heterogeneous effects in higher-accountability grades or in schools that improve their implementation of behavioral supports (see Tables 2 and 3). All estimates prove robust to a large set of DD robustness checks (see Figure 3).

Conclusions: Our findings contribute to the limited understanding of the effects of accountability policy inside the black-box of classroom practice. To those hoping for dramatic improvements in teaching practice as well as those concerned about unintended consequences of evaluation policy, our findings present another reminder of the loose-coupling between education policy and teacher behaviors.

References

Brehm, M., Imberman, S. A., & Lovenheim, M. F. (2017). Achievement effects of individual performance incentives in a teacher merit pay tournament. *Labour Economics, 44*, 133–150. https://doi.org/10.1016/j.labeco.2016.12.008

Chakrabarti, R. (2014). Incentives and responses under No Child Left Behind: Credible threats and the role of competition. *Journal of Public Economics, 110*, 124–146. https://doi.org/10.1016/J.JPUBECO.2013.08.005

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*(9–10), 1045–1057. https://doi.org/10.1016/J.JPUBECO.2009.06.002

Cullen, J. B., Koedel, C., & Parsons, E. (2019). The compositional effect of rigorous teacher evaluation on workforce quality. *Education Finance and Policy*, 1–85. https://doi.org/10.1162/edfp_a_00292

Dee, T. S., & Wyckoff, J. (2015). Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management, 34*(2), 267–297. https://doi.org/10.1002/pam.21818

Deming, D. J., Cohodes, S., Jennings, J., & Jencks, C. (2016). School accountability, postsecondary attainment, and earnings. *Review of Economics and Statistics, 98*(5), 848–862. https://doi.org/10.1162/REST_a_00598

Deming, D. J., & Figlio, D. (2016). Accountability in US education: Applying lessons from K–12 experience to higher education. *Journal of Economic Perspectives, 30*(3), 33–56. https://doi.org/10.1257/jep.30.3.33

Donaldson, M. L., & Papay, J. (2015). Teacher evaluation for accountability and development. In Helen Ladd & Margaret Goertz (Eds.), *Handbook of Research in Education Finance and Policy* (2nd ed., pp. 174–193). New York: Routledge.

Eren, O. (2019). Teacher incentives and student achievement: Evidence from an Advancement Program. *Journal of Policy Analysis and Management, 38*(4), 867–890. https://doi.org/10.1002/pam.22146

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics, 90*(4–5), 837–851. https://doi.org/10.1016/J.JPUBECO.2005.01.003

Hanushek, E. (2009). Teacher deselection. In D. D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession* (pp. 165–180). Washington, DC: Urban Institute Press.

Jackson, C. K., Rockoff, J. E., & Staiger, D. O. (2014). Teacher Effects and Teacher-Related Policies. *Annual Review of Economics, 6*(1), 801–825. https://doi.org/10.1146/annurev-economics-080213-040845

Jacobs, S., & Doherty, K. (2015). State of the States 2015: Evaluating Teaching, Leading and Learning.

Kraft, M. A., Brunner, E. J., Dougherty, S. M., & Schwegman, D. J. (2019). *Teacher evaluation reforms and the supply and quality of new teachers* (Brown University Working Paper).

Providence, RI. Retrieved from https://scholar.harvard.edu/files/mkraft/files/kraft_et_al._teacher_evaluation_-_updated_feb_2019.pdf

Ladd, H. F., & Lauen, D. L. (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management, 29*(3), 426–450. https://doi.org/10.1002/pam.20504

Liebowitz, D. D. (2019). *High-stakes teacher evaluation for accountability and growth: Should policy treat them as complements or substitutes?* (University of Oregon Working Paper).

Macartney, H. (2016). The Dynamic effects of educational accountability. *Journal of Labor Economics, 34*(1), 1–28. https://doi.org/10.1086/682333

Macartney, H., McMillan, R., & Petronijevic, U. (2018). *Teacher performance and accountability incentives* (NBER Working Paper Series No. No. 24747). Cambridge, MA.

Ozek, U. (2012). *One day too late? Mobile students in the era of accountability* (CALDER Working Paper Series No. No. 82). Washington, DC. Retrieved from https://caldercenter.org/sites/default/files/WP 82 Final.pdf

Pope, N. G. (2019). The effect of teacher ratings on teacher performance. *Journal of Public Economics, 172*, 84–110. https://doi.org/10.1016/J.JPUBECO.2019.01.001

Reback, R., Rockoff, J., & Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy, 6*(3), 207–241. https://doi.org/10.1257/pol.6.3.207

Rothstein, J. (2015). Teacher quality policy when supply matters. *American Economic Review, 105*(1), 100–130. https://doi.org/10.1257/aer.20121242

Stecher, B., Holtzman, D., Garet, M., Hamilton, L., Engberg, J., Steiner, E., … Chambers, J. (2018). *Improving Teaching Effectiveness: Final Report: The Intensive Partnerships for Effective Teaching Through 2015-2016*. Santa Monica: RAND Corporation. https://doi.org/10.7249/RR2242

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy, 11*(3), 340–359. https://doi.org/10.1162/EDFP_a_00186

Steinberg, M. P., & Sartain, L. (2015). Does Teacher Evaluation Improve School Performance? Experimental Evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy, 10*(4), 535–572. https://doi.org/10.1162/EDFP_a_00173

Strunk, K. O., Barrett, N., & Lincove, J. A. (2017). *When tenure ends: The short-run effects of the elimination of Louisiana's teacher employment protections on teacher exit and retirement*. Retrieved from https://educationresearchalliancenola.org/files/publications/041217-Strunk-Barrett-Lincove-When-Tenure-Ends.pdf

Winters, M. A., & Cowen, J. M. (2013). Would a value-added system of retention improve the distribution of teacher quality? A simulation of alternative policies. *Journal of Policy Analysis and Management, 32*(3), 634–654. https://doi.org/10.1002/pam.21705

Table 1. The effect of teacher evaluation reforms on Office Disciplinary Referrals, by location and subjectivity

	A. Class			B. Class, Subjective		
	I	II	III	IV	V	VI
Implement evaluation	-0.084	-0.086	-0.082	-0.041	-0.041	-0.054
	(0.063)	(0.063)	(0.072)	(0.049)	(0.050)	(0.043)
Implement evaluation * Trend			0.046			0.007
			(0.043)			(0.024)
Time trend			-0.017			0.004
			(0.032)			(0.022)
School composition controls		X	X		X	X
Grade-year observations (N)	107,458	107,458	107,458	107,458	107,458	107,458
School-year observations	20,135	20,135	20,135	20,135	20,135	20,135
R-squared	0.559	0.559	0.559	0.55	0.55	0.55

Notes: *p<.05, **p<.01, ***p<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. School controls include time-varying enrollment, proportion low-income and race/ethnicity. All models include grade, school and year fixed-effects and are weighted by grade enrollment.

Table 2. The effect of teacher evaluation reforms on Office Disciplinary Referrals, by grade-level accountability pressures, location and subjectivity

	A. Class (3-11 only)			B. Subjective (3-11 only)		
	I	II	III	IV	V	VI
Implement evaluation	-0.092	-0.096	-0.095	-0.052	-0.052	-0.074
	(0.066)	(0.067)	(0.076)	(0.056)	(0.057)	(0.044)
Implement evaluation * Trend			0.054			0.008
			(0.047)			(0.027)
Time trend			-0.017			0.009
			(0.033)			(0.024)
School composition controls		X	X		X	X
Grade-year observations (N)	64,431	64,431	64,431	64,431	64,431	64,431
School-year observations	19,630	19,630	19,630	19,630	19,630	19,630
R-squared	0.586	0.586	0.586	0.573	0.573	0.573

Notes: *p<.05, **p<.01, ***p<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. School controls include time-varying enrollment, proportion low-income and race/ethnicity. All models include grade, school and year fixed-effects and are weighted by grade enrollment.

Table 3. The moderating effect of Positive Behavioral Interventions and Supports (PBIS) on the effect of teacher evaluation reforms on Office Disciplinary Referrals, by location and subjectivity

	A. Classroom				B. Subjective			
	I	II	III	IV	V	VI	VII	VIII
Implement evaluation	-0.045	-0.083	-0.084	-0.245	-0.054	-0.075	-0.074	-0.194
	(0.071)	(0.108)	(0.109)	(0.198)	(0.067)	(0.083)	(0.085)	(0.109)
Implement PBIS well		-0.116	-0.117	-0.105		-0.086	-0.087	-0.081
		(0.063)	(0.063)	(0.063)		(0.045)	(0.045)	(0.045)
Implement evaluation * PBIS		0.036	0.036	0.171		0.018	0.018	0.117
		(0.097)	(0.097)	(0.188)		(0.059)	(0.059)	(0.100)
Implement evaluation * Trend				0.141				0.065
				(0.089)				(0.049)
Implement evaluation * Trend * PBIS				-0.088				-0.065
				(0.084)				(0.044)
Time trend				0.003				0.017
				(0.035)				(0.033)
School composition controls			X	X			X	X
Grade-year observations (N)	66,076	66,076	66,076	66,076	66,076	66,076	66,076	66,076
School-year observations	12,309	12,309	12,309	12,309	12,309	12,309	12,309	12,309
R-squared	0.602	0.602	0.602	0.602	0.584	0.584	0.584	0.585

Notes: *p<.05, **p<.01, ***p<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. School controls include time-varying enrollment, proportion low-income and race/ethnicity. All models include grade, school and year fixed-effects and are weighted by grade enrollment. Models I and V replicate results from main DD estimate on full sample. Fewer observations reflect subset of grade-year observations (61.5 percent) reporting PBIS implementation information.

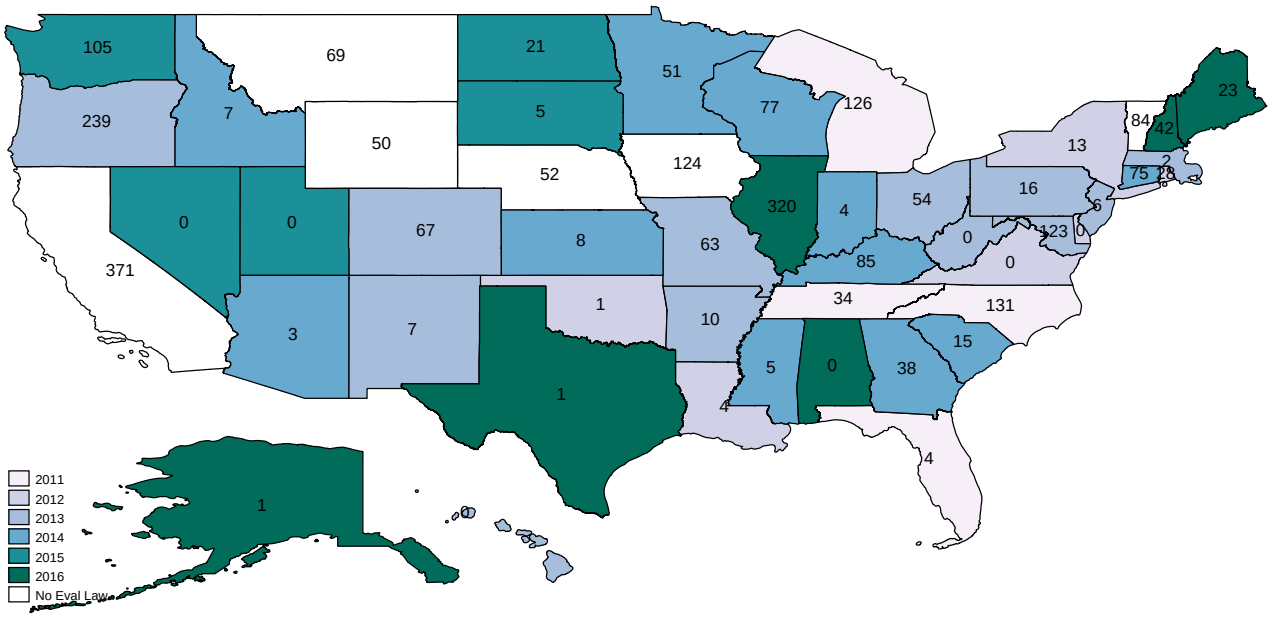
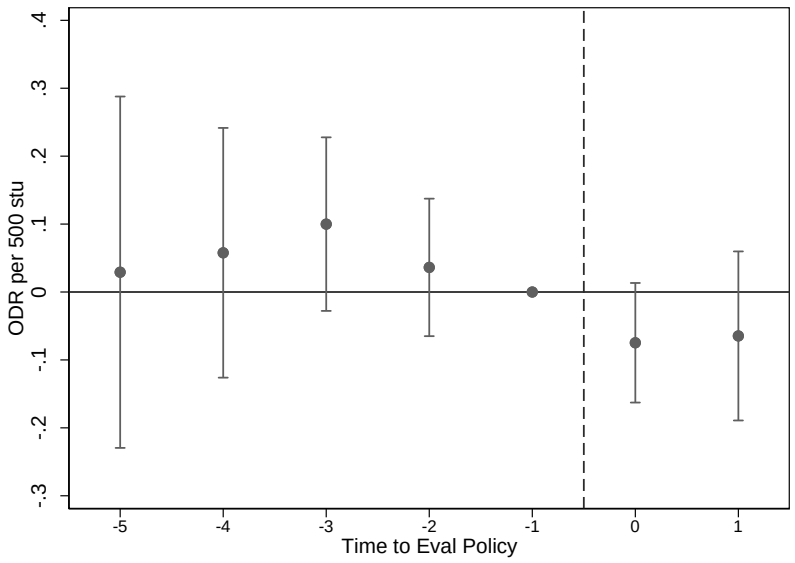
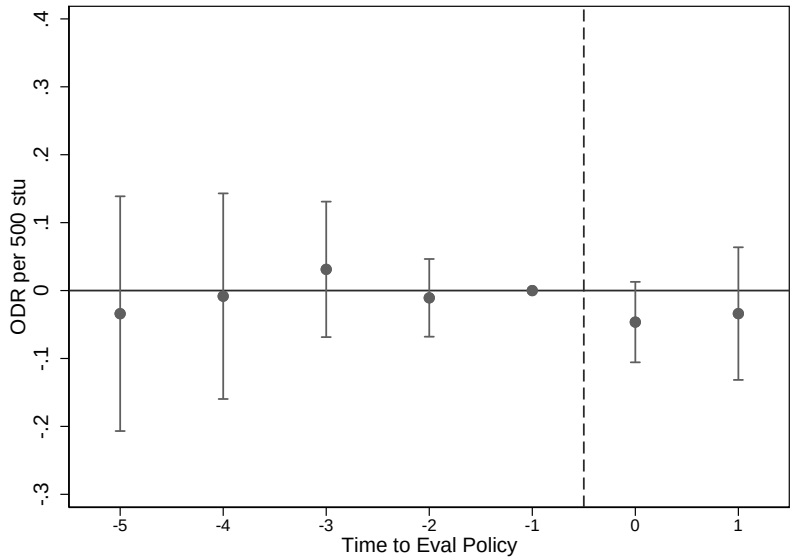


Figure 1. The timing of statewide teacher evaluation reforms and number of schools by state in analytic sample

Notes: the years above represent the fall of the academic year in which new evaluation systems were fully implemented statewide. Numbers inside each state represent total schools in analytic sample (n=2,564). Full list of states with schools in sample and timing of evaluation in Appendix Table A1.



Panel A. Classroom ODRs



Panel B. Subjective Classroom ODRs

Figure 2. Non-parametric event study displaying effect of high-stakes teacher evaluation reforms on rate of per-500-student, per-day Office Disciplinary Referrals (ODRs), by location and subjectivity

Notes: point estimates for years pre- and post-evaluation reforms and corresponding 95 percent confidence intervals derived from event study model describe in Equation 1 that is weighted by grade enrollment, includes grade, school and year fixed effects and time-varying school characteristics, with standard errors clustered at state level. Full coefficients reported in Columns IIa and IIc of Appendix Table A2.

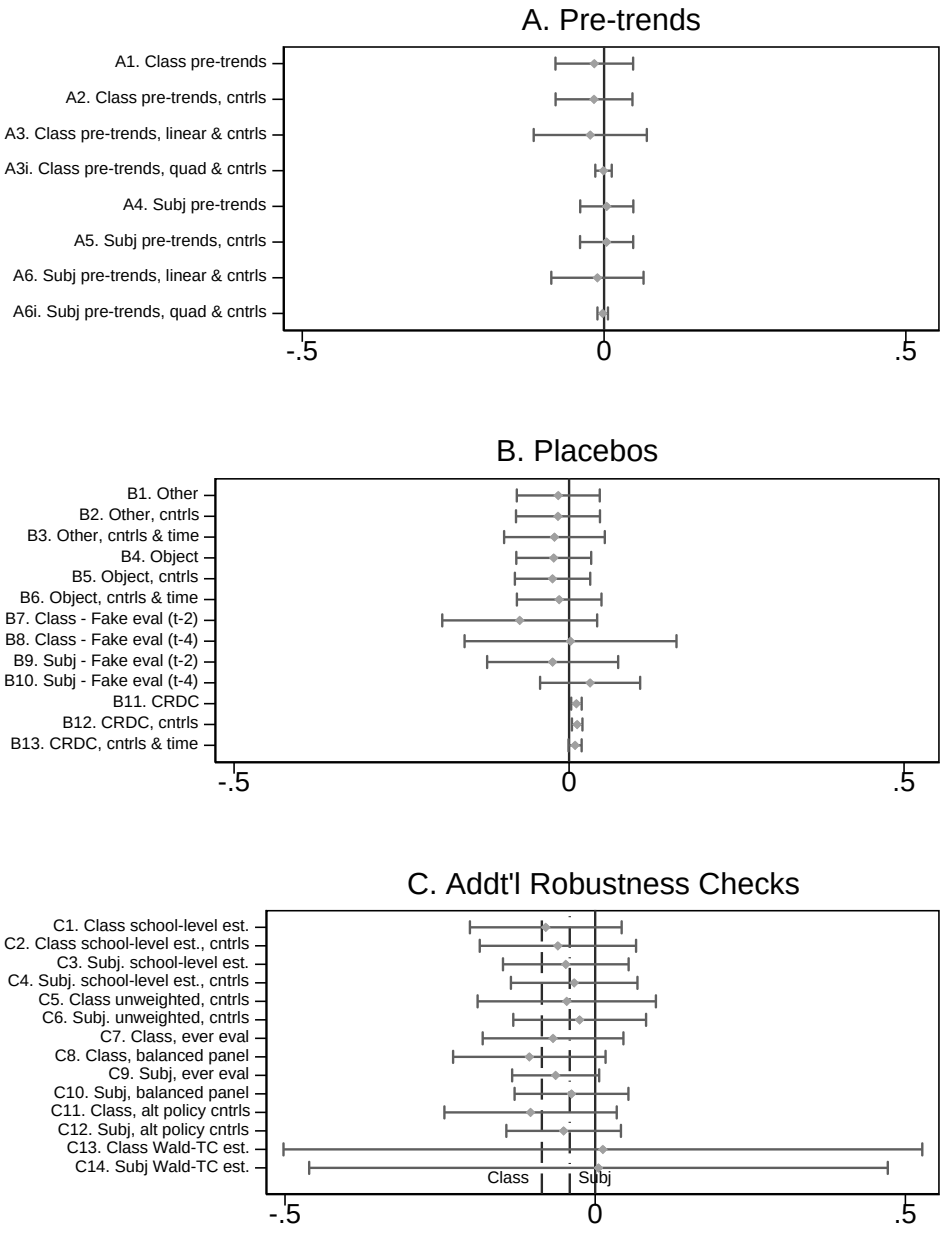


Figure 3. Tests of assumptions on all-grade difference-in-differences analysis.

Notes: Full set of point estimates available in Appendix Tables A4, A6, A7, A10 and A11, A13.