Appendix A. Additional Tables and Figures

Table A1. Education policy reforms by state, 2006-2018

	Schools in sample	Implement evaluation	Eliminate tenure	Weaken collective bargaining	Change teach. authority to remove stud. from class	Limit suspension/ exclusion
Alabama	0	2016				
Alaska	1	2016				
Arizona	3	2014				
Arkansas	10	2013				
California	371	None				2014
Colorado	67	2013				2012
Connecticut	75	2014			2018	2018
Delaware	0	2012				2018
Distr. of Columbia	0	2009			**2009**	**2009**; 2018
Florida	4	2011	**2011**			2009; 2018
Georgia	38	2014				**2014**
Hawaii	0	2013				2009
Idaho	7	2014	2011	2011		
Illinois	320	2016				**2016**
Indiana	4	2014			2009	
Iowa	124	None				
Kansas	8	2014	**2014**			
Kentucky	85	2014				
Louisiana	4	2012	**2012**		2009	2007; 2008; 2009; **2012**; 2015
Maine	23	2016				
Maryland	123	2013			2009	2014; 2017
Massachusetts	2	2013				
Michigan	126	2011				2017
Minnesota	51	2014			2016	
Mississippi	5	2014				
Missouri	63	2013				
Montana	69	None				
Nebraska	52	None				
Nevada	0	2015				**2015**
New Hampshire	42	2016				
New Jersey	6	2013			2012	2016
New Mexico	7	2013			2009	
New York	13	2012				
North Carolina	131	2011	2013			**2011**

North Dakota	21	2015			
Ohio	54	2013			2017; 2018
Oklahoma	1	2012			
Oregon	239	2013		2014	2014
Pennsylvania	16	2013			
Rhode Island	28	2012			2007; 2009; **2012**
South Carolina	15	2014			
South Dakota	5	2015			2014
Tennessee	34	2011	**2011**		2007; 2008; 2013; 2015; 2018
Texas	1	2016		2015	2011; 2017
Utah	0	2015			
Vermont	84	None			2011
Virginia	0	2012			2009; 2018
Washington	105	2015			2016
West Virginia	0	2013			2014
Wisconsin	77	2014	2011		
Wyoming	50	None			

Notes: Evaluation and teacher accountability policies drawn from Steinberg and Donaldson (2016), NCTQ (2016), and Kraft et al. (2020). Discipline policy changes draw from and Bezinque et al. (2018) from the National Center on Safe Supportive Learning Environments School Discipline Laws and Regulations Compendium database fields: "Teacher authority to remove students from classrooms" and "Limitations, conditions, or exclusions for use of suspension and expulsion." Covers 2006-2018. Time-varying implementation measures account for states in which policy was passed but never implemented. Years in bold are policies occurring in same year as teacher evaluation reform. Nine states reformed suspension/expulsion laws multiple times during this window. Main robustness checks use first year of reform, alternate tests use year closest to evaluation law.

Table A2. School characteristics and Office Disciplinary Referral (ODR) rates, by date of teacher evaluation implementation

	School Enrollment Characteristics								ODR Rates			
Implement evaluation in:	Avg. Enrollment	% Low Income	% AmInd/AK Native	% Asian/PI	% Black	% Hispanic	% White Non-Hispanic	% implement PBIS successfully	Class	Other	Subjective	Objective
2011-12	532.3	0.53	0.01	0.03	0.23	0.09	0.53	0.59	2.09	1.66	1.30	0.49
2012-13	462.7	0.42	0.01	0.04	0.07	0.10	0.73	0.47	3.11	1.55	1.94	0.45
2013-14	517.6	0.46	0.01	0.04	0.13	0.14	0.59	0.72	1.85	1.63	1.20	0.49
2014-15	574.0	0.56	0.01	0.04	0.15	0.12	0.60	0.70	2.35	1.62	1.51	0.45
2015-16	480.5	0.53	0.03	0.09	0.07	0.19	0.47	0.65	1.81	1.55	1.21	0.52
2016-17	439.3	0.46	0.00	0.04	0.19	0.15	0.53	0.69	2.77	1.45	1.65	0.44

Notes: Average school-level characteristics prior to implementing evaluation for schools in states ever implementing teacher evaluation reforms.

Table A3. Event study estimates of the effect of high-stakes teacher evaluation on ODRs, by location and subjectivity

	A. Class		B. Other		C. Subjective		D. Objective	
	Ia	IIa	Ib	IIb	Ic	IIc	Id	IId
-6 or more yrs pre	0.125	0.136	0.127	0.129	-0.013	-0.007	0.08	0.082
	(0.172)	(0.169)	(0.136)	(0.134)	(0.113)	(0.113)	(0.077)	(0.077)
5 yrs pre	0.028	0.029	0.109	0.109	-0.034	-0.033	0.028	0.029
	(0.135)	(0.132)	(0.094)	(0.093)	(0.089)	(0.089)	(0.064)	(0.063)
4 yrs pre	0.059	0.061	0.094	0.091	-0.006	-0.006	0.031	0.031
	(0.095)	(0.095)	(0.073)	(0.073)	(0.078)	(0.078)	(0.042)	(0.042)
3 yrs pre	0.1	0.104	0.06	0.058	0.033	0.034	0.028	0.029
	(0.066)	(0.066)	(0.040)	(0.040)	(0.051)	(0.052)	(0.032)	(0.033)
2 yrs pre	0.034	0.038	0.029	0.028	-0.01	-0.01	0.027	0.029
	(0.050)	(0.051)	(0.030)	(0.030)	(0.029)	(0.029)	(0.022)	(0.024)
1 yr pre	0	0	0	0	0	0	0	0
Evaluation introduced	-0.072	-0.076	-0.023	-0.025	-0.045	-0.047	-0.015	-0.016
	(0.045)	(0.046)	(0.020)	(0.020)	(0.031)	(0.031)	(0.022)	(0.022)
1 yr post	-0.063	-0.066	0.005	0.004	-0.033	-0.035	-0.018	-0.019
	(0.064)	(0.064)	(0.033)	(0.033)	(0.050)	(0.051)	(0.031)	(0.031)
2 yrs post	-0.027	-0.034	0.066	0.065	-0.016	-0.019	-0.011	-0.012
	(0.103)	(0.101)	(0.070)	(0.071)	(0.076)	(0.077)	(0.050)	(0.050)
3+ yrs post	0.007	0.002	0.186	0.186	-0.015	-0.018	0.007	0.006
	(0.181)	(0.179)	(0.116)	(0.117)	(0.124)	(0.125)	(0.079)	(0.078)
School composition controls		X		X		X		X
Grade-year observations (N)	107,468	107,468	107,468	107,468	107,468	107,468	107,468	107,468

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. Coefficients on 6 or more years pre, 2 yrs post and 3+ yrs post reported in table, but do not apply to all observations in sample due to differential timing; thus not reported in Figure 3.

Table A4. Event study estimates of the effect of high-stakes teacher evaluation on ODRs, by grade-level accountability pressures, location and subjectivity

Panel A. Classroom and Other Locations

	Class				Other			
	3-11		K-2, 12		3-11		K-2, 12	
-6 or more yrs pre	0.109	0.123	0.099	0.107	0.122	0.124	0.113	0.113
	(0.172)	(0.170)	(0.193)	(0.189)	(0.161)	(0.158)	(0.137)	(0.137)
5 yrs pre	0.042	0.045	-0.01	-0.011	0.125	0.125	0.074	0.072
	(0.147)	(0.146)	(0.133)	(0.130)	(0.095)	(0.093)	(0.105)	(0.104)
4 yrs pre	0.07	0.073	0.022	0.024	0.122	0.12	0.028	0.024
	(0.110)	(0.111)	(0.093)	(0.092)	(0.084)	(0.084)	(0.072)	(0.073)
3 yrs pre	0.106	0.11	0.083	0.085	0.064	0.062	0.052	0.05
	(0.071)	(0.071)	(0.076)	(0.075)	(0.040)	(0.041)	(0.060)	(0.060)
2 yrs pre	0.031	0.034	0.037	0.04	0.035	0.034	0.019	0.02
	(0.060)	(0.061)	(0.044)	(0.043)	(0.034)	(0.034)	(0.032)	(0.032)
1 yr pre	0	0	0	0	0	0	0	0
Evaluation introduced	-0.087	-0.093	-0.042	-0.045	-0.026	-0.028	-0.022	-0.023
	(0.054)	(0.055)	(0.045)	(0.045)	(0.021)	(0.021)	(0.033)	(0.033)
1 yr post	-0.061	-0.065	-0.075	-0.077	0.015	0.015	-0.019	-0.02
	(0.065)	(0.065)	(0.074)	(0.072)	(0.032)	(0.033)	(0.048)	(0.048)
2 yrs post	-0.021	-0.03	-0.049	-0.053	0.091	0.09	0.01	0.009
	(0.103)	(0.101)	(0.136)	(0.132)	(0.065)	(0.064)	(0.115)	(0.115)
3+ yrs post	0.019	0.011	-0.034	-0.036	0.213*	0.213*	0.119	0.121
	(0.179)	(0.177)	(0.203)	(0.198)	(0.102)	(0.103)	(0.181)	(0.181)
School composition controls		X		X		X		X
Grade-year observations (N)	64,437	64,437	43,013	43,013	64,437	64,437	43,013	43,013

Panel B. Subjective and Objective Reasons

	Subjective				Objective			
	3-11		K-2, 12		3-11		K-2, 12	
-6 or more yrs pre	-0.059	-0.052	0.043	0.047	0.098	0.101	0.028	0.03
	(0.112)	(0.113)	(0.118)	(0.116)	(0.090)	(0.090)	(0.064)	(0.063)
5 yrs pre	-0.04	-0.038	-0.031	-0.031	0.038	0.04	0.005	0.004
	(0.100)	(0.100)	(0.081)	(0.080)	(0.074)	(0.073)	(0.053)	(0.052)
4 yrs pre	-0.004	-0.004	-0.022	-0.022	0.032	0.033	0.025	0.025
	(0.095)	(0.096)	(0.058)	(0.057)	(0.046)	(0.045)	(0.043)	(0.042)
3 yrs pre	0.03	0.031	0.035	0.034	0.027	0.028	0.028	0.03
	(0.062)	(0.063)	(0.041)	(0.041)	(0.034)	(0.035)	(0.034)	(0.035)
2 yrs pre	-0.021	-0.021	0.008	0.008	0.034	0.036	0.012	0.014
	(0.036)	(0.036)	(0.024)	(0.023)	(0.026)	(0.028)	(0.020)	(0.021)

1 yr pre	0	0	0	0	0	0	0	0
Evaluation introduced	-0.06	-0.062	-0.017	-0.018	-0.008	-0.01	-0.027	-0.028
	(0.037)	(0.038)	(0.030)	(0.030)	(0.026)	(0.026)	(0.020)	(0.020)
1 yr post	-0.04	-0.041	-0.027	-0.028	-0.009	-0.01	-0.035	-0.036
	(0.054)	(0.056)	(0.047)	(0.047)	(0.035)	(0.034)	(0.029)	(0.029)
2 yrs post	-0.018	-0.022	-0.021	-0.024	-0.007	-0.009	-0.018	-0.019
	(0.080)	(0.082)	(0.086)	(0.084)	(0.057)	(0.057)	(0.049)	(0.049)
3+ yrs post	-0.011	-0.014	-0.039	-0.041	0.005	0.004	0.009	0.01
	(0.132)	(0.134)	(0.114)	(0.113)	(0.084)	(0.082)	(0.085)	(0.083)
School composition controls		X		X		X		X
Grade-year observations (N)	64,437	64,437	43,013	43,013	64,437	64,437	43,013	43,013

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. Coefficients on 6 or more years pre, 2 yrs post and 3+ yrs post reported in table, but do not apply to all observations in sample due to differential timing; thus not reported in Figure A2.

Table A5. The moderating effect of Positive Behavioral Interventions and Supports (PBIS) on the effect of teacher evaluation reforms on Office Disciplinary Referrals, by grade-level accountability pressures, location and subjectivity

	A. Class			B. Subjective		
	I	II	III	IV	V	VI
	3-11	3-11	K-2, 12	3-11	3-11	K-2, 12
Implement evaluation	-0.017	-0.084	-0.053	-0.044	-0.089	-0.029
	(0.087)	(0.133)	(0.104)	(0.086)	(0.106)	(0.076)
Implement PBIS well		-0.148	-0.01		-0.115	-0.003
		(0.082)	(0.043)		(0.058)	(0.029)
Implement evaluation * PBIS		0.070	-0.083		0.044	-0.063
		(0.130)	(0.088)		(0.082)	(0.065)
School composition controls	X	X	X	X	X	X
Grade-year observations (N)	39,640	39,640	26,405	39,640	39,640	26,405
School-year observations	12,020	12,020	9,521	12,020	12,020	9,521
R-squared	0.631	0.631	0.590	0.610	0.610	0.584

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-II and IV-V restricted to high-accountability grades (3-11). Models I and IV replicate results in the high accountability grades of the main effect of high-stakes evaluation in subset of schools reporting PBIS implementation information. Models II and V show moderating effect of PBIS implementation in grades 3-11. Models III and VI restricted to lower-accountability grades (K-2, 12). Fewer observations reflect subset of grade-year observations (61.5 percent) reporting PBIS implementation information.

Table A6. Parallel trends assumption checks, by location and subjectivity

	Class			Other		
	I	II	III	IV	V	VI
Panel A. Class and Other Locations						
Linear pre-trend	-0.016	-0.017	-0.024	-0.027	-0.027	-0.02
	(0.033)	(0.033)	(0.049)	(0.026)	(0.025)	(0.023)
Quadratic pre-trend			-0.001			0.001
			(0.007)			(0.004)
School composition controls		X	X		X	X
Grade-year observations (N)	107,468	107,468	107,468	107,468	107,468	107,468
School-year observations	20,137	20,137	20,137	20,137	20,137	20,137
R-squared	0.559	0.559	0.559	0.534	0.535	0.535

	Subjective			Objective		
Panel B. Subjective and Objective Reasons						
Linear pre-trend	0.004	0.004	-0.012	-0.011	-0.011	0.003
	(0.023)	(0.023)	(0.040)	(0.015)	(0.015)	(0.018)
Quadratic pre-trend			-0.002			0.002
			(0.005)			(0.002)
School composition controls		X	X		X	X
Grade-year observations (N)	107,468	107,468	107,468	107,468	107,468	107,468
School-year observations	20,137	20,137	20,137	20,137	20,137	20,137
R-squared	0.55	0.55	0.55	0.555	0.555	0.555

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 A7. Parallel trend assumption checks, by grade-level accountability pressures, location and subjectivity

	Class						Other					
	3-11			K-2, 12			3-11			K-2, 12		
Linear pre-trend	-0.017	-0.018	-0.042	-0.007	-0.008	0.001	-0.029	-0.03	-0.043	-0.019	-0.019	0.018
	(0.034)	(0.034)	(0.064)	(0.035)	(0.034)	(0.056)	(0.028)	(0.028)	(0.024)	(0.026)	(0.026)	(0.046)
Quadratic pre-trend			-0.004			0.001			-0.002			0.006
			(0.009)			(0.009)			(0.005)			(0.007)
School composition controls		X	X		X	X		X	X		X	X
Grade-year observations (N)	64,437	64,437	64,437	43,013	43,013	43,013	64,437	64,437	64,437	43,013	43,013	43,013
School-year observations	19,630	19,630	19,630	15,608	15,608	15,608	19,630	19,630	19,630	15,608	15,608	15,608
R-squared	0.586	0.586	0.586	0.546	0.546	0.546	0.56	0.561	0.561	0.553	0.553	0.554

	Subjective						Objective					
	3-11			K-2, 12			3-11			K-2, 12		
Linear pre-trend	0.009	0.008	-0.029	0.001	0.000	0.017	-0.013	-0.014	0.006	-0.004	-0.004	-0.009
	(0.024)	(0.025)	(0.050)	(0.021)	(0.021)	(0.036)	(0.018)	(0.017)	(0.020)	(0.013)	(0.012)	(0.023)
Quadratic pre-trend			-0.006			0.002			0.003			-0.001
			(0.005)			(0.006)			(0.003)			(0.003)
School composition controls		X	X		X	X		X	X		X	X
Grade-year observations (N)	64,437	64,437	64,437	43,013	43,013	43,013	64,437	64,437	64,437	43,013	43,013	43,013
School-year observations	19,630	19,630	19,630	15,608	15,608	15,608	19,630	19,630	19,630	15,608	15,608	15,608
R-squared	0.573	0.573	0.573	0.549	0.549	0.549	0.587	0.587	0.587	0.532	0.533	0.533

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 A8. Alternate outcome and evaluation implementation year placebo tests, by location and subjectivity

	I	II	III	IV	V
Panel A. Class and Other Locations		Other		Class (fake)	
Implement evaluation	-0.016	-0.017	-0.022		
	(0.032)	(0.032)	(0.039)		
Implement evaluation * Trend			0.096*		
			(0.043)		
False eval implementation (t-2)				-0.076	
				(0.060)	
False eval implementation (t-4)					0.004
					(0.081)
School composition controls		X	X	X	X
Grade-year observations (N)	107,468	107,468	107,468	107,468	107,468
School-year observations	20,137	20,137	20,137	20,137	20,137
R-squared	0.534	0.534	0.535	0.559	0.559
Panel B. Subjective and Objective Reasons					
		Objective		Subjective (fake)	
Implement evaluation	-0.023	-0.025	-0.015		
	(0.029)	(0.029)	(0.033)		
Implement evaluation * Trend			0.019		
			(0.027)		
False eval implementation (t-2)				-0.026	
				(0.051)	
False eval implementation (t-4)					0.032
					(0.038)
School composition controls		X	X	X	X
Grade-year observations (N)	107,468	107,468	107,468	107,468	107,468
School-year observations	20,137	20,137	20,137	20,137	20,137
R-squared	0.555	0.555	0.555	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 A9. CRDC difference-in-difference estimates of high-stakes teacher evaluation policies on out of school suspensions, 2005/06-2015/16

	A. Class	B. Subj	C. CRDC				
	I	II	III	IV	V	VI	VII
Implement evaluation	-0.038	-0.026	0.007	0.007	0.004	0.005	0.007
	(0.062)	(0.047)	(0.004)	(0.005)	(0.005)	(0.004)	(0.005)
Implement evaluation * Trend							-0.005
							(0.003)
Time trend							0.002**
							(0.001)
School composition controls	X	X		X	X	X	X
Grade-school-year observat.	88,401	88,401	NA	NA	NA	NA	NA
School-year observations	16,562	16,562	343,015	343,015	293,250	343,015	343,015
R-squared	0.585	0.571	0.717	0.718	0.720	0.751	0.719

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, Title I Status, and race/ethnicity. Models I and II re-estimate main results excluding years 2016-17 and 2017-18 to demonstrate comparability with time-frame of CRDC data. Models III-VII estimate rates of suspension in CRDC data. Model V restricted to schools in states that ever implemented high stakes evaluation reform. All CRDC models include school and year fixed-effects. Model I through V are weighted by school enrollment while Model VI is unweighted. Data obtained from Civil Rights Data Collection. All CRDC sample sizes (school-year observations) rounded to nearest 5 per IES requirements.

Table A10. Alternate outcome placebo tests, by high-accountability grade, location and subjectivity

	A. By Location & Grade			B. By Type & Grade		
	I Other 3-11	II Class K-2, 12	III Other K-2, 12	IV Obj 3-11	V Subj K-2, 12	VI Obj K-2, 12
Implement evaluation	-0.015	-0.078	-0.026	-0.018	-0.024	-0.039
	(0.031)	(0.064)	(0.042)	(0.036)	(0.039)	(0.024)
School composition controls	X	X	X	X	X	X
Grade-year observations (N)	64,437	43,013	43,013	64,437	43,013	43,013
School-year observations	19,630	15,608	15,608	19,630	15,608	15,608
R-squared	0.56	0.546	0.553	0.587	0.549	0.533

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 A11. Alternate evaluation year implementation placebo tests, by location and subjectivity (grades 3-11 only)

	I	II	III	IV	V	VI	VII	VIII
	Class	Other	Class	Other	Subj	Obj	Subj	Obj
False eval implementation (t-2)	-0.093	-0.098			-0.037	-0.012		
	(0.072)	(0.064)			(0.064)	(0.033)		
False eval implementation (t-4)			0.006	-0.064			0.051	-0.032
			(0.095)	(0.073)			(0.040)	(0.053)
School composition controls	X	X	X	X	X	X	X	X
Grade-year observations (N)	64,437	64,437	64,437	64,437	64,437	64,437	64,437	64,437
School-year observations	19,630	19,630	19,630	19,630	19,630	19,630	19,630	19,630
R-squared	0.586	0.56	0.586	0.56	0.573	0.587	0.573	0.587

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 A12. Alternate specifications testing robustness to alternate samples and weighting of average treatment effects

	I	II	III	IV	V	VI	VII	VIII
	School estimates			Wald-TC	Ever eval		Balanced panel	
Panel A. Class and Other Locations								
	Class				Class	Other	Class	Other
Implement evaluation	-0.080	-0.062	-0.067	0.011	-0.070	-0.020	-0.108	-0.040
	(0.066)	(0.068)	(0.076)	(0.259)	(0.059)	(0.027)	(0.064)	(0.030)
Implement evaluation * Time trend			0.043					
			(0.046)					
Time trend			-0.011					
			(0.034)					
School composition controls		X	X	X	X	X	X	X
Grade-year observations (N)	N/A	N/A	N/A	N/A	82,969	82,969	83,455	83,455
School-year observations	20,405	20,405	20,405	20,405	15,803	15,803	15,599	15,599
R-squared	0.706	0.709	0.709	N/A	0.548	0.517	0.594	0.570
Panel B. Subjective and Objective Reasons								
	Subjective				Subject	Object	Subject	Object
Implement evaluation	-0.048	-0.035	-0.053	0.004	-0.066	0.004	-0.039	-0.036
	(0.055)	(0.056)	(0.046)	(0.235)	(0.037)	(0.028)	(0.048)	(0.028)
Implement evaluation * Time trend			-0.001					
			(0.029)					
Time trend			0.010					
			(0.026)					
School composition controls		X	X	X	X	X	X	X
Grade-year observations (N)	N/A	N/A	N/A	N/A	82,969	82,969	83,455	83,455
School-year observations	20,405	20,405	20,405	20,405	15,803	15,803	15,599	15,599
R-squared	0.724	0.727	0.727	N/A	0.526	0.577	0.582	0.583

* 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. Models I-III estimated at school-level which includes 268 more school-year observations than main analytic sample. Models I-III include school and year fixed effects and are weighted by school enrollment. Models IV-VIII also include grade fixed effects and are weighted by grade enrollment. Models V and VI restricted to schools that experienced high-stakes evaluation. Models VII and VIII are a balanced panel restricted to school-year observations 5 years before through 1 year after introduction of high-stakes evaluation or never experienced it.

Table A13. Triple-difference estimates of the effects of teacher evaluation reforms on Office Disciplinary Referrals

	I	II	III
Panel A. Class and Other Locations			
Implement evaluation * classroom	-0.070	-0.072	-0.044
	(0.045)	(0.044)	(0.042)
Implement evaluation	-0.015	-0.017	-0.030
	(0.032)	(0.033)	(0.047)
Implement evaluation * classroom * Trend			-0.044
			(0.024)
Implement evaluation * Trend			0.093*
			(0.040)
Time trend			-0.022
			(0.026)
School composition controls		X	X
Grade-year observations (N)	107,468	107,468	107,468
School-year observations	20,137	20,137	20,137
R-squared	0.554	0.554	0.554
	IV	V	VI
Panel B. Subjective and Objective Reasons			
Implement evaluation * Subjective	-0.020	-0.021	-0.019
	(0.051)	(0.051)	(0.039)
Implement evaluation	-0.022	-0.023	-0.025
	(0.028)	(0.028)	(0.034)
Implement evaluation * Subjective * Trend			-0.003
			(0.029)
Implement evaluation * Trend			0.015
			(0.023)
Time trend			-0.003
			(0.015)
School composition controls		X	X
Grade-year observations (N)	107,468	107,468	107,468
School-year observations	20,137	20,137	20,137
R-squared	0.574	0.574	0.574

* 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, year, classroom/subjective, classroom/subjective-by-school and classroom/subjective-by-year fixed effects and are weighted by grade enrollment. Double-difference models available in Table 2 (class and subjective) and Table A8 (other and objective).

Table A14. Triple-difference estimates of the effects of teacher evaluation reforms on Office Disciplinary Referrals, grades 3-11 only

	I	II	III
Panel A. Class and Other Locations			
Implement evaluation * classroom	-0.079	-0.081	-0.052
	(0.059)	(0.059)	(0.055)
Implement evaluation	-0.014	-0.016	-0.034
	(0.031)	(0.031)	(0.048)
Implement evaluation * classroom * Trend			-0.046
			(0.028)
Implement evaluation * Trend			0.104*
			(0.045)
Time trend			-0.024
			(0.025)
School composition controls		X	X
Grade-year observations (N)	64,437	64,437	64,437
School-year observations	19,630	19,630	19,630
R-squared	0.585	0.586	0.586
	IV	V	VI
Panel B. Subjective and Objective Reasons			
Implement evaluation * Subjective	-0.038	-0.039	-0.041
	(0.065)	(0.065)	(0.050)
Implement evaluation	-0.015	-0.016	-0.019
	(0.035)	(0.035)	(0.039)
Implement evaluation * Subjective * Trend			0.003
			(0.038)
Implement evaluation * Trend			0.012
			(0.028)
Time trend			-0.003
			(0.015)
School composition controls		X	X
Grade-year observations (N)	64,437	64,437	64,437
School-year observations	20,137	20,137	20,137
R-squared	0.609	0.609	0.609

* 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, year, classroom/subjective, classroom/subjective-by-school and classroom/subjective-by-year fixed effects and are weighted by grade enrollment. Double-difference models available in Tables 3 (class and subjective, grades 3-11) and Table A10 (other and objective, grades 3-11)

Table A15. The effects of teacher evaluation, other accountability, and discipline policy reforms on ODRs, by location and subjectivity

	Classroom ODRs				Subjective ODRs			
	Separate Models	Joint Account.	Joint discipline	Full joint model	Separate Models	Joint Account.	Joint discipline	Full joint model
	I	II	III	IV	V	VI	VII	VIII
Implement evaluation	-0.089	-0.091	-0.101	-0.104	-0.042	-0.047	-0.045	-0.051
	(0.064)	(0.067)	(0.067)	(0.072)	(0.051)	(0.053)	(0.045)	(0.048)
Eliminate tenure	-0.315	-0.316		-0.306	-0.121	-0.124		-0.105
	(0.205)	(0.204)		(0.205)	(0.153)	(0.154)		(0.156)
Weaken collective bargaining	0.105	0.134		0.156	0.193	0.207		0.212
	(0.385)	(0.373)		(0.370)	(0.248)	(0.241)		(0.235)
Alter teacher authority to remove student from class	0.110		0.119	0.109	0.070		0.094	0.094
	(0.195)		(0.199)	(0.195)	(0.144)		(0.153)	(0.151)
Alter limits to suspension/expulsion	0.032		0.018	0.03	-0.029		-0.044	-0.037
	(0.095)		(0.077)	(0.078)	(0.065)		(0.049)	(0.052)
School composition controls	X	X	X	X	X	X	X	X
Grade-year observations (N)	107,468	107,468	107,468	107,468	107,468	107,468	107,468	107,468
School-year observations	20,137	20,137	20,137	20,137	20,137	20,137	20,137	20,137

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. Coefficients on Suspension Limit models that use the year of discipline reform closest to evaluation reform are 0.004 (0.094) and -0.041 (0.063) for class and subjective ODRs, respectively.

Table A16. The effects of teacher evaluation, other accountability and discipline policy on ODRs, by location and subjectivity (grades 3-11 only)

	Classroom ODRs (3-11 only)				Subjective ODRs (3-11 only)			
	Separate Models	Joint Account.	Joint discipline	Full joint model	Separate Models	Joint Account.	Joint discipline	Full joint model
	I	II	III	IV	V	VI	VII	VIII
Implement evaluation	-0.098	-0.100	-0.107	-0.109	-0.054	-0.058	-0.054	-0.059
	(0.068)	(0.072)	(0.072)	(0.076)	(0.059)	(0.061)	(0.049)	(0.052)
Eliminate tenure	-0.342	-0.341		-0.332	-0.13	-0.133		-0.108
	(0.251)	(0.251)		(0.255)	(0.189)	(0.191)		(0.196)
Weaken collective bargaining	0.077	0.109		0.126	0.189	0.206		0.203
	(0.446)	(0.431)		(0.429)	(0.284)	(0.275)		(0.269)
Alter teacher authority to remove student from class	0.085		0.100	0.089	0.061		0.099	0.099
	(0.194)		(0.205)	(0.201)	(0.154)		(0.171)	(0.169)
Alter limits to suspension/exclusion	0.014		0.005	0.017	-0.064		-0.079	-0.072
	(0.104)		(0.097)	(0.097)	(0.071)		(0.060)	(0.062)
School composition controls	X	X	X	X	X	X	X	X
Grade-year observations (N)	64,437	64,437	64,437	64,437	64,437	64,437	64,437	64,437
School-year observations	19,630	19,630	19,630	19,630	19,630	19,630	19,630	19,630

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. Coefficients on Suspension Limit models that use the year of discipline reform closest to evaluation reform are -0.025 (0.105) and -0.082 (0.069) for class and subjective ODRs, respectively.

Table A17. Unweighted OLS estimates of the effect of teacher evaluation reforms on Office Disciplinary Referrals, by location and subjectivity

	A. Class			B. Subjective		
	I	II	III	IV	V	VI
Implement evaluation	-0.045	-0.048	-0.032	-0.025	-0.026	-0.02
	(0.072)	(0.073)	(0.088)	(0.054)	(0.055)	(0.060)
Implement evaluation * Trend			0.038			0.006
			(0.039)			(0.026)
Time trend			-0.019			-0.005
			(0.032)			(0.024)
School composition controls		X	X		X	X
Grade-year observations (N)	107,468	107,468	107,468	107,468	107,468	107,468
School-year observations	20,137	20,137	20,137	20,137	20,137	20,137
R-squared	0.497	0.497	0.497	0.49	0.49	0.49

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.

Table A18. The effect of teacher evaluation reforms on Office Disciplinary Referrals using alternate standard error clustering strategies, by location and subjectivity

	Classroom ODRs				Subjective ODRs			
	State-year	Twoway cluster	State-year	Twoway cluster	State-year	Twoway cluster	State-year	Twoway cluster
	I	II	III	IV	V	VI	VII	VIII
Implement evaluation	-0.084	-0.084	-0.089	-0.089	-0.041	-0.041	-0.042	-0.042
	(0.048)	(0.064)	(0.048)	(0.069)	(0.031)	(0.050)	(0.031)	(0.051)
School composition controls			X	X			X	X
Grade-year observations (N)	107,468	107,468	107,468	107,468	107,468	107,468	107,468	107,468
School-year observations	20,137	20,137	20,137	20,137	20,137	20,137	20,137	20,137

Notes: *p<.05, **p<.01, ***p<.001. Cells report estimates and associated standard errors. Odd-numbered models cluster at level of state-year observation. Even-numbered models implement two-way cluster on state and year following Cameron, Gelbach and Miller (2011). 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 A19. Alternate specifications testing robustness to alternate samples, by location and subjectivity (grades 3-11 only)

	Class, 3-11		Subject, 3-11	
	I	II	III	IV
	Ever eval	Balanced panel	Ever eval	Balanced panel
Implement evaluation	-0.071	-0.121	-0.087*	-0.043
	(0.069)	(0.064)	(0.041)	(0.051)
School composition controls	X	X	X	X
Grade-year observations (N)	49,578	50,097	49,578	50,097
School-year observations	15,392	15,207	15,392	15,207
R-squared	0.583	0.621	0.558	0.603

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 A20. Alternate specification tests of 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

	Class		Subjective	
	Ever eval	Balanced Panel	Ever eval	Balanced Panel
Implement evaluation	-0.129	-0.076	-0.135*	-0.061
	(0.107)	(0.173)	(0.060)	(0.109)
Implement PBIS well	0.044	0.018	0.033	0.012
	(0.101)	(0.179)	(0.063)	(0.097)
Implement evaluation * PBIS	-0.122	-0.114	-0.099	-0.081
	(0.083)	(0.059)	(0.062)	(0.041)
School composition controls	X	X	X	X
Grade-year observations (N)	49,708	52,536	49,708	52,536
School-year observations	9,418	9,762	9,418	9,762
R-squared	0.594	0.623	0.558	0.608

* 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 and are weighted by grade enrollment. Ever eval are restricted to schools that experienced high-stakes evaluation. Balanced panel are restricted to school-year observations 5 years before through 1 year after introduction of high-stakes evaluation or never experienced it.

Table A21. Alternate specification tests of 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 (grades 3-11 only)

	Class, 3-11		Subj, 3-11	
	I	II	III	IV
	Ever eval	Balanced panel	Ever eval	Balanced panel
Implement evaluation	-0.110	-0.077	-0.159*	-0.075
	(0.136)	(0.185)	(0.077)	(0.116)
Implement PBIS well	0.054	0.048	0.047	0.047
	(0.136)	(0.201)	(0.088)	(0.110)
Implement evaluation * PBIS	-0.129	-0.142	-0.115	-0.108
	(0.108)	(0.079)	(0.080)	(0.054)
School composition controls	X	X	X	X
Grade-year observations (N)	29,670	31,547	29,670	31,547
School-year observations	9,195	9,535	9,195	9,535
R-squared	0.63	0.653	0.592	0.634

* 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 and are weighted by grade enrollment. Ever eval are restricted to schools that experienced high-stakes evaluation. Balanced panel are restricted to school-year observations 5 years before through 1 year after introduction of high-stakes evaluation or never experienced it.

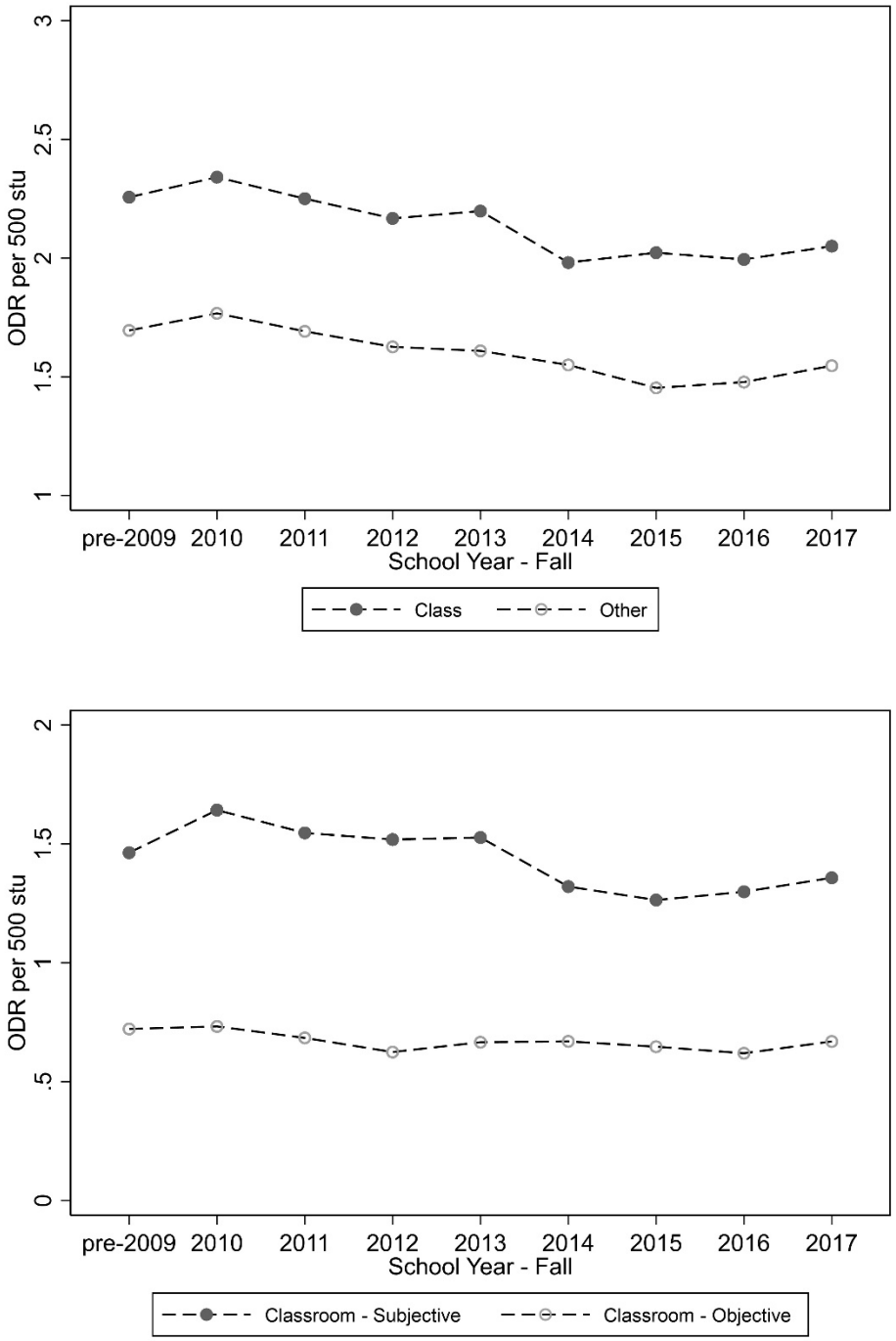
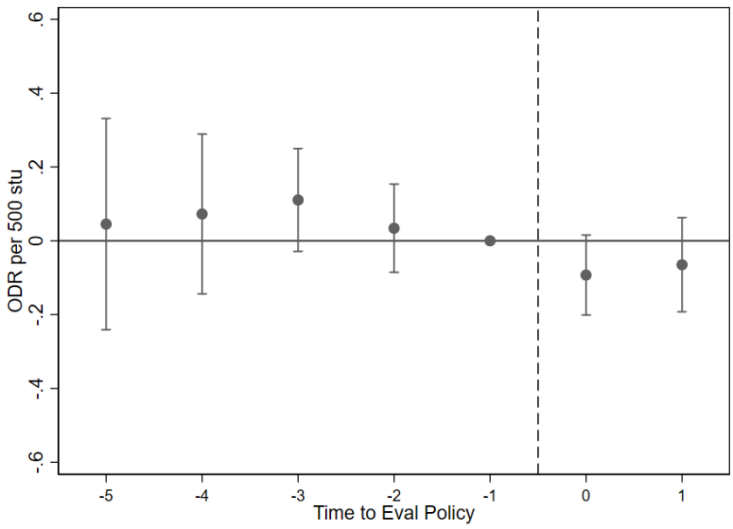
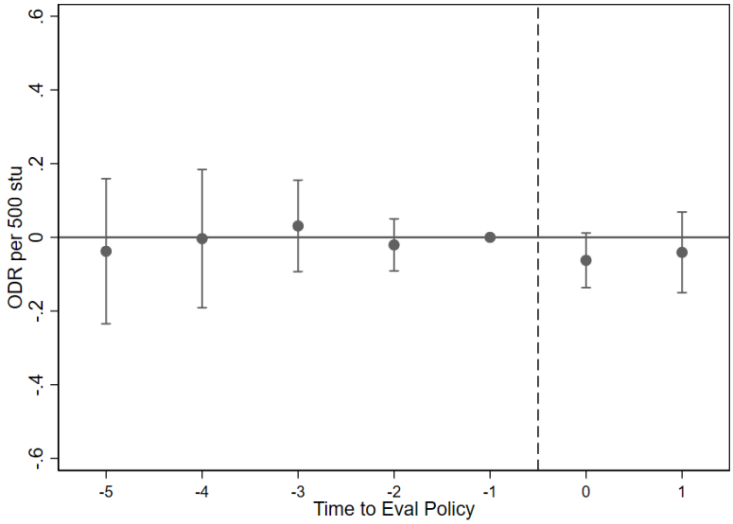


Figure A1. Office Disciplinary Referral (ODR) trends for states that never experienced evaluation reform, by location and type of referral



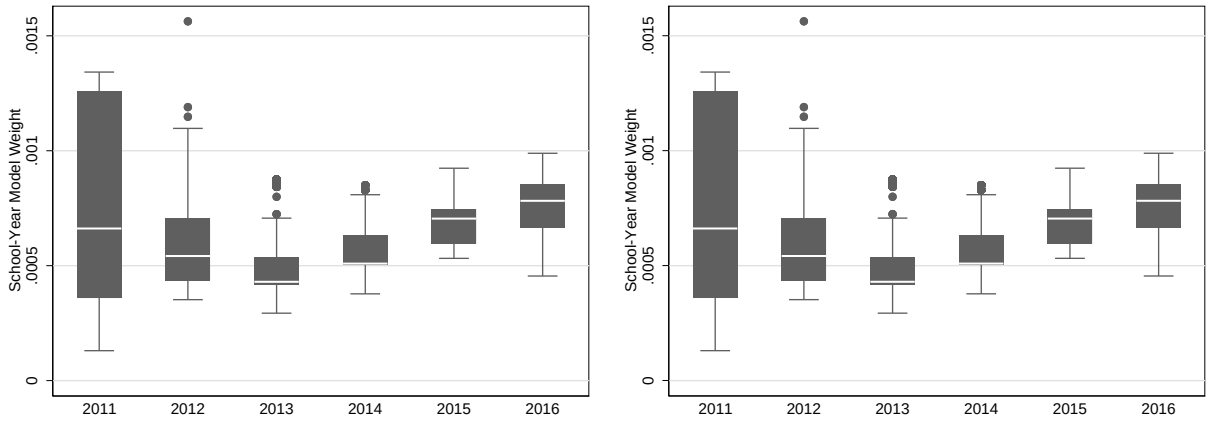
Panel A. Classroom ODRs



Panel B. Subjective Classroom ODRs

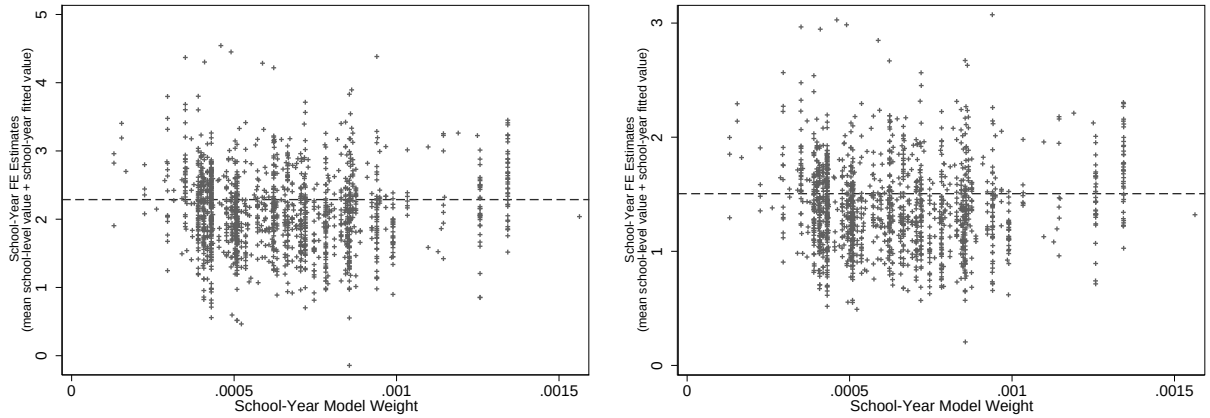
Figure A2. Non-parametric event study displaying effect of high-stakes teacher evaluation reforms on rate per-500-student, per-day Office Disciplinary Referrals (ODRs) in grades 3-11, by location and subjectivity

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 school size, includes grade, school and year fixed effects and time-varying school characteristics, with standard errors clustered at state level. Full coefficients reported in Models IIa and IIc of Appendix Table A4.



Classroom Subjective

Panel A. Fixed effect weights by year of teacher evaluation implementation

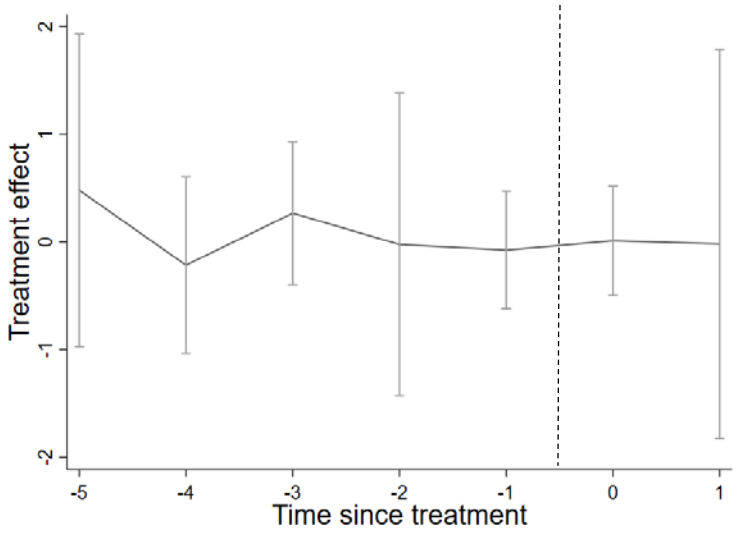


Classroom Subjective

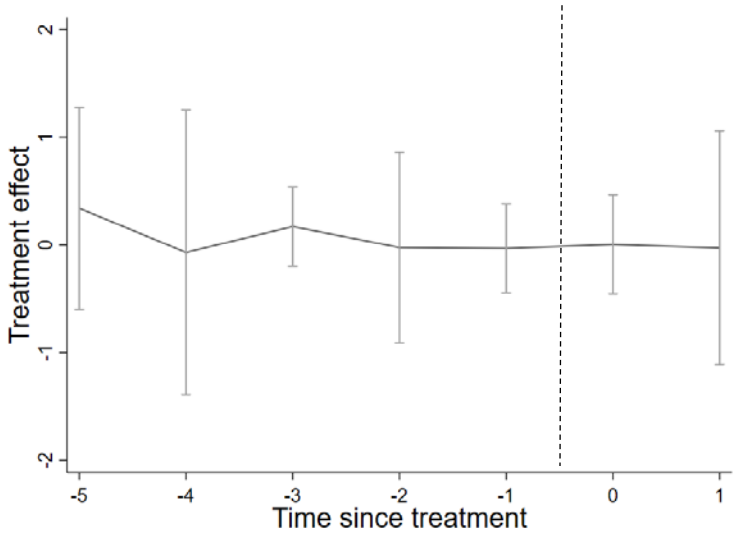
Panel B. School year fixed effect fitted values by fixed effect weight

Figure A3. Tests of difference-in-difference fixed effect weights variability by year and size of treatment effect

Notes: Dotted lines in Panel B are the unweighted, school-level sample fitted value of the effect of evaluation implementation, combining the mean ODR rate and the average treatment effect (2.29 for class and 1.51 for subjective)



Panel A. Class



Panel B. Subjective

Figure A4. Difference-in-difference estimates with Wald-TC estimator with prior-year placebos and post-implementation trend

Appendix B. Data Description

B.1 SWIS PBIS Sample and Data

The School-wide Information System (SWIS) data on Positive Behavioral Interventions and Supports come from the Education and Community Supports research unit of the University of Oregon (Bragg, 2019). A key strategy to improve behavioral supports in schools implementing PBIS is to track behavioral data. As such, each behavioral incident that prompts a student to be referred to an administrator responsible for addressing misbehavior is recorded. Approximately 25,000 schools in the 2016-17 school year were attempting to implement PBIS in some form. About 11,000 of these schools used the SWIS data management system and 5,745 of these schools agreed to have their data used for research purposes (Hoselton, 2018).

We begin by restricting our sample to observations with valid data that were subject to the policies of interest. Due to data inconsistencies in which schools sharing the same ID appear in different states, we drop 27 school-year observations. We then restrict our sample to school-years appearing in and after the 2006-07 school year. At this point, we have 75,066 school-year observations. We exclude all Alternative and Juvenile Justice schools as they have different governing regulations, resulting in dropping 5,105 school-year observations. We also drop 103 school-year observations that are exclusively pre-schools as are may not subject to the same evaluation policy. For similar reasons, we exclude Bureau of Indian Affairs (134 school-year observations), schools in Guam and the Virgin Islands (172 school-year observations), and charter schools (924 school-year observations).

We then require that we observe schools in states that experience evaluation reform for four years prior to the adoption of high-stakes teacher evaluation and one year following the initial policy implementation. For school in states that never experience evaluation, we require that we observe them four times between 2006 and 2017. This substantially reduces our sample to 23,538 school year observations.

The SWIS data include enrollment data from the school's NCES Common Core Data (CCD) record in most cases, but some schools do not report this October 1 count data and instead report a local tally of students. In the case of schools with highly transient populations, the October 1 count data in the CCD may represent enrollment of differences of 20 percent or more than a February 1 count. We attempt to capture the most complete enrollment data. We use the CCD enrollment data or the SWIS-reported enrollment when only one of the two is available. Otherwise, we use the average of the two. 1,512 school-year observations are either missing all enrollment data or have enrollment below 20 students and we exclude these.

We describe our imputations of race/ethnicity and low-income measures in the notes to Table 1. Even after these imputations, we still have 40 school-year observations without race/ethnicity data and 227 school-year observations without free- or reduced-price lunch (FRPL) data. We exclude these observations.

We then limit our sample to school-years which have recorded Office Disciplinary Referrals (ODRs) originating in the classroom or other locations, and are subjective or objective in nature. We restrict our sample to schools that have measured outcomes in all four of these areas. Excluding these school-years missing one or more outcomes results in a further reduction of our sample, resulting in a school-level sample that includes 20,405 school-year observations. This is our sample for our school-level estimates in Appendix Table A12.

From this school-level sample, we further identify our main grade-school-year analytic sample. We restrict from our sample any school-year observation that does not have an outcome recorded for all of the four primary outcomes recorded at the grade level. This results in the exclusion of an additional 384 school-year observations and results in a main sample of 20,137 school-years. Embedded in these school-year observations are 107,468 grade-school-year observations. We re-shape our data to use these grade-school-year observations as our primary analytic units.

Our outcome measures are the grade-school-year count of a particular category of ODR, divided by the total grade enrollment, divided by the number of school days in the year for that school. Then, we scale the outcome to the approximate average school size in our sample by multiplying this ratio by 500. As we discuss in the notes to Table 1, due to a small number of outlying values, we cap all outcome measures above the 99th percentile to the value of the 99th percentile.

We construct school-level measures of racial composition and the proportion of low-income students (measured by their receipt of free- or reduced-price lunch). These measures are constructed by dividing enrollment by race by total enrollment. All models adjusting for racial composition include the following racial/ethnic group percentages from the CCD and school reports: American Indian/Native Alaskan, Asian/Pacific Islander, Black, Hispanic and White, Non-Hispanic. As we discuss in the notes to Table 1, we observe a small number of school-year observations in which a particular demographic group represents more than the total school enrollment. These instances generally arise when total and sub-group enrollment figures differentially reflect in- and out-flows of students over the school year. We cap these values at 1. This affects 563 school-year observations.

We assign grade-school-year observations a value of 1 for PBIS implementation in years in which they meet self- and externally assessed criteria for successful implementation.

B.2 CRDC Data

As a placebo test of our main results, we estimate models using restricted-use Civil Rights Data Collection (CRDC) data set. CRDC data are collected biennially (for the most part) by the U.S. Department of Education, dating back to 1968. These data are primarily focused on civil rights issues such as discipline, bullying, and access for students with disabilities (U.S. Department of Education, 2018). In this study, we draw on five waves of data, from the 2005-2006 school year to the most recently-collected 2015-2016 school year. Within these waves of data, the U.S. Department of Education altered who was included in the sample. For the 2005-2006 and 2009-2010 waves, data were collected from a large, nationally representative sample of public schools and districts, while for the following three waves of data (2011-2012, 2013-2014, 2015-2016), data were collected from every public school and district in the country. These data do not include the last year of

evaluation implementation covering schools in six states that implemented evaluation reforms in 2016-17. Thus, we re-estimate our results in our main sample restricting observations to only those appearing prior to 2016-17. As these results are consistent, we feel confident concluding that the CRDC sample provides an informative check for our main estimates.

To prepare the data, we first merge the school-level out-of-school suspension counts, enrollment, demographic, and Title I status data from all five waves of data to create a longitudinal dataset. There are differences in how the CRDC collected demographic data across the different waves, moving from five distinct racial/ethnic categories to seven. In order to create stable categories across the waves, we create five categories, Asian-American/Hawaiian students, African-American students, Hispanic students, White non-Hispanic students, and an Other category that includes American Indian and Multi-racial students. We impute missing enrollment values using the average enrollment values from the prior and proceeding wave of data, or in the case of the initial wave the two waves after, and the final wave the two waves prior. We use the same imputation approach for demographic enrollment values.

We then calculate the suspension rate by dividing the total number of students suspended by the schools' total enrollment. Importantly, the out-of-school suspension rate measures the total number of students who were suspended throughout the year, not the total number of suspensions throughout the year. Therefore, the measure does not account for the additional number of suspensions that occurred if a student was suspended more than once in the school year. We do not include in-school suspensions nor expulsions in these analyses. We cap all suspension rates at 1, as there were rare cases in which schools reported a total number of students suspended that was larger than the school enrollment, which accounted for 1,435 observations. This has been previously documented in CRDC data (Losen & Gillespie, 2012), and may reflect that enrollment values are measured in Fall, while the count of students suspended is measured at the end of the school year, creating the opportunity to have more students suspended over the year's course than the initial Fall enrollment count.

The CRDC data collection does not have a measure of school or district percent of students who qualify for Free or Reduced Price Lunch, so as a measure of socioeconomic status we use whether or not a school qualified as a Title I school. Values are not available for the 2005-2006 school year, so we impute Title I status based on the following wave, or in the case of schools that did not have a 2009-2010 value, the 2011-2012 value. We also impute missing values for following waves, using the maximum (0/1) from the prior and proceeding waves.

Aligned with how we prepare the referral data, we exclude schools according to a number of criteria. For schools in states that did implement high-stakes teacher evaluation policies, we retain only schools for which we have an observation prior to, and after, the evaluation implementation. For schools in states that did not implement high-stakes teacher evaluation policies within the time span covered by our data, we retain only those with two or more observations. The above exclusion rules result in an exclusion of 38,180 observations, a product of factors such as schools closing or providing incomplete data across years. We further exclude Alternative, Juvenile Justice, and Charter schools, for a total of 22,870 observations excluded. We also exclude schools that offered only pre-K (another 4,350 observations), as well as schools with enrollment below 20 students

(another 4,565 observations). We also, after imputation, exclude three observations for which we are unable to impute total enrolment, another 6,930 observations for which we ae unable to impute Title I status, and 515 that have missing suspension rates. Note that the preceding description of the number of observations excluded will not align precisely with the total CRDC observations and those in our sample due to the fact that we round all reported values to the nearest 5 per Institute for Education Sciences requirements.

Table B1. CRDC descriptive statistics, 2005/06 – 2015/16

	Full Sample	Never Under Evaluation	Ever Under Evaluation
Total Schools	79,065	25,770	53,300
School-Year Observations	343,015	109,750	233,270
Total Districts	11,640	4,115	7,520
Average School Enrollment	595.21	615.74	585.55
(SD)	(450.48)	(508.30)	(420.19)
Racial/Ethnic Composition			
% Asian/Hawaiian	0.05	0.07	0.04
(SD)	(0.10)	(0.12)	(0.08)
% African American	0.16	0.11	0.18
(SD)	(0.22)	(0.18)	(0.24)
% Hispanic	0.24	0.42	0.15
(SD)	(0.28)	(0.32)	(0.20)
% White	0.52	0.37	0.59
(SD)	(0.32)	(0.31)	(0.30)
% Other	0.03	0.03	0.03
(SD)	(0.06)	(0.06)	(0.07)
Average Suspension Rate	0.06	0.06	0.06
(SD)	(0.09)	(0.08)	(0.09)

Notes: Standard deviations, where applicable, in parentheses. 15 school-year observations have total enrollment data imputed, 215 values were imputed for race/ethnicity, and 1,435 school-year observations have suspension rates capped at 1. Data obtained from Civil Rights Data Collection. All sample sizes (schools, school-year and district observations) rounded to nearest 5.

B.3 Teacher Evaluation and Accountability Policy Reform Data

We draw all data on teacher evaluation and accountability policies from Kraft et al. (2020) and refer our readers to their paper for details on this data collection process

B.4 Concurrent Discipline Policy Change Data

We compile data on concurrent discipline policy changes from the Compendium of School Discipline Laws and Regulations (Bezinque et al., 2018) on whether any state-level policies related

to *Teacher authority to remove students from the classroom* and *Limitations, conditions, or exclusions for use of suspension and expulsion* were enacted between 2006 and 2018. We identify all relevant statute and regulation for each of these two categories in the online compendium. We do not capture any changes to disciplinary statute or regulation not recorded in the Compendium:

https://safesupportivelearning.ed.gov/school-discipline-compendium

We then review each associated section of the state statute/regulation to identify any dates of revisions during the 2006-2018 window. Our default approach is to code any change in policy even minor ones; however, there are some instances when the language of the statute was revised to reflect the renaming of an agency or other minor shift. We exclude these from our reform tallies. We also exclude changes that focused exclusively on discipline policy for students with disabilities. In Table B2, we list the substance of these reforms with direct links to the statute. We code all of these policy changes based on the first fall of the school year under which the policy was implemented. There are six states that include schools in our sample that implemented multiple changes to limit suspension or expulsion (LA, MD, OH, RI, TN and TX). In our main robustness checks, we use the first observed policy change. We also test using the year closest to the implementation of evaluation reforms.

Table B2. Content of discipline policy reforms

	Tchr auth to remove	Limit suspension	Description (teacher authority)	Description (limit suspension/expulsion)
Alabama				
Alaska				
Arizona				
Arkansas				
California		2014		Limits on cause of suspension and age
Colorado		2012		Numerous changes to suspension reasons, disruptions/removal, and requiring that LEAs craft conduct and discipline codes
Connecticut	2018	2018	"Act Concerning Classroom Safety and Disruptive Behavior"	"Act Concerning Classroom Safety and Disruptive Behavior" focuses on reducing punitive/exclusionary discipline, which changed policy around both classroom removal and suspension/exclusion
Delaware		2018		Require LEA discipline reports and improvement plans (based on restorative justice) for schools under certain thresholds

Distr. of Columbia	**2009**	**2009**; 2018	Outline tiers of behavior for classroom discipline action and the accompanying discipline responses	2009: tiers include certain behaviors that cannot result in a suspension, such as absence ; 2018: limit length of suspension and who can be suspended
Florida		2009; 2018		2009: Revise zero tolerance policy to define out low-level offenses; 2018: provide alternatives to suspension and referral to law enforcement
Georgia		**2014**		Zero tolerance policy is only applicable to firearms, also clarifies the ability of local education boards to modify discipline policy for those who violate zero tolerance policy
Hawaii		2009		Interventions as alternatives to suspension required; limit on suspension due to truancy
Idaho				
Illinois		**2016**		Requires school officials to limit the number and duration of expulsions/suspensions, disallows zero-tolerance policies, and other requirements around how OSS may be used
Indiana	2009		Teacher Protection Act of 2009 had protections for teachers' disciplinary actions	
Iowa				
Kansas				Update to law was to reflect change in name from secretary of social and rehabilitation services to children and families
Kentucky				Changes to the suspension and expulsion policy defines what constitutes a threat, does not change grounds for suspension

Louisiana	2009	2007; 2008; 2009; **2012**; 2015	2009 permits principal to counsel alternatives to class removal; provide makeup work	Changes in suspension/exclusion policy from 2007-08 increase penalty for behavioral infractions. 2009: requires makeup work during suspension; 2012: requires alt education during suspension, adds provisions for bullying; 2015: prohibits suspension in K-5 for uniform violation;
Maine				
Maryland	2009	2014; 2017	Amend teachers' use of exclusion	2014: Require revisions of local student discipline policies to reflect a number of elements, including positive behavioral supports,; 2017: prohibit Prek-2nd grade suspensions/expulsions (w/specific exceptions)
Massachusetts				
Michigan		2017		"Rethink Discipline" law limits expulsion, requires consideration of alternatives to suspension, sets presumption that suspension longer than 10 days NOT justified
Minnesota	2016		Add in language that teachers may "may remove students from class under section 121A.61, subdivision 2, for violent or disruptive conduct."	
Mississippi				
Missouri				
Montana				Defines term of expulsion as 20+ days in 2009; requires annual review of policies in 2013, no substantive changes
Nebraska				
Nevada		**2015**		Outlines the circumstances under which students can be suspended/expelled for different firearm/weapon incidents
New Hampshire				Makes assignments available to students during period suspended

New Jersey	2012	2016	Harassment, intimidation, bullying grounds for removal from classroom	Limited suspension for K-2 students
New Mexico	2009		Revise procedure teachers go through for detention, suspension and expulsions.	
New York				
North Carolina		**2011**		Changes in who has authority to assign long-term suspensions and what services/opportunities are provided to those suspended
North Dakota				
Ohio		2017; 2018		2017: cannot be suspended for truancy; 2018: Limiting the use of out-of-school suspensions and expulsions for pre-K-third graders, money to support alternative discipline approaches
Oklahoma				
Oregon	2014	2014	Revise code to require LEAs to plan for reducing exclusionary discipline use	Major changes to discipline policy, focused on reducing suspensions and expulsions
Pennsylvania				
Rhode Island		2007; 2009; **2012**		2007: Changes to weapons/alcohol policy; 2009: adopt the "1.3 Safe, Healthy, and Supportive Learning Environment" policy ; 2012: Can no longer suspend students for truancy,
South Carolina				
South Dakota		2014		Adds phrase "No local school board may impose a lesser consequence than those established in § 13-32-9, but a local school district may adopt a policy (…) with more strict consequences to meet the needs of the district"

Tennessee		2007; 2008; 2013; 2015; 2018		2007 closed suspension hearing, fight=suspension, defines threat; 2008 discipline data reporting required; 2013 allows for self-defense, adjusts language around assault of staff leading to suspension; 2015 allows consequences for off-school ground behavior; 2018 specifies zero tolerance
Texas	2015	2011; 2017	Lists the campus behavior coordinator as a person to whom teachers can send students after student removal from classroom	2017: Limit grade of suspension for certain infractions, allowance for positive behavioral programs; 2011: outline what "serious misbehaviors" warrant expulsion
Utah				
Vermont		2011		Allows suspension for off-school events
Virginia		2009; 2018		2009: no suspension for truancy, 2018: limit suspensions for students grade 3 and below and outline time length limits on long-term suspensions
Washington		2016		Limits on length of suspension and use of suspension outside explicit circumstances
West Virginia		2014		Add section on weapon/substance possession procedures
Wisconsin				
Wyoming				