Appendix Table of Contents:

Appendix Tables and Figures

Appendix Table 1: The Effect of Treatment on Individual State Test Scores

	ITT			2SLS (Ever)			2SLS (Years)			2SLS (Trainings)		
	2015	2016	Pooled	2015	2016	Pooled	2015	2016	Pooled	2015	2016	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)	(13)
High-Stakes Math	0.059***	0.003	0.032***	0.059***	0.003	0.034***	0.064***	0.002	0.028***	0.094***	0.005	0.052***
	(0.008)	(0.008)	(0.006)	(0.008)	(0.010)	(0.007)	(0.009)	(0.006)	(0.005)	(0.013)	(0.014)	(0.010)
N	25,703	26,584	52,287	25,703	26,584	52,287	25,703	26,584	52,287	25,703	26,584	52,287
First stage coefficient				0.994***	0.843***	0.916***	0.916***	1.302***	1.112***	0.626***	0.592***	0.610***
				(0.001)	(0.003)	(0.002)	(0.001)	(0.006)	(0.003)	(0.002)	(0.001)	(0.001)
High-Stakes Reading	0.050***	0.030***	0.046***	0.050***	0.036***	0.050***	0.054***	0.023***	0.041***	0.077***	0.050***	0.073***
	(0.008)	(0.008)	(0.006)	(0.008)	(0.009)	(0.006)	(0.008)	(0.006)	(0.005)	(0.012)	(0.013)	(0.009)
N	34,794	35,512	70,306	34,794	35,512	70,306	34,794	35,512	70,306	34,794	35,512	70,306
First stage coefficient				0.994***	0.847***	0.919***	0.917***	1.329***	1.125***	0.646***	0.608***	0.628***
				(0.001)	(0.003)	(0.002)	(0.001)	(0.005)	(0.003)	(0.002)	(0.001)	(0.001)

Notes: This table reports ITT estimates of the effects of our management experiment in Houston on student achievement on state-mandated tests. Treatment is assigned in the same way in Table 4. Samples are limited to students with a valid outcome subject test score; unlike Table 4, students are not required to have a valid test score in every subject in order to enter the sample. Testing variables are drawn from district test score files and are standardized to have a mean of zero and a standard deviation of one within each year and grade among students with valid test scores. Columns (1)-(3) report Intent-to-Treat (ITT) estimates in each year. Columns (4)-(6) report 2SLS estimates that use treatment assignment to instrument for ever having attended a treatment school. Columns (7)-(9) report 2SLS estimates that use treatment assignment to instrument for the number of years spent in a treatment school. Columns (10)-(12) report 2SLS estimates that use treatment assignment to instrument for the percent of our management training sessions attended by a student's school principal, measured by attendance sheets at each training over the summer of 2014 and 2015. All specifications control for 3 years of baseline math and reading scores and their squares, indicators for whether the baseline test was taken in Spanish, indicators for whether baseline scores are from high- or low-stakes exams, and matched pair fixed effects. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Appendix Table 2: The Effect of Treatment on Individual ITBS Test Scores

	ITT	2SLS (Ever)	2SLS (Years)	2SLS (Trainings)
	(1)	(2)	(3)	(4)
Math	0.046***	0.046***	0.050***	0.076***
	(0.008)	(0.008)	(0.009)	(0.014)
N	24,208	24,208	24,208	24,208
First stage coefficient		0.995***	0.926***	0.605***
		(0.001)	(0.001)	(0.002)
Reading	0.041***	0.041***	0.044***	0.068***
	(0.008)	(0.008)	(0.009)	(0.013)
N	24,348	24,348	24,348	24,348
First stage coefficient		0.995***	0.926***	0.606***
		(0.001)	(0.001)	(0.002)
Science	0.048***	0.048***	0.051***	0.078***
	(0.009)	(0.009)	(0.010)	(0.015)
N	24,442	24,442	24,442	24,442
First stage coefficient		0.995***	0.926***	0.606***
		(0.001)	(0.001)	(0.002)
Social Studies	0.042***	0.043***	0.046***	0.070***
	(0.009)	(0.009)	(0.010)	(0.015)
N	24,434	24,434	24,434	24,434
First stage coefficient		0.995***	0.926***	0.605***
		(0.001)	(0.001)	(0.002)

Notes: This table reports ITT estimates of the effects of our management experiment in Houston on student achievement on the Iowa Test of Basic Skills (ITBS), a nationally-normed low-stakes test. The sample includes all students enrolled in grades 1-8 in one of the 40 experimental elementary and middle schools at the beginning of the 2014-15 school year. HISD stopped administering low-stakes exams after the 2014-15 school year. Treatment is assigned as the first school attended in 2014-15. Samples are limited to students with a valid outcome subject test score; unlike Table 4, students are not required to have a valid test score in every subject in order to enter the sample. Testing variables are drawn from district test score files and are standardized to have a mean of zero and a standard deviation of one within each year and grade among students with valid test scores. Column (1) reports Intent-to-Treat (ITT) estimates with treatment assigned based on the first school attended in 2014-15. Column (2) reports 2SLS estimates that use treatment assignment to instrument for ever having attended a treatment school. Column (3) reports 2SLS estimates that use treatment assignment to instrument for the number of years spent in a treatment school. Column (4) reports 2SLS estimates that use treatment assignment to instrument for the percent of our management training sessions attended by a student's school principal, measured by attendance sheets at each training over the summer of 2014. All specifications control for 3 years of baseline math and reading scores and their squares, indicators for whether the baseline test was taken in Spanish, indicators for whether baseline scores are from high- or low-stakes exams, and matched pair fixed effects. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Appendix Table 3: Actual High/Low Implementation and Actual Returning/New Principal Treatment Effects on High-Stakes Tests

	Full Sample		Actual Returning Principal		Actual New Principal	
	2015	2016	2015	2016	2015	2016
	(1)	(2)	(3)	(4)	(5)	(6)
Full Sample	0.101***	0.020	0.142***	0.117***	0.011	-0.136***
	(0.014)	(0.015)	(0.017)	(0.018)	(0.023)	(0.024)
N	25,397	26,379	17,210	17,102	8,187	9,277
Above-Med. Actual Impl. Index	0.187***	0.044**	0.202***	0.130***	0.120***	-0.182***
	(0.020)	(0.022)	(0.023)	(0.025)	(0.039)	(0.041)
N	13,493	13,427	10,325	10,047	3,168	3,380
Below-Med. Actual Impl. Index	0.015	-0.009	0.065**	0.089***	-0.045	-0.103***
	(0.020)	(0.020)	(0.026)	(0.027)	(0.028)	(0.029)
N	11,904	12,952	6,885	7,055	5,019	5,897
Above-Med. Actual Pct. Trainings	0.224***	0.123***	0.286***	0.243***	-0.003	-0.177***
	(0.022)	(0.023)	(0.026)	(0.028)	(0.042)	(0.041)
N	10,725	11,468	8,464	8,643	2,261	2,825
Below-Med. Actual Pct. Trainings	0.019	-0.060***	0.018	-0.028	0.010	-0.103***
	(0.017)	(0.019)	(0.022)	(0.025)	(0.027)	(0.029)
N	14,672	14,911	8,746	8,459	5,926	6,452

Notes: This table reports ITT estimates of the effects of our management experiment in Houston on student achievement on high-stakes test scores for high- and low-implementing principals who either stayed for the second year of the experiment or left. The sample is the same as in Table 4. Testing variables are drawn from district test score files and are standardized to have a mean of zero and a standard deviation of one within each year and grade among students with valid test scores. High-stakes tests are the State of Texas Assessments of Academic Readiness (STAAR) exams in math and reading (administered in grades 3-12). The dependent variable is the sum of standardized math and reading scores. Columns (1)-(2) report ITT estimates of the effect of treatment for the full sample. Columns (3)-(4) report ITT estimates of the effect of treatment for principals who returned to their schools in the second year of treatment, and Columns (5)-(6) report ITT estimates in schools with principal turnover between the two years of treatment. The rows further limit the sample by a school's fidelity of implementation. For example, Row (2), Columns (3)-(4) contain the ITT estimates for schools that are are high-implementers and have the same principal in both years of the treatment. For details on all variables used to subset the sample, see the Online Appendix. All specifications control for 3 years of baseline math and reading scores and their squares, indicators for whether the baseline test was taken in Spanish, indicators for whether baseline scores are from high- or low-stakes exams, and matched pair fixed effects. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Appendix Table 4: Subsample Analysis by Implementation Level, Alternative Cutoffs

	High-Stakes			Low-Stakes		
	50th pctile (Main)	40th pctile	60th pctile	50th pctile (Main)	40th pctile	60th pctile
	(1)	(2)	(3)	(4)	(5)	(6)
High Implementation Index (Actual)	0.120***	0.106***	0.204***	0.283***	0.179***	0.419***
	(0.015)	(0.013)	(0.017)	(0.039)	(0.035)	(0.045)
Low Implementation Index (Actual)	-0.001	-0.008	-0.032**	0.044	0.190***	0.010
	(0.014)	(0.016)	(0.013)	(0.042)	(0.048)	(0.037)
High Implementation Index (Predicted)	0.180***	0.161***	0.235***	0.258***	0.315***	0.347***
	(0.015)	(0.013)	(0.016)	(0.039)	(0.035)	(0.043)
Low Implementation Index (Predicted)	-0.054***	-0.103***	-0.055***	0.096**	-0.108**	0.048
	(0.014)	(0.016)	(0.013)	(0.041)	(0.049)	(0.038)
High Percent Trainings (Actual)	0.178***	0.141***	0.315***	0.545***	0.448***	0.751***
	(0.016)	(0.013)	(0.020)	(0.047)	(0.039)	(0.060)
Low Percent Trainings (Actual)	-0.022*	-0.037**	-0.048***	-0.024	-0.066	-0.031
	(0.013)	(0.016)	(0.012)	(0.035)	(0.041)	(0.032)
High Percent Trainings (Predicted)	0.142***	0.140***	0.189***	0.390***	0.344***	0.499***
	(0.014)	(0.013)	(0.016)	(0.041)	(0.037)	(0.047)
Low Percent Trainings (Predicted)	-0.037***	-0.078***	-0.031**	0.005	-0.000	0.001
	(0.014)	(0.017)	(0.013)	(0.038)	(0.043)	(0.035)
Principal Returns (Predicted)	0.113***	0.073***	0.007	0.380***	0.258***	0.185***
	(0.014)	(0.012)	(0.015)	(0.039)	(0.035)	(0.045)
Principal Leaves (Predicted)	-0.011	0.018	0.101***	-0.014	0.025	0.183***
	(0.015)	(0.018)	(0.014)	(0.040)	(0.048)	(0.036)

Notes: This table reports ITT estimates of the average yearly effects of our management experiment in Houston on student achievement on high- and low-stakes test scores for subgroups of the sample based on the fidelity of implementation of our management training. Columns (1) and (4) present the main results where high implementation refers to above-median. Columns (2) and (5) present results where high implementation refers to scores above the 40th percentile and Columns (3) and (6) present results where high implementation refers to scores above the 60th percentile. High-stakes tests are the State of Texas Assessments of Academic Readiness (STAAR) exams in math and reading (administered in grades 3-12), and low-stakes tests are the Iowa Test of Basic Skills (ITBS) exams in math, reading, science, and social studies (administered in grades 1-8). Each dependent variable is the sum of standardized scores in all subjects administered (i.e. 2 subjects in STAAR and 4 subjects in ITBS). Specifications and samples in Columns (1)-(3) are analogous to the pooled spcification in Column (3) of Table 4, and specifications and samples in Columns (3)-(6) are analogous to the first year specification in Column (1) of Table 4. Testing variables are drawn from district test score files and are standardized to have a mean of zero and a standard deviation of one within each year and grade among students with valid test scores. All variables used to partition the sample into subgroups are defined in the Online Appendix. All specifications control for 3 years of baseline math and reading scores and their squares, indicators for whether the baseline test was taken in Spanish, indicators for whether baseline scores are from high- or low-stakes exams, and matched pair fixed effects. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Appendix Table 5: The Effect of Treatment on Attrition for Administrative Testing Outcomes

	2015		2016		Pooled	
	Control Mean	Treatment Effect	Control Mean	Treatment Effect	Control Mean	Treatment Effect
	(1)	(2)	(3)	(4)	(5)	(6)
Full Sample						
Missing Math STAAR Score	0.129	-0.003	0.132	0.002	0.131	-0.001
		(0.004)		(0.004)		(0.003)
Missing Reading STAAR Score	0.075	0.002	0.081	0.002	0.078	0.002
		(0.003)		(0.003)		(0.002)
Took Math STAAR-L	0.037	-0.005***	0.038	-0.013***	0.038	-0.009***
		(0.002)		(0.002)		(0.001)
Elementary Schools						
Missing Math STAAR Score	0.088	0.015**	0.096	-0.003	0.092	0.006
		(0.007)		(0.008)		(0.005)
Missing Reading STAAR Score	0.089	0.013*	0.096	-0.003	0.093	0.005
		(0.007)		(0.008)		(0.005)
Took Math STAAR-L	0.008	0.020***	0.004	0.013***	0.006	0.016***
		(0.003)		(0.003)		(0.002)
Middle Schools						
Missing Math STAAR Score	0.088	0.002	0.091	0.007	0.090	0.004
		(0.004)		(0.004)		(0.003)
Missing Reading STAAR Score	0.086	0.002	0.091	0.006	0.088	0.004
		(0.004)		(0.004)		(0.003)
Took Math STAAR-L	0.043	-0.013***	0.044	-0.019***	0.044	-0.016***
		(0.003)		(0.003)		(0.002)
High Schools						
Missing Math STAAR Score	0.261	-0.041***	0.258	-0.009	0.259	-0.025***
		(0.010)		(0.012)		(0.008)
Missing Reading STAAR Score	0.037	-0.007	0.041	-0.008	0.039	-0.009**
		(0.005)		(0.006)		(0.004)
Took Math STAAR-L	0.044	0.001	0.050	-0.015**	0.047	-0.007*
		(0.006)		(0.006)		(0.004)

Notes: This table reports ITT estimates of the effects of our management experiment in Houston on whether a student is missing the test scores used in the main analysis. A student can exit the sample in one of two ways; by missing the exam entirely or by taking the STAAR L exam offered in math to students with limited English proficiency. The dependent variables are dummy variables that are a one if the student exited the sample by not taking a standard STAAR exam and zero otherwise. Students are considered eligible to take end-of grade exams if they were enrolled in grades 3-8. Students are considered eligible to take end-of-course math and reading exams if they were enrolled in Algebra I, or English I or II, respectively and were enrolled in grades 9-12. The sample is defined analogously as in Table 4 but students are no longer required to have valid test scores. Columns (1), (3), and (5) report the means of the control group. Columns (2), (4), and (6) report Intent-to-Treat (ITT) estimates of the effects of treatment on attrition. All specifications control for 3 years of baseline math and reading scores and their squares, indicators for whether the baseline test was taken in Spanish, indicators for whether baseline scores are from high- or low-stakes exams, and matched pair fixed effects. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Appendix Table 6: Cost Benefit Analysis

	Cost per student per year (Treatment)	Cost per student per year (Control)	Gains per year (Sum 2 Subjects)	Sample Size	IRR (%)
	(1)	(2)	(3)	(4)	(5)
Panel A: Studies from Fryer 2016					
Early Childhood					
Head Start Impact Study	$ 9,985.96	$ 3,146.91	0.323	4,700	9.19
Charter Schools					
Injecting Best Practices ES	$ 360.76	$ 0.00	0.256	75,500	35.13
Injecting Best Practices SS	$ 1,866.80	$ 0.00	0.134	95,400	18.41
Harlem Children's Zone ES	$ 21,266.15	$ 13,730.53	0.305	700	10.84
Harlem Children's Zone MS	$ 21,266.15	$ 13,730.53	0.276	1,400	11.92
SEED Schools	$ 43,184.66	$ 22,566.03	0.419	200	8.64
Teacher Incentives					
Talent Transfer Initiative	$ 656.18	$ 0.00	0.235	3,200	27.80
Teacher Certification					
Teach For America	$ 3,706.74	$ 0.00	0.180	1,800	11.73
Class Size					
Tenessee STAR	$ 5,084.74	$ 0.00	0.240	11,600	9.76
Managed Professional Development					
Success for All	$ 852.38	$ 0.00	0.090	2,100	14.15
Tutoring					
Experience Corps	$ 869.58	$ 0.00	0.075	900	13.81
Curriculum					
Enhanced Reading Opportunities	$ 2,096.42	$ 0.00	0.180	2,200	22.04
Financial Incentives					
Coshocton Incentive Program	$ 76.23	$ 0.00	0.118	900	49.14
New York, Dallas, Chicago	$ 356.94	$ 0.00	0.000	26,900	-
Panel B: Management Experiement					
Overall	$ 9.61	$ 0.35	0.060	51,800	79.47
Predicted High Implementation	$ 9.61	$ 0.35	0.120	26,500	95.59
Predicted Principal Staying	$ 9.61	$ 0.35	0.113	23,700	94.12

Notes: This table presents a summary of the costs, treatment effects, and calculated internal rates of return (IRRs) for our management experiment and 14 other major education interventions as summarized in Fryer (2016). In Panel A, we have included the experiments that mark major education policy interventions for which we could find reliable cost estimates. In Panel B, we present summaries for our overall experiment, the subset of matched pairs with predicted high implementation levels, and the subset of matched pairs where principals are predicted to return in the second year of the experiment. All specifications are identical to the pooled specification in Column (3) of Table 4. The IRR is the discount rate that sets the cost of each intervention equal to the discounted stream of future income benefits associated with increased student achievement. For additional details on the calculation for each experiment, please see the Online Appendix.

Appendix Table 7: The Effect of Treatment on State Test Scores; Retakes and Highest Scores

	Precedence to First Score (Main)			Highest Score			Precedence to Retakes		
	2015	2016	Pooled	2015	2016	Pooled	2015	2016	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Sum High Stakes	0.101***	0.020	0.060***	0.109***	0.018	0.062***	0.109***	0.020	0.064***
	(0.014)	(0.015)	(0.010)	(0.014)	(0.015)	(0.010)	(0.014)	(0.015)	(0.010)
N	25,397	26,379	51,776	25,397	26,379	51,776	25,397	26,379	51,776
Math	0.059***	0.003	0.032***	0.059***	-0.000	0.030***	0.059***	0.001	0.031***
	(0.008)	(0.008)	(0.006)	(0.008)	(0.008)	(0.006)	(0.008)	(0.008)	(0.006)
N	25,703	26,584	52,287	25,703	26,584	52,287	25,703	26,584	52,287
Reading	0.050***	0.030***	0.046***	0.056***	0.031***	0.049***	0.059***	0.033***	0.051***
	(0.008)	(0.008)	(0.006)	(0.008)	(0.008)	(0.006)	(0.008)	(0.008)	(0.006)
N	34,794	35,512	70,306	34,794	35,512	70,306	34,794	35,512	70,306

Notes: This table reports ITT estimates of the effects of our management experiment in Houston on student achievement on state-mandated tests. The sample and specifications are identical to those in Columns (1)-(3) of Table 4. In Columns (1)-(3), precedence is given to the first non-missing score. Remaining duplicates are then dealt with using the procedure outlined in the Online Appendix. In Columns (4)-(6), we use the highest score attained by each student (of duplicate test score entires, on-time tests, and retakes). In Columns (7)-(9), precedence is given to retake scores and remaining duplicates are then dealt with using the same procedure as in (1)-(3). All specifications control for 3 years of baseline math and reading scores and their squares, indicators for whether the baseline test was taken in Spanish, indicators for whether baseline scores are from high- or low-stakes exams, and matched pair fixed effects. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Appendix Table 8: Pre-Treatment Summary Statistics, New to HISD in 2015-16

	Non-Exp Mean	Exp Mean	*p-value*	Control Mean	Treatment Mean	*p-value*
	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.469	0.475	0.756	0.493	0.455	0.179
Black	0.350	0.406	0.213	0.370	0.444	0.347
Hispanic	0.520	0.545	0.563	0.587	0.502	0.269
White	0.062	0.024	0.000	0.026	0.022	0.794
Asian	0.053	0.013	0.000	0.013	0.013	0.970
Other Race	0.015	0.011	0.458	0.004	0.018	0.184
Limited English Proficient	0.323	0.228	0.004	0.266	0.189	0.124
Special Education Services	0.069	0.099	0.050	0.100	0.098	0.959
Gifted and Talented	0.034	0.025	0.298	0.021	0.029	0.559
Ecnomically Disadvantaged	0.716	0.709	0.863	0.713	0.706	0.906
Number of Students	2454	906		460	446	
p-value from joint F-test			*0.000*			*0.105*

Notes: This table reports student and school-level pre-treatment summary statistics for our management experiment for students who entered the district in the second year of treatment. Students are only included in the sample if they have at least one valid outcome test score variable in 2015-16. Column (1) reports the mean of the non-experimental group. Column (2) reports the mean of the experimental group. Column (3) reports the p-value on the null hypothesis of equal means in the experimental and non-experimental groups. Similarly, Columns (4)-(6) report the mean of the control and treatment groups and the p-value on the null hypothesis of euqal means in the treatment and control groups, respectively. The tests in Columns (3) and (6) use standard errors clustered at the school level. All demographic measures are culled from administrative data collected pre-treatment.

Appendix Table 9: The Effect of Treatment on State Test Proficiency Levels

	ITT			2SLS (Ever)			2SLS (Years)			2SLS (Trainings)		
	2015	2016	Pooled	2015	2016	Pooled	2015	2016	Pooled	2015	2016	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)	(13)
Panel A: Math												
Satisfactory (Phased)	0.024***	0.001	0.012***	0.024***	0.001	0.014***	0.026***	0.001	0.011***	0.038***	0.002	0.020***
	(0.006)	(0.006)	(0.004)	(0.006)	(0.007)	(0.004)	(0.006)	(0.004)	(0.004)	(0.009)	(0.009)	(0.007)
N	24,538	25,523	50,061	24,538	25,523	50,061	24,538	25,523	50,061	24,538	25,523	50,061
Control Mean	*0.598*	*0.617*	*0.608*									
Satisfactory (Recom.)	0.028***	0.003	0.015***	0.028***	0.003	0.016***	0.030***	0.002	0.013***	0.044***	0.005	0.024***
	(0.005)	(0.005)	(0.003)	(0.005)	(0.006)	(0.004)	(0.005)	(0.004)	(0.003)	(0.007)	(0.008)	(0.006)
N	24,538	25,523	50,061	24,538	25,523	50,061	24,538	25,523	50,061	24,538	25,523	50,061
Control Mean	*0.222*	*0.284*	*0.253*									
Advanced	0.006**	0.005	0.005**	0.006**	0.006	0.006**	0.007**	0.004	0.005**	0.010**	0.008	0.008**
	(0.003)	(0.003)	(0.002)	(0.003)	(0.004)	(0.002)	(0.003)	(0.003)	(0.002)	(0.005)	(0.006)	(0.004)
N	24,538	25,523	50,061	24,538	25,523	50,061	24,538	25,523	50,061	24,538	25,523	50,061
Control Mean	*0.064*	*0.095*	*0.080*									
Panel B: Reading												
Satisfactory (Phased)	0.025***	0.012**	0.022***	0.025***	0.014**	0.024***	0.027***	0.009**	0.020***	0.038***	0.020**	0.035***
	(0.005)	(0.005)	(0.003)	(0.005)	(0.006)	(0.004)	(0.005)	(0.004)	(0.003)	(0.007)	(0.008)	(0.005)
N	34,745	35,512	70,257	34,745	35,512	70,257	34,745	35,512	70,257	34,745	35,512	70,257
Control Mean	*0.576*	*0.586*	*0.581*									
Satisfactory (Recom.)	0.033***	0.010**	0.025***	0.033***	0.012**	0.027***	0.036***	0.008**	0.022***	0.051***	0.017**	0.039***
	(0.004)	(0.004)	(0.003)	(0.004)	(0.005)	(0.003)	(0.005)	(0.003)	(0.003)	(0.006)	(0.007)	(0.005)
N	34,745	35,512	70,257	34,745	35,512	70,257	34,745	35,512	70,257	34,745	35,512	70,257
Control Mean	*0.303*	*0.337*	*0.320*									
Advanced	0.007**	-0.003	0.003	0.007**	-0.003	0.003	0.008**	-0.002	0.003	0.011**	-0.005	0.005
	(0.003)	(0.003)	(0.002)	(0.003)	(0.004)	(0.002)	(0.003)	(0.002)	(0.002)	(0.004)	(0.005)	(0.003)
N	34,745	35,512	70,257	34,745	35,512	70,257	34,745	35,512	70,257	34,745	35,512	70,257
Control Mean	*0.085*	*0.105*	*0.095*									

Notes: This table reports ITT estimates of the effects of our management experiment in Houston on student proficiency levels on state-mandated tests. Treatment is assigned the same way as in Table 4, and students are included in the sample if they have a valid outcome variable (subject proficiency level). Proficiency level variables are drawn from district test score files. New STAAR minimum performance standards are being phased in from 2015-2021; the first variable in each panel is an indicator for meeting the phased satisfactory level in each year, the second variable in each panel is an indicator for meeting the recommended satisfactory level that will be in place in 2021, and the third variable in each panel is an indicator for achieving an advanced score in each subject (already phased in). Columns (1)-(3) report Intent-to-Treat (ITT) estimates in each year. Columns (4)-(6) report 2SLS estimates that use treatment assignment to instrument for ever having attended a treatment school. Columns (7)-(9) report 2SLS estimates that use treatment assignment to instrument for the number of years spent in a treatment school. Columns (10)-(12) report 2SLS estimates that use treatment assignment to instrument for the percent of our management training sessions attended by a student's school principal, measured by attendance sheets at each training over the summer of 2014 and 2015. All specifications control for 3 years of baseline math and reading scores and their squares, indicators for whether the baseline test was taken in Spanish, indicators for whether baseline scores are from high- or low-stakes exams, and matched pair fixed effects. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Appendix Table 10: The Effect of Treatment on Student Test Scores (ITT) by School Level

	Baseline Regressions			Fully Controlled Regressions		
	2015	2016	Pooled	2015	2016	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
Elementary Schools						
High Stakes (Sum 2 Subjects)	-0.028	-0.089***	-0.057**	-0.038	-0.114***	-0.075***
	(0.029)	(0.034)	(0.023)	(0.028)	(0.033)	(0.022)
N	5,003	5,086	10,089	5,003	5,086	10,089
Low Stakes (Sum 4 Subjects)	0.079	—	—	0.041	—	—
	(0.051)			(0.048)		
N	8,547			8,547		
Middle Schools						
High Stakes (Sum 2 Subjects)	0.138***	0.035**	0.085***	0.101***	-0.004	0.047***
	(0.016)	(0.017)	(0.012)	(0.015)	(0.017)	(0.011)
N	15,482	16,044	31,526	15,482	16,044	31,526
Low Stakes (Sum 4 Subjects)	0.247***	—	—	0.149***	—	—
	(0.033)			(0.031)		
N	15,331			15,331		
High Schools						
High Stakes (Sum 2 Subjects)	0.037	0.099***	0.073***	-0.018	0.091***	0.046**
	(0.032)	(0.035)	(0.024)	(0.030)	(0.032)	(0.022)
N	4,912	5,249	10,161	4,912	5,249	10,161

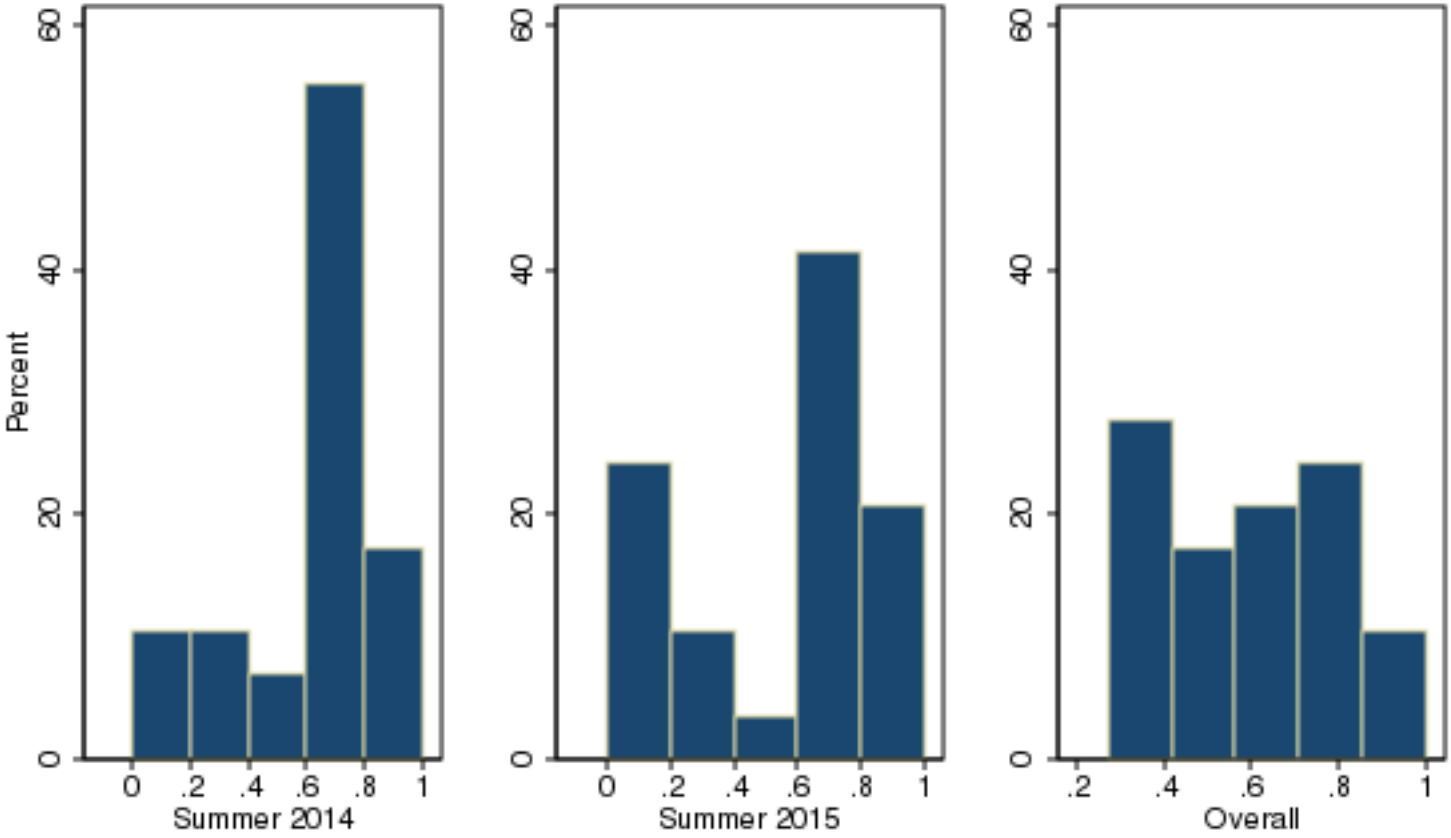
Notes: This table reports ITT estimates of the effects of our management experiment in Houston on student achievement on high- and low-stakes test scores for each school level. Samples are the same as in Table 4. Testing variables are drawn from district test score files and are standardized to have a mean of zero and a standard deviation of one within each year and grade among students with valid test scores. High-stakes tests are the State of Texas Assessments of Academic Readiness (STAAR) exams in math and reading (administered in grades 3-12), and low-stakes tests are the Iowa Test of Basic Skills (ITBS) exams in math, reading, science, and social studies (administered in grades 1-8). Each dependent variable is the sum of standardized scores in all subjects administered (i.e. 2 subjects in STAAR and 4 subjects in ITBS). Colunms (1)-(3) report ITT estimates of the effect of treatment, controlling only for matched-pair fixed effects and three years of baseline reading and math scores and their squares, indicators for whether the baseline test was taken in Spanish, and indicators for whether baseline scores are from high- or low-stakes exams. Columns (4)-(6) report ITT estimates of the effect of treatment, controlling for matched-pair fixed effects, grade-year fixed effects, and the student-level demographics summarized in Table 2 plus three years of baseline reading and math scores and their squares, indicators for whether the baseline test was taken in Spanish, and indicators for whether baseline scores are from high- or low-stakes exams. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Appendix Table 11: Subsample Analysis by Principal Characteristics, Subset Implementation Index

	High-Stakes	*p-value* on group diff	Low-Stakes	*p-value* on group diff
	(1)	(2)	(3)	(4)
Above-Median Predicted Number of Teacher Obs	0.220***		0.368***	
	(0.015)		(0.039)	
Below-Median Predicted Number of Teacher Obs	-0.099***	*0.000*	-0.070*	*0.000*
	(0.014)		(0.041)	
Above-Median Predicted Data Action Plan Submission Rate	0.156***		0.269***	
	(0.014)		(0.037)	
Below-Median Predicted Data Action Plan Submission Rate	-0.054***	*0.000*	0.068	*0.000*
	(0.015)		(0.044)	
Above-Median Predicted Lesson Plan Submission Rate	0.143***		0.255***	
	(0.014)		(0.037)	
Below-Median Predicted Lesson Plan Submission Rate	-0.036***	*0.000*	0.049	*0.000*
	(0.014)		(0.044)	

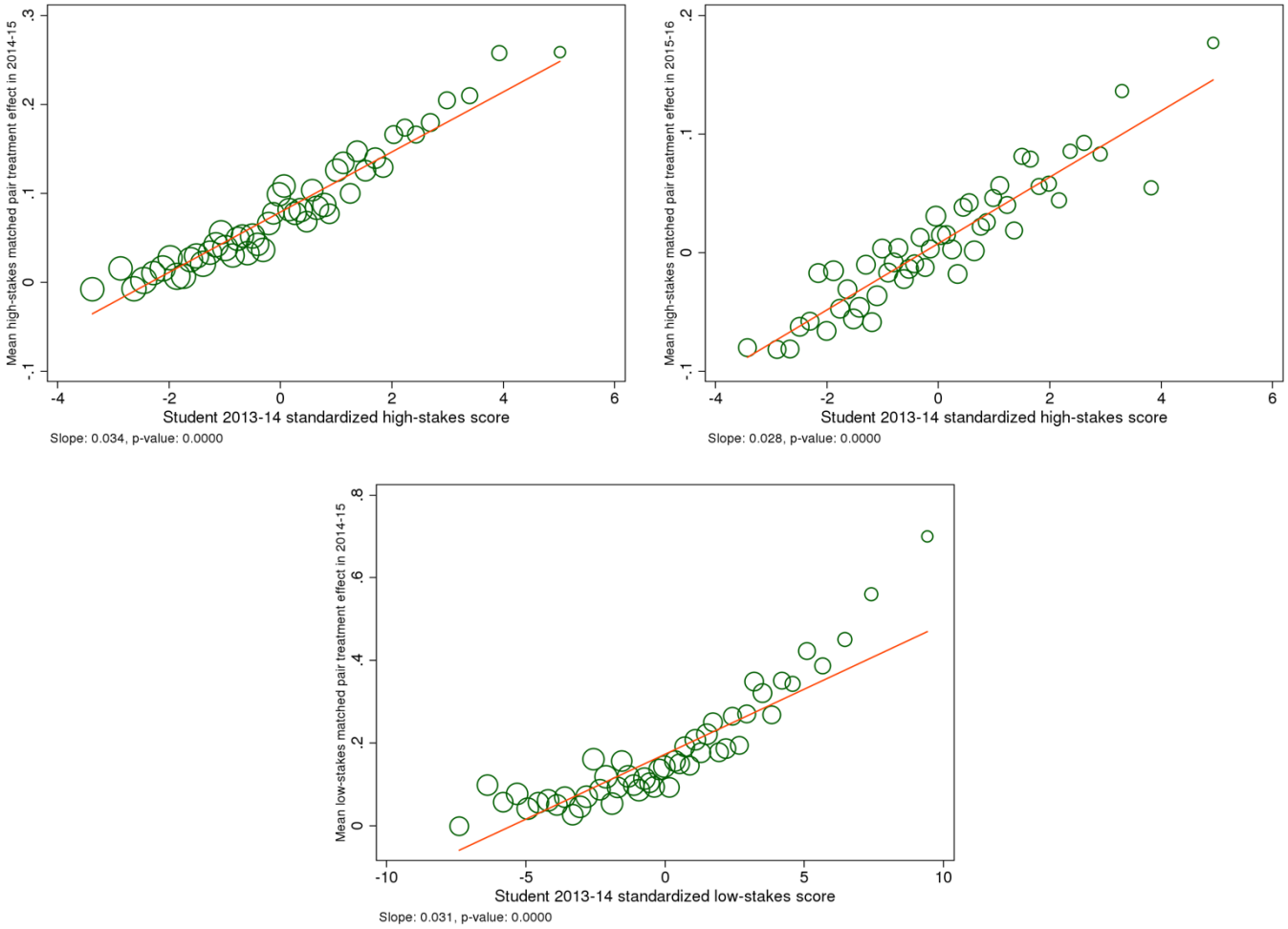
Notes: This table reports ITT estimates of the average yearly effects of our management experiment in Houston on student achievement on high- and low-stakes test scores for subgroups of the sample based on principal and school characteristics. Samples are analogous to those in Table 4. High-stakes tests are the State of Texas Assessments of Academic Readiness (STAAR) exams in math and reading (administered in grades 3-12), and low-stakes tests are the Iowa Test of Basic Skills (ITBS) exams in math, reading, science, and social studies (administered in grades 1-8). Testing variables are drawn from district test score files and are standardized to have a mean of zero and a standard deviation of one within each year and grade among students with valid test scores. Each dependent variable is the sum of standardized scores in all subjects administered (i.e. 2 subjects in STAAR and 4 subjects in ITBS). Columns (1) and (3) report ITT estimates of the effect of treatment. Specifications and samples in Column (1) are analogous to the pooled spcification in Column (3) of Table 4, and specifications and samples in Column (3) are analogous to the first year specification in Column (1) of Table 4. All variables used to partition the sample into subgroups are defined in the Online Appendix. Columns (2) and (4) report the p-value on the null hypothesis that the treatment effect is the same across all subgroups within a given category. All specifications control for 3 years of baseline math and reading test scores and their squares, indicators for whether the baseline test was taken in Spanish, indicators for whether baseline scores are from high- or low-stakes exams, and matched pair fixed effects. Standard errors, reported in parentheses, are robust to heteroskedasticity. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Appendix Figure 1: Treatment Schools' Percent of Trainings Attended



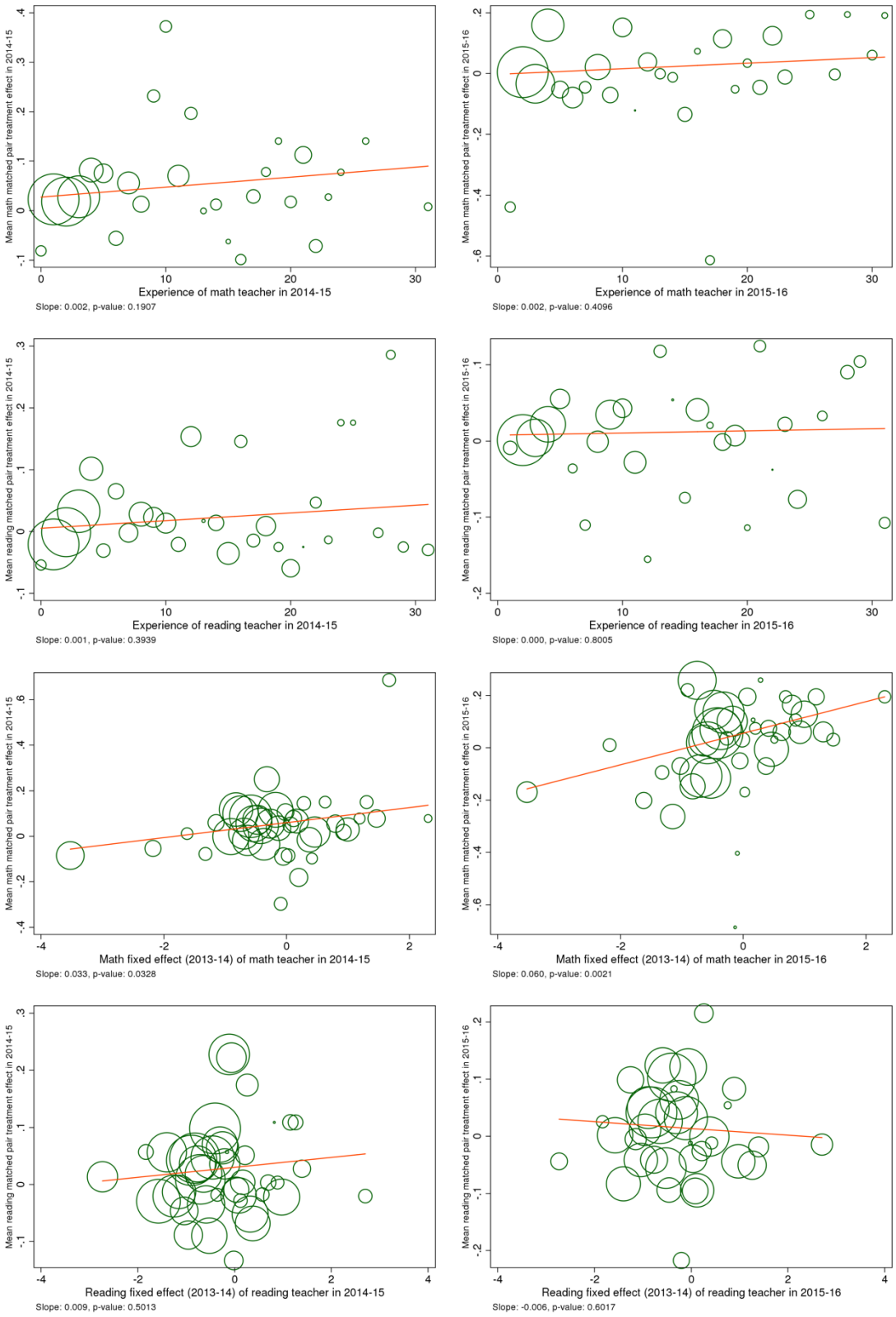
This figure plots the distribution of the percent of trainings attended in each summer and over both summers of our management experiment. The pooled percent of trainings is at the school level - if the principal changed between summers, the school has a value that is the total percent of trainings attended by the principal of that school regardless of whether or not it is the same principal.

Appendix Figure 2A: Heterogeneous Effects by Student Characteristics



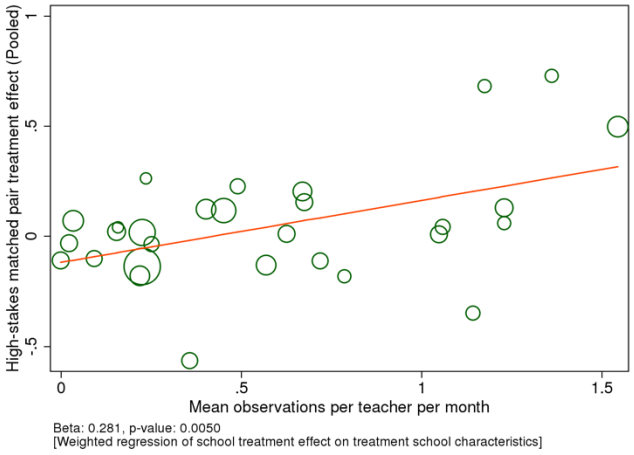
Note: This figure plots average matched pair treatment effects over 50 bins of students' previous year test scores. The slope (and its p-value) is calculated using a regression of average treatment effects on the average student standardized pre-treatment subject test score in each bin, weighted by the number of students in each bin with valid outcome test scores, with standard errors robust to heteroscedasticity (N= 50).

Appendix Figure 2B: Heterogeneous Effects by Teacher Characteristics

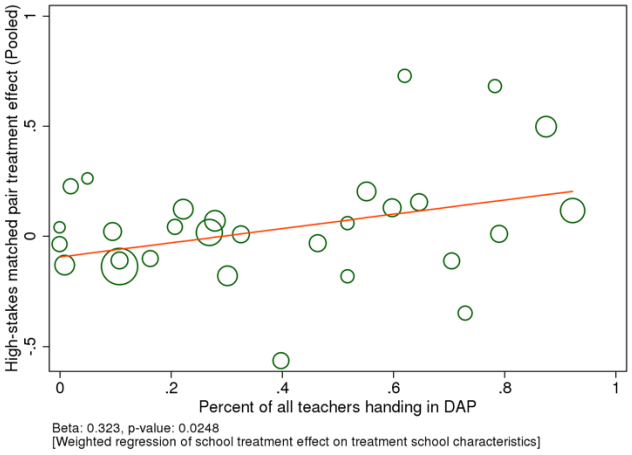


Note: This figure plots average matched pair treatment effects over (i) bins of teacher experience and (ii) bins of teachers' pre-treatment effect on students, standardized to have a mean of zero and standard deviation one over the entire district. The slope (and its p-value) is calculated using a regression of treatment effects on the independent variable, weighted by the number of students with valid test scores in each bin and with standard errors robust to heteroskedasticity. Teacher experience is split by year, and teachers with more than 30 years of experience are binned together. For details on the definition of teacher effects, see the Online Appendix.

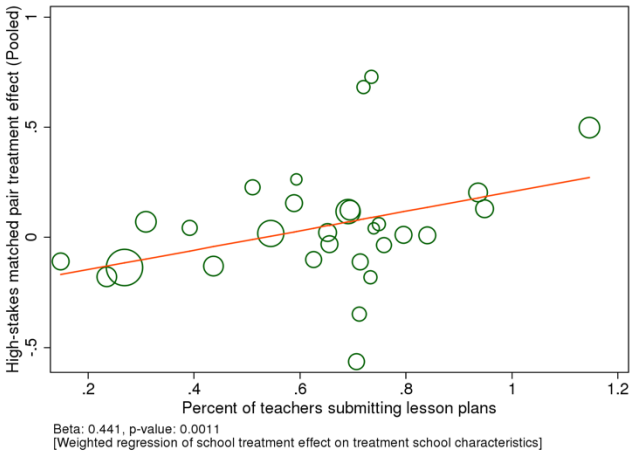
Appendix Figure 2C: Heterogeneous Effects by Principal Characteristics



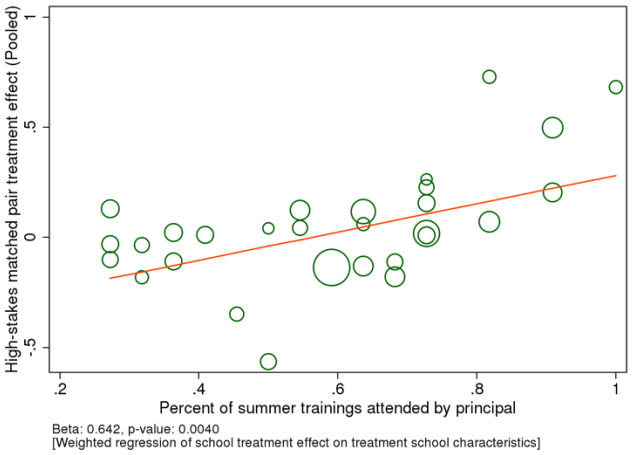
Beta: 0.281, p-value: 0.0050
[Weighted regression of school treatment effect on treatment school characteristics]



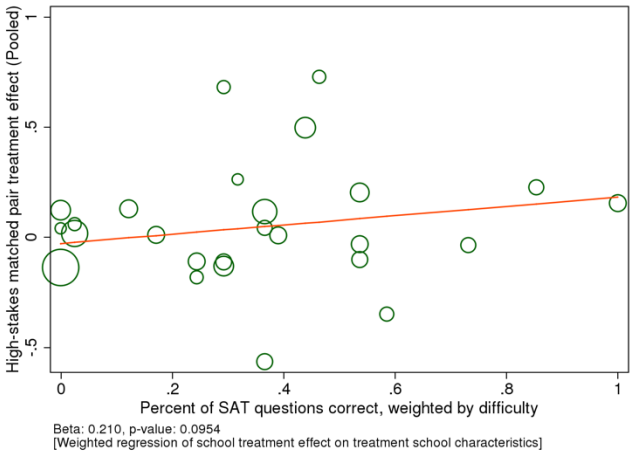
Beta: 0.323, p-value: 0.0248
[Weighted regression of school treatment effect on treatment school characteristics]



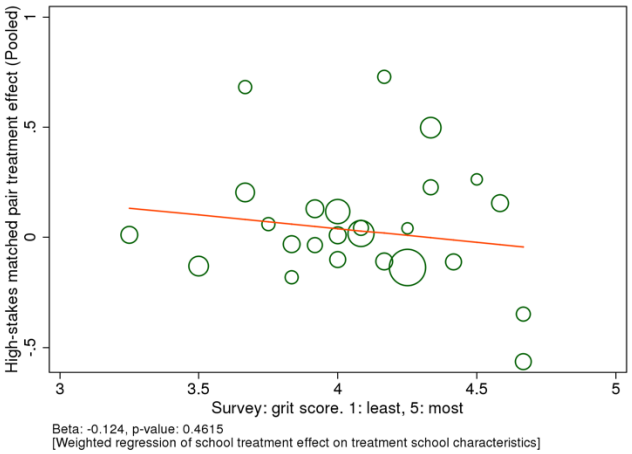
Beta: 0.441, p-value: 0.0011
[Weighted regression of school treatment effect on treatment school characteristics]



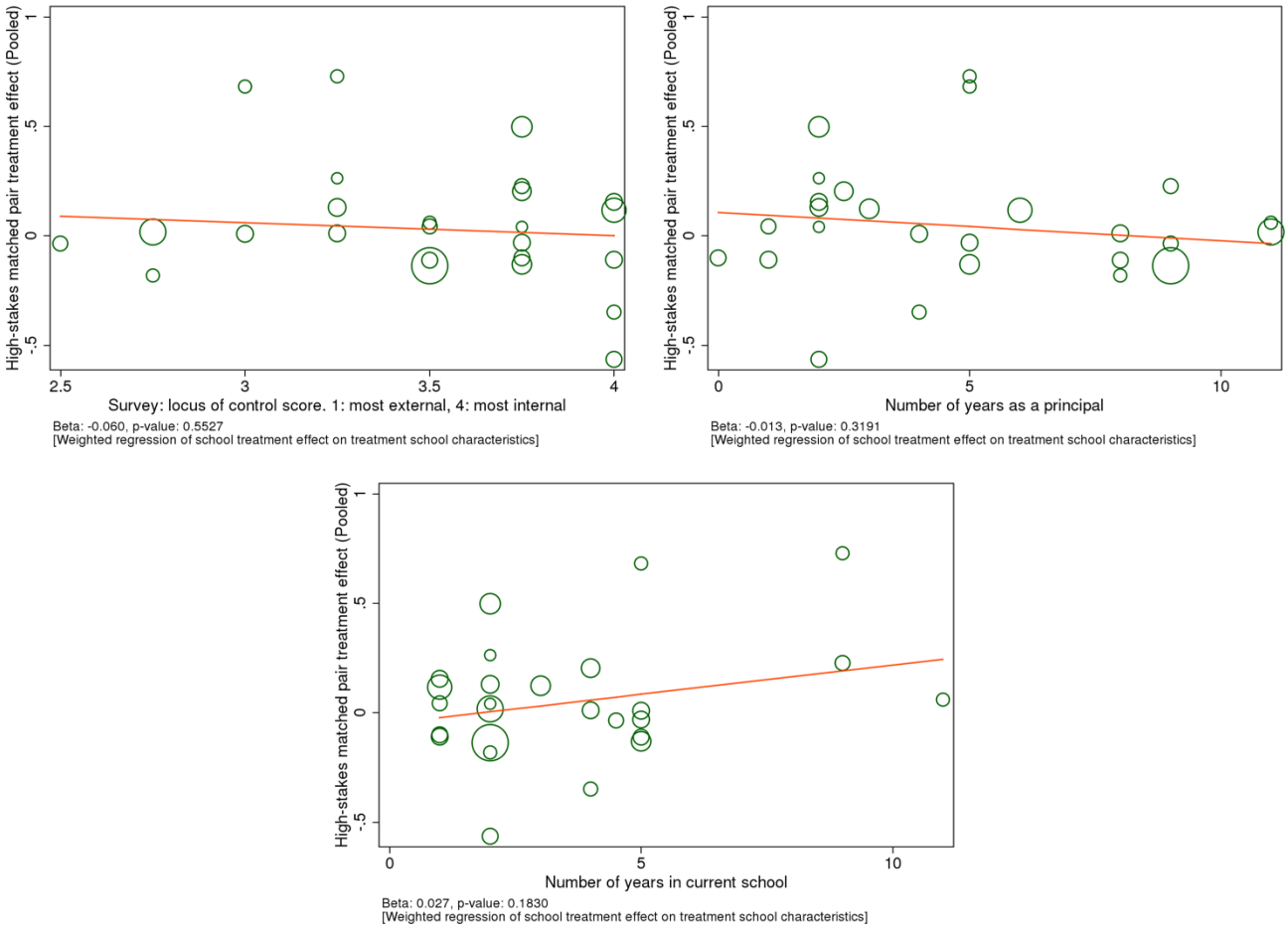
Beta: 0.642, p-value: 0.0040
[Weighted regression of school treatment effect on treatment school characteristics]



Beta: 0.210, p-value: 0.0954
[Weighted regression of school treatment effect on treatment school characteristics]

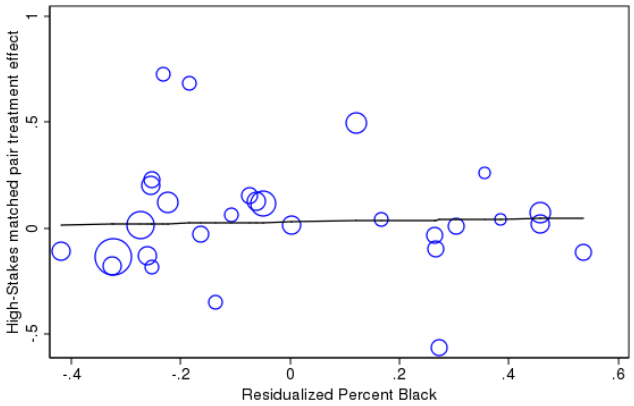


Beta: -0.124, p-value: 0.4615
[Weighted regression of school treatment effect on treatment school characteristics]

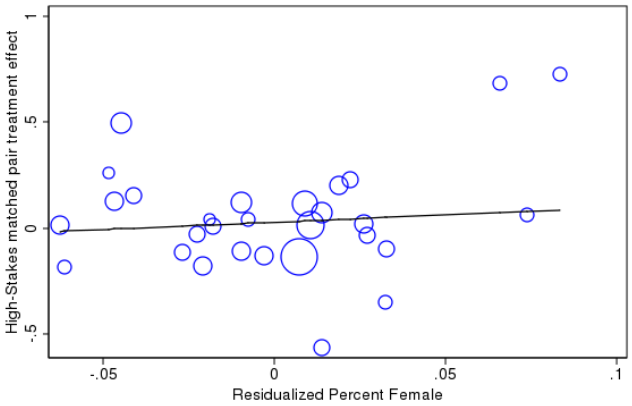


Note: This figure plots matched pair treatment effects against school and principal characteristics. The slope (and its p-value) is calculated using a regression of treatment effects on the dependent variable, weighted by school size and with standard errors robust to heteroscedasticity (N= 29). For details on the construction of dependent variables, see the Online Appendix.

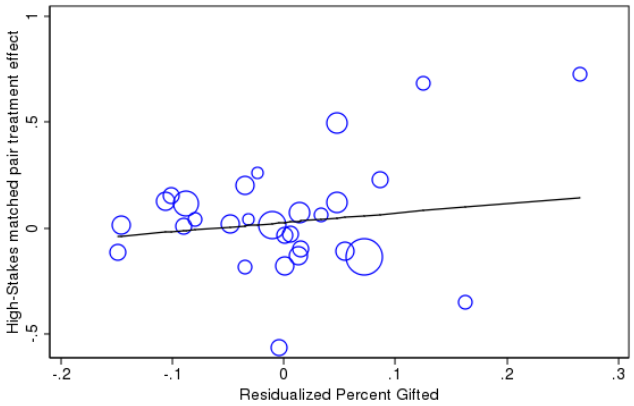
Appendix Figure 3: Robustness of Splitting Sample by Implementation Fidelity



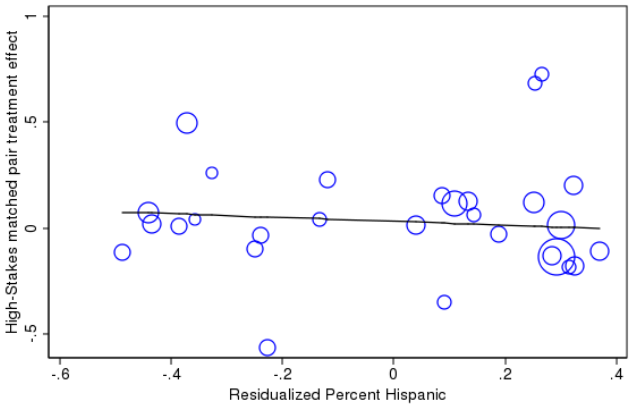
Beta: 0.035, p-value: 0.8000
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



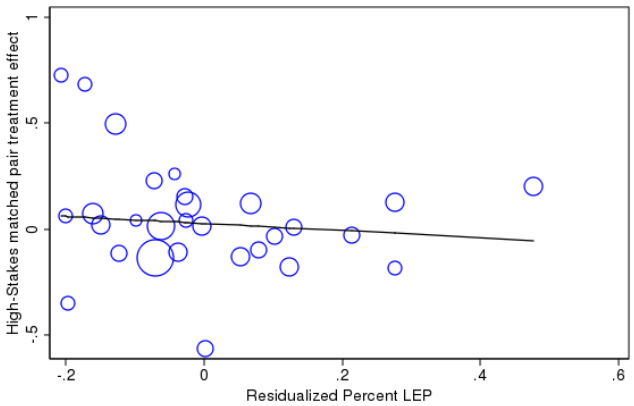
Beta: 0.689, p-value: 0.7105
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



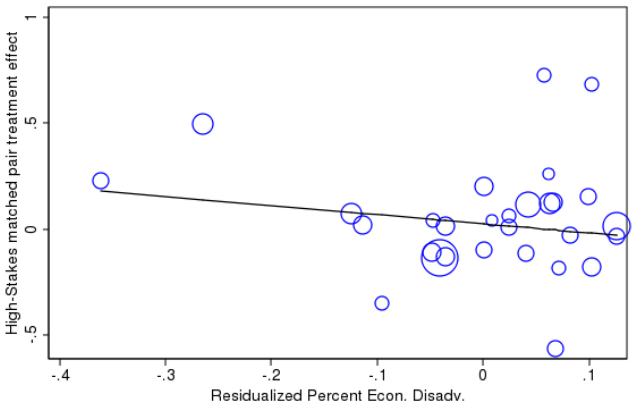
Beta: 0.438, p-value: 0.5404
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



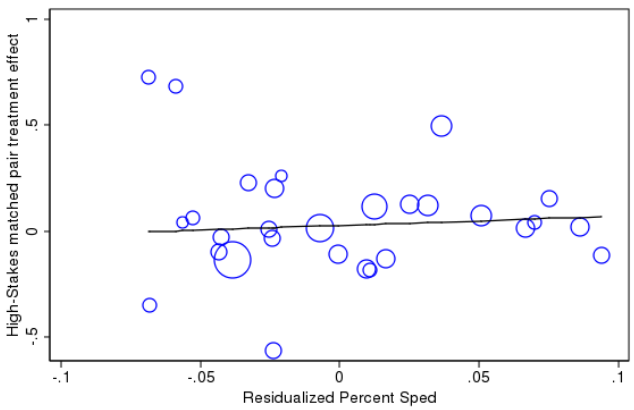
Beta: -0.090, p-value: 0.5658
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



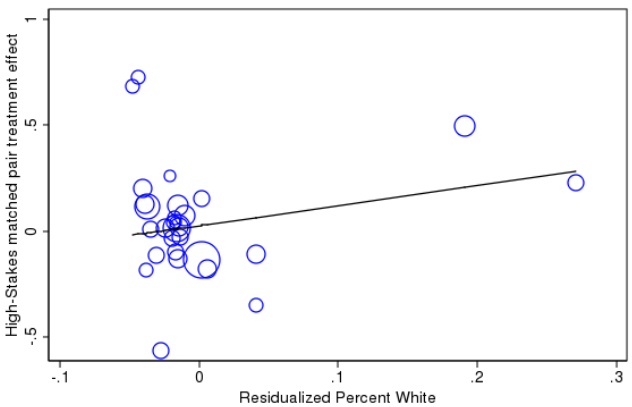
Beta: -0.167, p-value: 0.5989
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



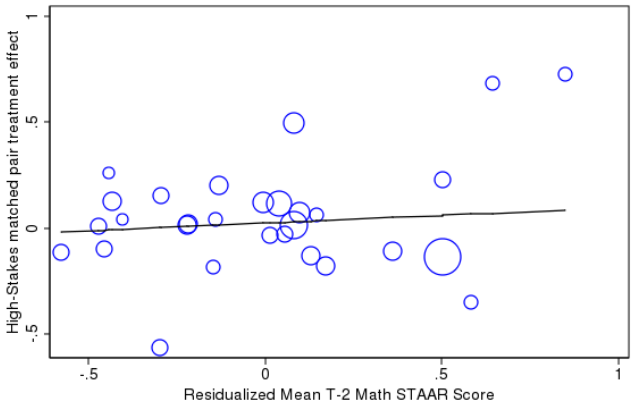
Beta: -0.434, p-value: 0.3331
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



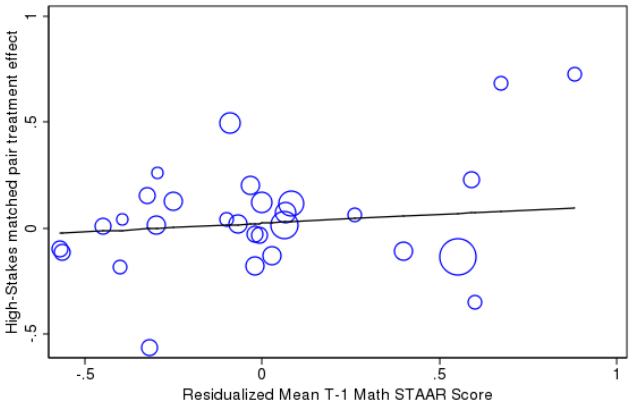
Beta: 0.445, p-value: 0.6615
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



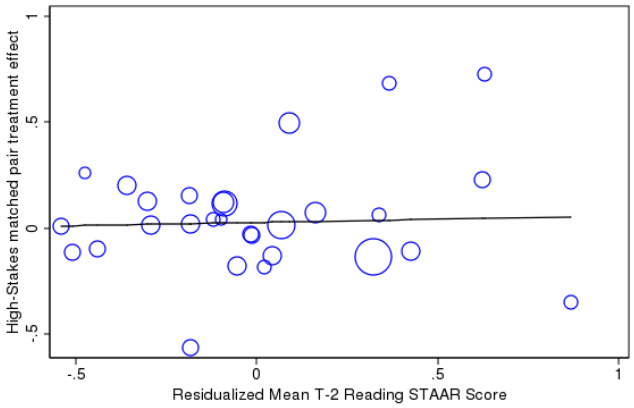
Beta: 0.945, p-value: 0.1955
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



Beta: 0.073, p-value: 0.6695
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]

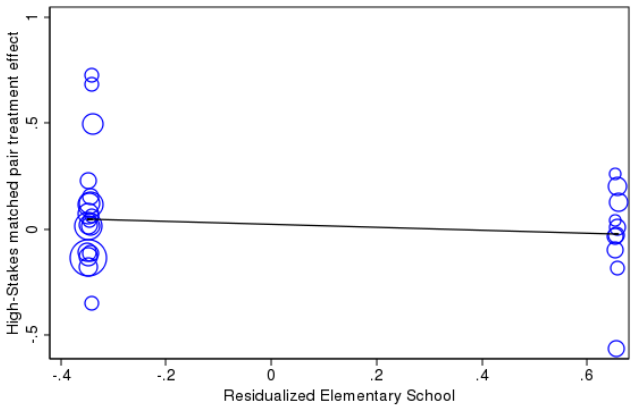


Beta: 0.083, p-value: 0.6294
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]

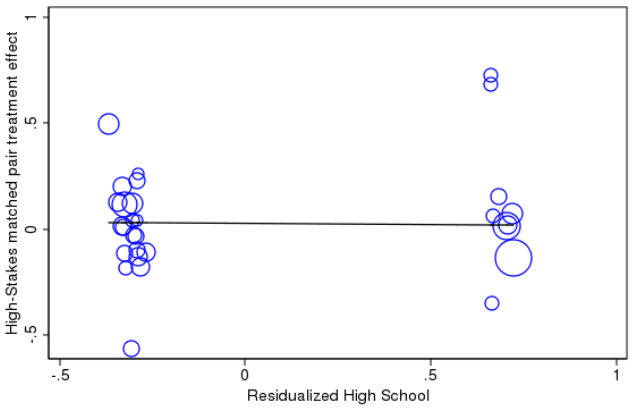


Beta: 0.030, p-value: 0.8497
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]

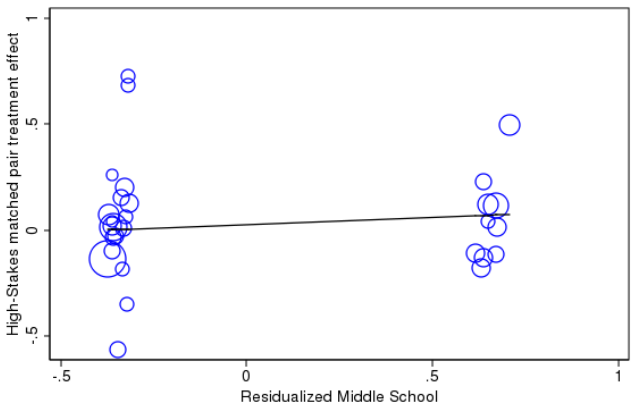
Beta: 0.050, p-value: 0.7352
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



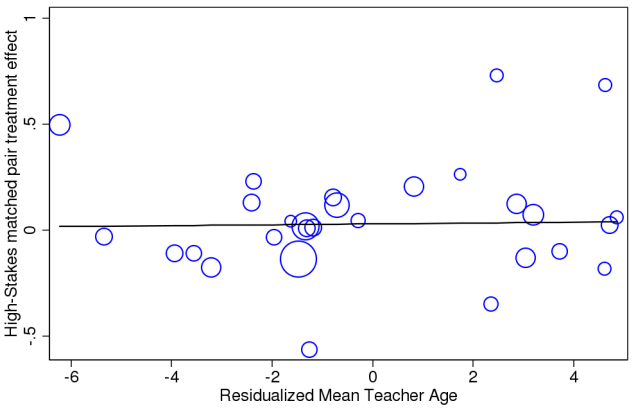
Beta: -0.070, p-value: 0.4588
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



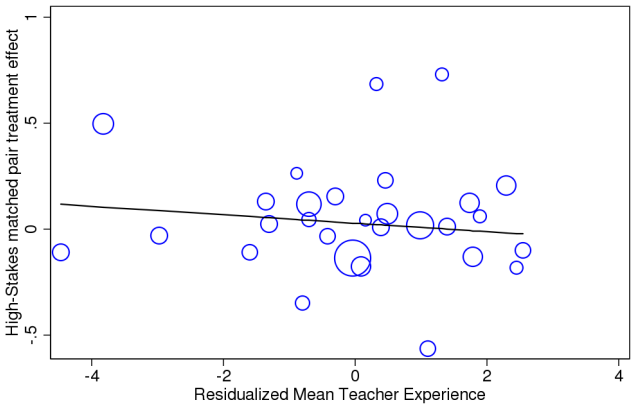
Beta: -0.011, p-value: 0.9097
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



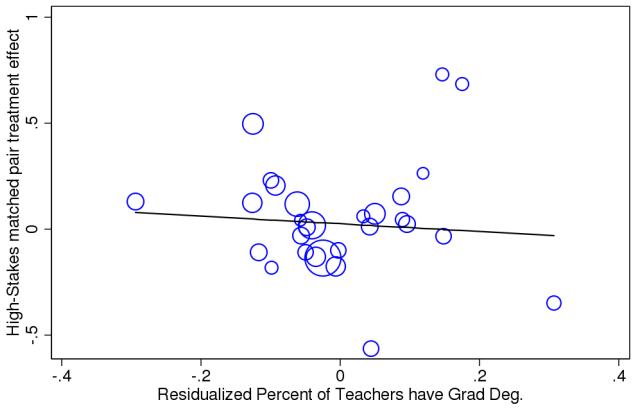
Beta: 0.068, p-value: 0.4540
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



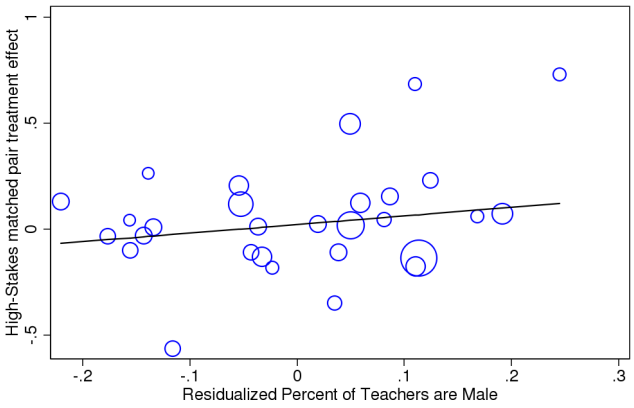
Beta: 0.002, p-value: 0.9158
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



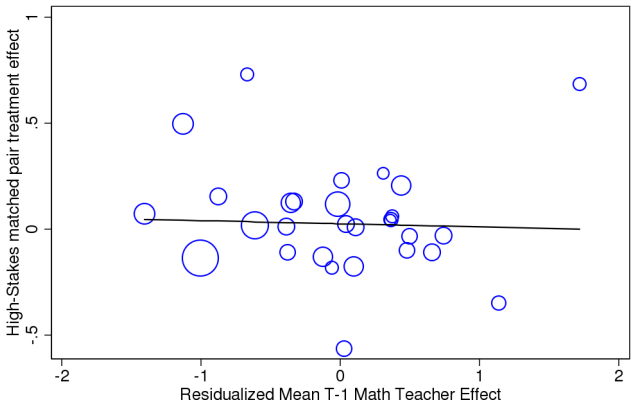
Beta: -0.020, p-value: 0.5237
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



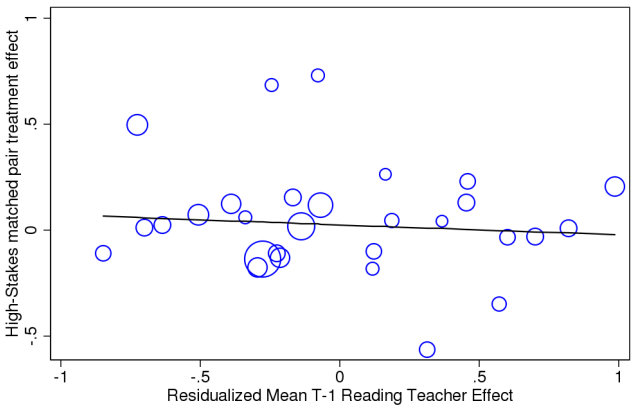
Beta: -0.182, p-value: 0.7077
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



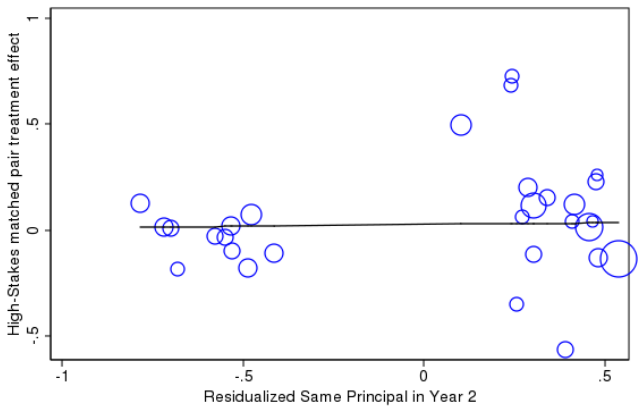
Beta: 0.405, p-value: 0.3373
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



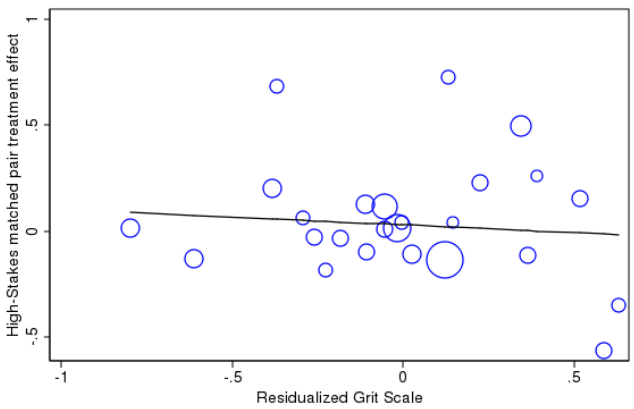
Beta: -0.015, p-value: 0.8665
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



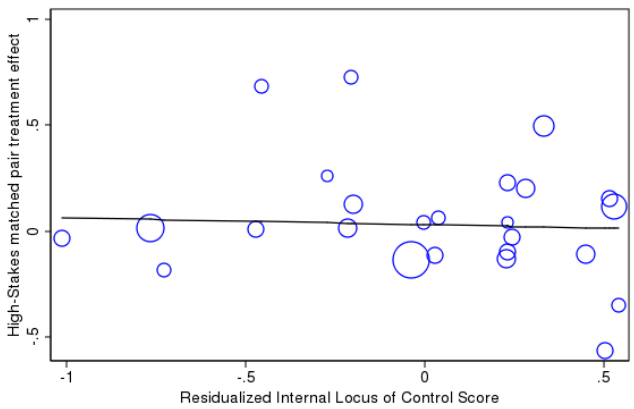
Beta: -0.047, p-value: 0.6184
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



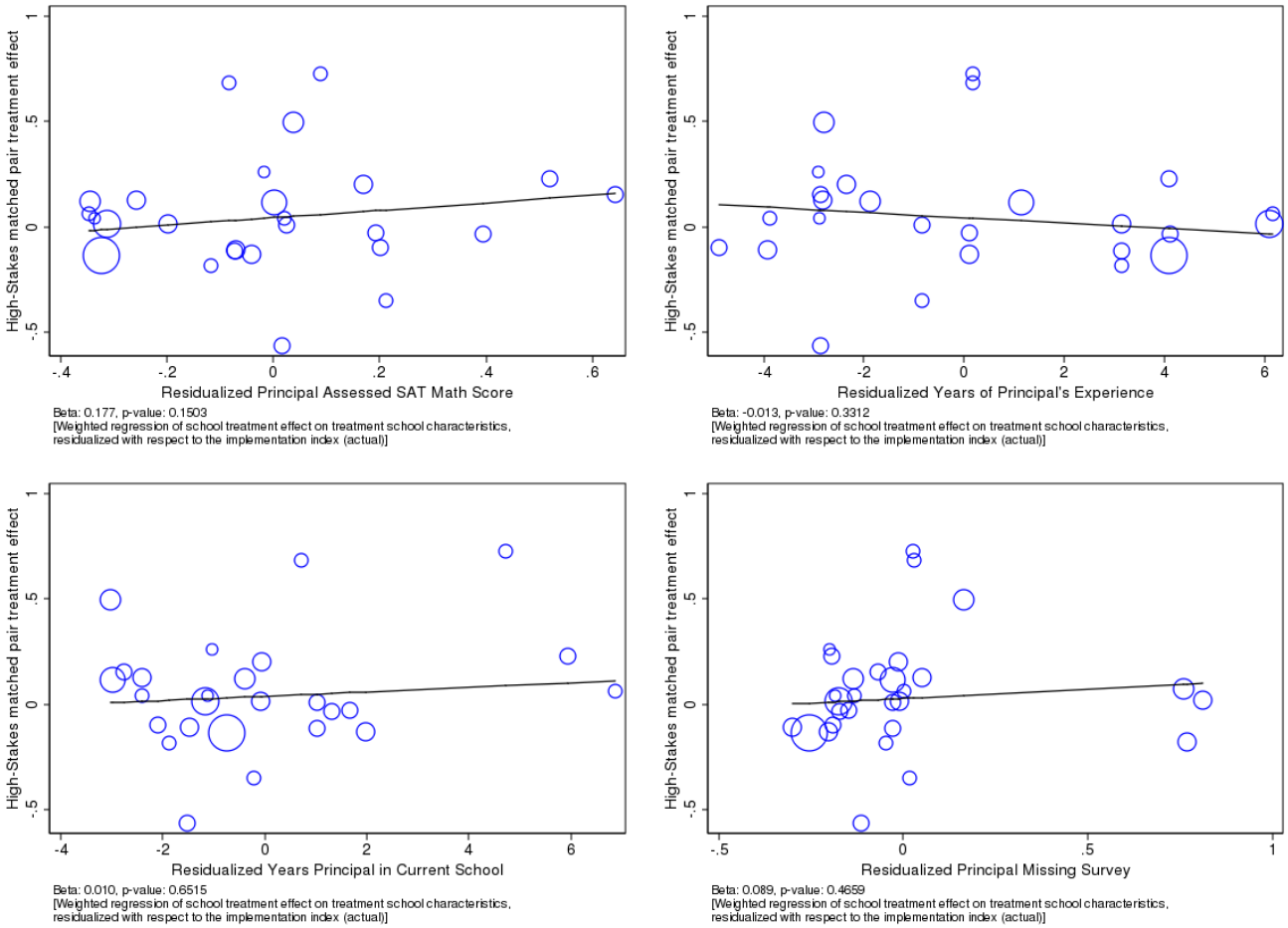
Beta: 0.016, p-value: 0.8175
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



Beta: -0.074, p-value: 0.6774
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



Beta: -0.032, p-value: 0.7481
[Weighted regression of school treatment effect on treatment school characteristics, residualized with respect to the implementation index (actual)]



Note: This figure plots matched pair treatment effects against school and principal characteristics that have been residualized with respect to the fidelity of implementation in each school. The slope (and its p-value) is calculated using a regression of treatment effects on the dependent variable, weighted by school size and with standard errors robust to heteroskedasticity (N= 29). For details on the construction of dependent variables, see the Online Appendix.

Appendix A: Implementation Guide

I. Training and Organizational Structure

Organizational Structure

The HISD in-district supervisory structure was modified to support the implementation of the management protocol on the twenty-nine treatment campuses. Three School Support Officers (SSOs) supervised and supported these twenty-nine schools – one over the elementary schools, one over the middle schools, and one over the high schools. These SSOs were overseen by a Chief Management Officer who was identified and hired by the district due to his experience implementing the practices detailed in *Leverage Leadership* (Bambrick-Santoyo 2012). The Chief Management Officer reported directly to the district superintendent, and interfaced frequently with our program team. The Chief Management Officer was the only marginal hire for the project.

In addition, the Chief Management Officer was supported in his role by the Director of the Office of School Leadership, who provided implementation support by conducting trainings with principals and instructional leaders, making school visits and coaching leaders on aspects of the implementation, and supporting the gathering of monitoring data. Midway through year one, a Program Manager was brought onto the team to oversee the collection and monitoring of project deliverables from schools. Using information supplied by the Program Manager, the Chief Management Officer was able to provide much more targeted support of schools struggling to meet all aspects of implementation, resulting in a sharp increase in the quality of implementation among schools.

In year two, the SSO structure was altered and the treatment schools were returned to the traditional HISD management structure. Instead of having three SSOs who supervised only treatment schools, schools were supervised by district SSOs who also supervised other district schools. The new SSOs were trained in the management model; therefore SSOs who supervised treatment schools did not also supervise any control schools.

Training

To prepare for the implementation of the new management practices, principals engaged in two book study sessions during the spring of 2014 to discuss potential opportunities and challenges in using Bambrick-Santoyo (2012) as a guide for implementation.

In the summer of 2014, all principals were required to participate in two weeks of training focused on the seven levers detailed in Bambrick-Santoyo (2012). This training was led by the Chief Management Officer with support from the three SSOs overseeing schools as well as the Office of School Leadership. Principals were encouraged to invite other members of their leadership teams, including Assistant Principals, Deans of Curriculum and Instruction, Deans of Students, and other instructional leaders. Attendance for each day of the training was approximately 100 school-based personnel. The two weeks of training were split up, with three weeks of work time provided between the first and second week of training to allow leadership teams to adapt materials for their campuses. The specific focus of the training is detailed in the implementation section below.

Each session of the training was focused on a particular management lever from Bambrick-Santoyo (2012). Most of the time was dedicated to direct content training and to giving principals the opportunity to practice the skills they would be expected to implement. There was also time dedicated to setting up the systems that principals would use to monitor the implementation of these management levers at their schools.

Additional training was held in summer 2015. The training was shorter – one full week of content, a working session for principals, and a day of training on monitoring systems. The content for the training focused on a few key areas, such as creating and providing feedback for rigorous

lesson plans, identifying the highest leverage action step during a classroom observation, and setting up systems for effective team planning.

II. Three Levers of School Management

Fusing the best practices described in Bambrick-Santoyo (2012) with the World Management Survey (Bloom et al. 2012) and the political realities of Houston, its school board, and other local considerations, we developed the following five-pronged intervention designed to test whether (and the extent to which) the relationship between school management and student achievement is causal. After assessing the skills and knowledge of principals, the Chief Management Officer decided to narrow the focus of ongoing principal training to ensure that principals understood the markers of high-quality instruction and could effectively manage teachers toward this goal. Trainings therefore focused on levers I-III.

Management Lever I: Instructional Planning

In order to ensure that teachers in treatment schools were designing high-quality instruction by backward induction and to provide instructional coaches with a reference point for classroom observations, all teachers were expected to turn in weekly lesson plans that included specific required lesson components to principals and/or instructional leaders. Leaders were expected to provide teachers with feedback on these lesson plans before the plans were to be implemented in the classroom.

During summer training, leaders received explicit training on the process of backward planning, as well as on how to provide high-quality feedback on teacher lesson plans and how to lead a planning meeting with teachers. Leaders were also given examples of lesson plan templates and each school adapted the templates to be used by their school.

In year 1, we did not have a method for monitoring lesson plan submission or feedback given by leaders on lesson plans. In year 2, HISD implemented a platform called the HUB. Teachers were able to submit lesson plans and leaders were able to provide feedback within the platform. The Chief Management Officer and his team were able to monitor schools' implementation of this lever through the HUB. Submission of lesson plans varied by school; we examine the differential impact of the treatment based on lesson plan submission rates in Appendix Table 11.

Management Lever II: Data-Driven Instruction

For principals to improve their management practices, they needed access to data that allows them to make strategic decisions with their teachers. To assist principals in improving their management through data, all students within treatment schools were assessed every 6-8 weeks in conjunction with the HISD Scope and Sequence to allow principals to work with teachers on re-teaching strategies and differentiated instruction in response to data.

The interim assessments were developed through a collaboration between the HISD curriculum department and our implementing project team. Through the course of the 2014-15 school year, assessments in Grades 1-11 were developed for Reading/ELA, Math, Science, and Social Studies. Additionally, assessments in Spanish Language Arts were developed for Grades 1-5, and Writing for Grades 1-7. For each grade and subject, a minimum of 4 and a maximum of 6 interim assessments were developed, and these were administered on a common timeline approximately 6-8 weeks apart from each other.

Administration of these assessments was tracked through upload of data from assessments to either HISD's data analysis platform, EdPlan, or, in the case of one school, an alternative data

platform, Kickboard. Administration of these assessments by school typically exceeded 90%, but never reached 100% for any single assessment.[1]

After each interim assessment, teachers were expected to analyze their students' performance data and draft an action plan based on the data. Principals (or another member of the school leadership team) would then meet with each teacher, individually or in subject- or grade-level teams, to discuss these plans and modify as necessary. This requirement was monitored through submission of data action plans created through the analysis process. Schools were expected to submit data action plans to the project team within one week following administration of an interim assessment.

Compliance with this requirement was inconsistent, with approximately 50% of schools submitting at least one required data action plan at the start of the year, and this number rising to 75% by the end of year one. Quality of data action plans improved over the course of the year but many schools did not meet the minimum standards expected. Overall completion and submission of data action plans also varied greatly from school to school, with some schools submitting a data action plan for every teacher in the school, and some schools only submitting two data action plans following an assessment. The inconsistency continued in year two, with twenty schools never meeting the expectation of all core subject teachers creating a high-quality data action plan after each interim assessment.[2]

In order for both principals and teachers to make data-driven decisions, schools were also expected to implement weekly formative assessments. These typically took the form of short quizzes at the end of the week to assess standards covered that week, and informed the feedback meetings between instructional leaders and teachers following observations. It took several months for weekly formative assessments to be implemented across treatment schools. The expectation for all teachers of core subjects in all schools to gather formative assessment data from students was not monitored until late February of 2015, with monitoring challenges persisting through March. By the end of March, nearly every school was participating in the weekly data capture expectation, with schools choosing different data platforms to analyze the data, including the district platform EdPlan, Kickboard, Schoolrunner, and Dropbox. Even at the point in which most schools were participating in some capacity, only about 30% of schools reached the point in which all or nearly all teachers in the school were administering a formative assessment on a weekly basis. Additionally, throughout year one, approximately 25% of schools never were able to produce any evidence of any teachers performing the weekly formative assessment. In year two, the expectations for weekly data collection and analysis were maintained. Implementation was variable, with ten campuses never meeting the expectations of weekly data analysis in the 2015-16 school year.

Management Lever III: Observation and Feedback

A key tenet of management best practices is that the performance of employees is regularly monitored. The implementation of the data-driven instruction lever provided leaders with valuable but incomplete performance information. To supplement this data, principals were expected to ensure that all teachers were observed during classroom instruction at least once every other week for 15-20 minutes per observation. The observations were conducted either by the principal or by

[1] One middle school, a magnet school, frequently did not administer as they reported that they were administering internally-developed interim assessments. Three magnet high schools also inconsistently administered these assessments for the same reason.

[2] In the second half of year two, it was also difficult to track whether schools were administering interim assessments. Initially schools were expected to scan and upload student assessment data to EdPlan; however, there were substantial technical difficulties with this process and many schools ceased using EdPlan for this purpose. It is difficult to say if schools that stopped using EdPlan for data analysis continued to administer and analyze interim assessments.

another instructional leader. The leader and teacher then had a face-to-face feedback meeting after the observation to discuss key takeaways and identify a key action step for the teacher to implement in order to improve instruction.

Leadership teams in schools were given significant training and support in what to look for during observations, how to track observations, and how to hold themselves accountable for meeting the goal. School leaders were taught a specific 6-step protocol for conducting the feedback meetings which was taken directly from Bambrick-Santoyo (2012). Several hours during the summer training with leaders and multiple professional development sessions throughout the school year were devoted to learning and implementing the 6-step protocol that focused on identifying one key action step and helping the teacher identify and practice the key action step.

During summer training, schools were provided with an example of an observation tracker, the required components of the tracker, and were given time to set up their tracker for the school year. By January of 2015, every school was utilizing an observation tracker of some form for their campus. Schools developed many different trackers, with some schools using a separate tracker for each instructional leader, some using a common Google spreadsheet for all teachers and leaders, and some using a Google form that would automatically populate a Google spreadsheet. In early Spring, our project team took the observation information currently available from each school and created new observation trackers for every school that included all necessary information in an easily navigable form. These new trackers ensured that leaders could easily track when a teacher was last observed and what the action step was from that observation, thus allowing leaders to take a more systematic approach to talent management on their campuses. These trackers also allowed the Chief Management Officer and his team to deliver targeted feedback to schools on how to improve implementation of this component of the project on their campuses.

For the 2015-16 school year, based on feedback from principals after the 2014-15 school year, the observation and feedback monitoring system was tied into the existing district Teacher Appraisal and Development System (TADS). This enabled school leaders to enter both formal and informal teacher observations into a single platform and allowed all relevant leaders and coaches to access observation data.

Because the TADS platform is for all schools in HISD, we were able to collect data on teacher observations in 2015-16 for both treatment and control schools, providing an important alternative measure of treatment implementation.

Appendix B: Data Description and Variable Construction

Demographic Variables

Demographic variables that should not vary over time (race, gender) were pulled from the 2014-15 HISD enrollment file and were filled in with values from the 2014-15 attendance file if missing in the enrollment file. They were filled in with previous years' enrollment and attendance files (through 2007-2008) if missing, with enrollment always given precedence over attendance files and with more recent data always given precedence over previous years. Demographic variables that may vary from year to year (economically disadvantaged status, LEP status, a special education indicator and a gifted and talented indicator) were only pulled from the 2014-15 enrollment and attendance file (with precedence to the enrollment file). These designations were assigned in the first two months of the first year of treatment. Students who entered the district in 2015-16 were assigned all demographic variables from the 2015-16 HISD enrollment file and were filled in with values from the 2015-16 attendance file if missing in the enrollment file. These designations were assigned in the first two months of the second year of treatment.

- *Race/Ethnicity:* We code the race variables such that the five categories – white, black, Hispanic, Asian, and other – are collectively exhaustive and mutually exclusive. Hispanic ethnicity is an absorbing state. Hence "white" implies non-Hispanic white, "black" non-Hispanic black, and so on.
- *Gender:* Gender was coded as male, female, or missing.
- *Economically Disadvantaged:* A student is considered economically disadvantaged if he is eligible for Free or Reduced price lunch or is flagged as economically disadvantaged without Free or Reduced price lunch. A student is income-eligible for free lunch if her family income is below 130 percent of the federal poverty guidelines, or categorically eligible if (1) the student's household receives assistance under the Food Stamp Program, the Food Distribution Program on Indian Reservations (FDPIR), or the Temporary Assistance for Needy Families Program (TANF); (2) the student was enrolled in Head Start on the basis of meeting that program's low-income criteria; (3) the student is homeless; (4) the student is a migrant child; or (5) the student is a runaway child receiving assistance from a program under the Runaway and Homeless Youth Act and is identified by the local educational liaison. Economic disadvantage is categorically determined by the income level and number of members of a student's household.
- *Limited English Proficient and Special Education:* These statuses are determined by the HISD Language Proficiency Assessment Committee and HISD Special Education Services, respectively; they enter into our regressions as dummy variables. We do not consider students who have recently transitioned out of LEP status to be of limited English proficiency.
- *Gifted and Talented:* HISD offers two Gifted and Talented initiatives: Vanguard Magnet, which allows advanced students to attend schools with peers of similar ability, and Vanguard Neighborhood, which provides programming for gifted students in their local school. We consider a student gifted if he or she is involved in either of these programs.
- *New to District:* Students are considered new to the district if they appear in the enrollment file in a given year but not in the enrollment file in the previous year.

- *New to School:* Students are considered new to their school if their first school attended in a given year is different from the last school they attended and they attended their first school during the first two months of the school year.

Test Scores

The mandated state testing program for students enrolled in grades 3-12 is the State of Texas Assessments of Academic Readiness, or STAAR. Students in grades 3-8 take end-of-grade (EOG) math and reading exams in each year. Fifth grade students take science exams and 8th grade students take social studies exams. Students in the 5th and 8th grade are required to pass their math and reading exams in order to move to the next grade and are allowed two attempts to retake the test. Students in grades 9-12 are required to pass end-of-course (EOC) exams in Algebra I, English I, English II, History, and Biology in order to graduate from high school. State proficiency levels are currently being gradually phased from a less rigorous standard to a more rigorous standard, which will be fully implemented in the 2020-2021 school year. For each student, there is a variable that indicates whether or not they met each year's phased minimum performance standard and whether or not they would have met the recommended standard that will be used in 2021. There is also a variable that indicates whether or not a student achieved a commended performance. There are multiple versions of the STAAR exam. STAAR-A is an accommodated version of the test, introduced in 2014-15, that is comparable with the standard version of the test (STAAR-S). STAAR-A scores are included in this analysis. The STAAR-L is a linguistically modified version of the exam that is not comparable to STAAR-S; scores on this exam are not included in this analysis.

HISD also administered a low-stakes exam in each year leading up to and including 2014-15; in all years before 2014-15 HISD administered the Stanford 10 and in 2014-15 HISD administered the Iowa Test of Basic Skills (ITBS). Both are nationally-normed tests. In 2015-16, HISD did not administer a low-stakes exam. In STAAR, Stanford, and ITBS test files, students with a raw score of zero (i.e. zero questions answered correctly) were assumed to have not taken the test. In all files, the scaled score, which converts raw scores to be comparable across test forms given on different days within a given year, is used as the main score.

Some students in some years have multiple entries in the various testing files. Multiple entries were dealt with using the following procedure:

- STAAR EOG: If a student took both an on-time test and a retest, use the on-time test. If a student took both STAAR-S and STAAR-A, use the STAAR-S score. If a student took a test in both English and Spanish, use the English score. If a student took a test with accommodations (extended time, etc.) and without, use the non-accommodated score. If a student took multiple grade-level tests, use the one from the student's enrolled grade. If a student took tests in multiple grades and is enrolled in neither, take the one from the grade closest to the enrolled grade or the lower of two grades if they are equidistant from the enrolled grade. If there remain multiple test scores that have the same accommodations, language, and grade, take the maximum score.
- STAAR EOC: If a student took both an on-time test and a retest, use the on-time test. If a student took both STAAR-S and STAAR-A, use the STAAR-S score. If you took multiple English tests in one year, give precedence to English I if the

student is in grade 9 or lower and precedence to English II if the student is in grade 10 or above. Take the maximum of all remaining multiple test scores. In the July 2016 retest file, scaled scores were missing. They were filled in using raw score-scale score conversion charts for each test type, administration type, and subject available on the TEA website.

- TAKS 11th grade exit level exams: Students who entered high school before 2011 are still required to take the TAKS exit level exams (the test used before STAAR). If students have multiple TAKS scores, they are dealt with using the same procedure as the STAAR EOG scores.
- If a student took both an EOG and EOC test in a given subject, give precedence to the EOG score if a student is enrolled in grades 3-8 and to the EOC score if a student is enrolled in grades 9-12. If a student took both TAKS and EOC, give precedence to EOC. Proficiency levels are assigned based on the final score assigned to each student after multiples have been dealt with.
- Stanford/ITBS: If a student took multiple grade-level tests, use the one from the student's enrolled grade. If a student took tests in multiple grades and is enrolled in neither, take the one from the grade closest to the enrolled grade or the lower of two grades if they are equidistant from the enrolled grade. If there remain multiple scores from the same grade, take the mean score. If a student took both a Spanish and English test, give precedence to the English test.

Scaled scores are standardized to have a standard deviation one and mean of zero within each subject, grade, and year across HISD after all multiple testing entries have been dealt with. STAAR EOC scores are standardized by subject and year. Stanford/ITBS scores are standardized within testing language by year and grade, since those English and Spanish tests are not comparable.

All results are robust to using simply the highest score a student ever achieved, and by giving precedence to retakes and then following the same procedure.

Administrative Measures of Principal Management

- *Percent of trainings attended:* There were nine principal training sessions over the course of the summer of 2014 and 2015. Each session had a sign-in sheet used to determine attendance. Given principal turnover, this measure is coded at the school level: if a principal showed up to the training who is at that time working at school X, school X gets a 1 in a series of attendance indicators. The final variable is the percent of trainings attended. Note that one treatment school vice principal in year one became a control school principal in year 2 after having attended a training in the summer of 2014 and thus the mean number of trainings attended for control school principals is not zero.
- *Teacher observations:* Principals and other administrators who observe teachers are required to fill out the Teacher Appraisal and Development System (TADS) for each teacher observation they complete. We use data from TADS for 14 control schools and all treatment schools and the number of teachers employed at the school at the beginning of treatment to calculate the average number of observations per teacher per school. Since the data was obtained at different times in the school year for treatment and control schools (control schools was two months later than treatment) we scale this number to be the average number of observations per teacher per month. Fifteen control schools did not submit data to TADS. Their number of

observations is filled in with zero. Replacing this value with (a) one observation per year, as required by HISD, or (b) the mean value of all other controls schools does not change results.

- *Lesson Plan Submission Rates:* Each week, teachers were expected to submit their daily lesson plans through an online platform called the HUB, available through HISD. Instructional leaders were expected to review and provide feedback on those plans within the platform. We were able to monitor the submission on plans using the HUB, and each week we recorded how many sets of lesson plans were submitted for a school, relative to the number of teachers assigned to that school. The number of teachers to submit a lesson plan was averaged over each week of the school year and divided by the number of teachers in each school who were required to be submitting lesson plans at the beginning of the first year of the experiment to calculate the percent of teachers submitting lesson plans.
- *Data Action Plan Submission Rates:* After each snapshot assessment window (every 6-8 weeks) the HISD project team tracked the submission rate of data action plans for each subject at each school. Teachers were expected to submit their data action plans for their instructional leaders to review on the HUB; the HISD project team was able to access these plans through the same platform. The percent of all teachers to submit a data action plan was averaged over the total collection periods for each treatment school. Middle and high schools had five total collection periods; elementary schools had three.
- *Administrative Implementation Index:* The administrative measures of teacher observations, DAP submission, and lesson plan submission were standardized to have a mean of zero and standard deviation one over the sample of experimental schools for which we had data. The implementation index is the mean of those three standardized measures. Control schools receive the value of the implementation index that their matched pair had earned.

Principal Survey Responses

- *Effectiveness of training:* Principals were asked, "how effective was any training you received for the 2014-15 school year compared to any training received for the 2013-14 school year?" and responded on a scale of 1-5 where 1 indicated significantly less effective and 5 indicated significantly more effective. The analysis uses a binary indicator that is a 1 if the principal answered 4 or 5 (slightly or significantly more effective) and a zero otherwise.
- *Average number of teacher observations:* Each principal was asked how many times he or another member of his leadership team observed each of four core teachers (randomly selected from the appropriate school). An average of these responses makes up the survey measure of teacher observation.
- *Percent of teachers handing in lesson plans:* Principals were asked, "approximately what percentage of teachers on your campus submitted weekly lesson plans during the 2014-15 school year?"
- *Years as principal/in current school:* Principals self-reported the number of years they had been a principal and the number of years they had been in their current school.
- *Locus of control:* Principals answered four multiple-choice questions (the Rotter abbreviated scale (Rotter 1966; Valecha and Ostrom 1974) concerning their locus of control which were coded such that a (1) indicated the most external and (4) indicated the most internal locus of control. The mean answer to these four questions is our measure of locus of control.

- *Grit:* Principals answered 12 multiple-choice questions concerning their level of grit (see Duckworth et al. (2007)) which were coded such that a (1) indicated the least grit and (5) indicated the most grit. The mean answer to these twelve questions is our measure of grit.
- *Score on Math SAT Questions:* Principals were asked 13 multiple-choice SAT Math questions and 2 free-response SAT Math questions. These questions were graded (as right or wrong) and weighted by difficulty level. The weighted percent of questions correct is used as our measure of principal IQ/ability. All unanswered questions are coded as wrong.

Principal Shadowing – Time Use Diaries

Principals in each school were shadowed for up to two days by aspiring principals as a part of their training program. The shadowers kept detailed time diaries of their time spent in each school and coded principals' time use into 27 distinct tasks. Those tasks were aggregated to the following 7 types of activities. We use the percent of total time a principal spent doing each of these types of activities as the dependent variable in Figure 3.

- *Investing in human capital:* time spent meeting with teachers and school-based staff, reviewing teacher lesson plans, observing classroom instruction, leading professional development for teachers, or attending professional development for themselves.
- *Community and parent communication:* time spent meeting with parents, on phone calls with parents, or meeting with alumni/community members.
- *Meeting with school or district leadership:* time spent meeting with the school administrative or leadership team, or central HISD employees (e.g. School Support Officers, Teacher Development Specialists).
- *Administrative tasks:* time spent completing administrative paperwork, emails, dealing with vendors, facilities or school technology systems, planning school events, updating bulletin boards, or making announcements.
- *Whole-school activities:* time spent observing whole-school transitions, leading whole-school events, or attending student extracurricular activities.
- *Examining student performance data:* time spent using student performance data
- *Student interaction:* time spent teaching a class, addressing discipline problems, or meeting with students.

Assignment to Treatment

In the first year of treatment, students are assigned to the first school that they attended in HISD in the 2014-15 school year if they attended that school before November 7 (the end of the second reporting period). This was taken from the attendance file; if a student did not appear in the attendance file but did appear in the enrollment file (a snapshot taken in mid-October) then the student was assigned to the school that they attend according to the enrollment file. In the second year of treatment, there are three subsets of students. Students in non-entry grades (1, 2, 3, 4, 5, 7, 8, 10, 11, 12) are assigned to their same school as the first year of treatment. Students in entry grades (6, 9) are assigned to their zoned middle or high school according to their address *before* the first year of treatment; i.e. the middle or high school that they were supposed to attend in 2015-16 given their address in 2014-15. Thus 6th and 9th graders both enter and exit the experimental sample in the second year. Students who repeat 6th or 9th grade or regress in grades are assigned to their same school as the first year of treatment. Students who enter the school district in the second year of

treatment are assigned to the first school they attended in that year, analogously as in the first year of treatment. In each year, we assign students in one of the 29 treatment schools to the treatment group and students in one of the 29 control schools to the control group.

Teacher Value Added and Employee Files

Teacher Demographics
Students are split into subsamples based on their teacher's gender, experience, graduate degree level, which come from the HISD-provided employee file from 2014-15 and 2015-16.

Student-Teacher Linkage
We link students to their teachers using a file of course grades in each year of treatment. Course grades are dropped if they are not quarterly academic grades. Courses that were taken in a school other than the school a student is assigned to for the purposes of treatment assignment are also not included (i.e., students are only linked to teachers in their experimental ITT school).

- Elementary and middle school students: A student is linked to his or her teacher in math, reading, science, and social studies in the first reporting quarter of the year. ESL and Language Arts were considered reading courses. Students are linked to their teacher in each subject only if they were enrolled in the course during the first quarterly reporting period. Within each subject, it is possible for a student to have multiple teachers. Students were linked to one teacher per subject using the following procedure: Precedence was given to the teacher who taught a student in the most courses (e.g. a teacher who taught both reading and language arts was given precedence over a teacher who taught just ESL). Precedence was then given to teachers who taught the most relevant course/s, using the HISD course catalogue and course descriptions.
- High school students: Since high school students testing outcomes are only in Algebra and English, students are linked only to their teachers in those specific subjects. Since these courses can be semester-long rather than yearlong, students are linked to their first teacher in either subject for the year. Within each subject, it is possible for a student to have multiple teachers. Students were linked to one teacher per subject using the following procedure: Precedence was given to the teacher who taught a student in the most courses. In Algebra, precedence was given to the teacher who taught Algebra I, then to Algebra A over B (at any level), and then to regular classes over modified, alternate, or pre-AP Algebra classes. In English, precedence was given to the teacher who taught the course that the student tested in (either English I or English II), then to the lowest level English course, then to regular courses over modified, advanced, or pre-AP English courses.

Teacher Value Added
HISD officials provided us with 2013-14 value-added data – district-calculated Cumulative Gain Indices. However, due to the nature of official TVA calculations in the district, only 17 (19) percent of teachers in the district have TVA measures in math (reading). In order to use more of the sample, we calculate our own measures of teacher effects in the year previous to treatment. We regress standardized student test scores in 2013-14 on test scores in 2012-13 and their squares, student demographics (gender, race, and indicators for LEP, special education, gifted and talented, and economically disadvantaged status) and grade fixed effects plus a *full range of teacher fixed effects* (for teachers linked to the students that they teach in the subject of the test). Students are linked to teachers using the course grades file from 2013-14 and are linked to any teacher who taught them in

a math or reading course throughout the year. Students with multiple teachers in a given subject enter the regression more than once. The coefficients on the teacher fixed effects are considered a gain-based measure of a teacher's "effect," controlling for student demographics and previous year test scores – these are standardized across the district to have a mean of zero and standard deviation one. We are able to calculate this measure for more than twice as many teachers as we have official TVA calculations for (32% of all teachers have math effects, 36% of all teachers have reading effects). If we limit to teachers who teach either math or reading, we have self-calculated measures for 58% of math teachers and 66% of reading teachers. Among teachers with non-missing values of both measures, the correlation between the official measure of TVA and our calculated teacher effect is 0.65 in math and 0.49 in reading. Throughout the paper, we use our calculated measure of teacher effects rather than official TVA measures. The average school in our experimental sample has 10 teachers with teacher effects in math and 12 teachers with teacher effects in reading.

Appendix C: Return on Investment Calculations

We calculate back-of-the-envelope Internal Rates of Return (IRRs) based on the expected income benefits associated with increased student achievement. We follow Krueger (2003) to calculate the IRRs. Let E_t denote an individual's real annual earnings at time t and β denote the percentage increase in earnings resulting from a one standard deviation increase in test scores. The IRR is the discount rate r^* that sets costs equal to the discounted stream of future benefits:

$$C_0 = \sum_{t=T_0}^{T_N} E_t * \beta(\tau_m + \tau_r) * (\frac{1+g}{1+r})^t$$

where T_0 is the time period in which the individual turns 18 and enters the labor market, T_N is the time period in which the individual turns 65 and retires, τ_m and τ_r denote the treatment effects for math and reading, respectively, and g is the annual rate of real wage growth.

According to the literature on the relationship between test score gains and lifetime earnings, β lies somewhere between 8 percent and 12 percent (Krueger, 2003 and Chetty et al., 2014). Krueger also notes that real earnings and productivity have historically grown at rates between 1 percent and 2 percent, which are plausible rates for g. For the purpose of this cost benefit analysis, we set $\beta = 0.12$ and g equal to 0.02, and approximate E_t using the Current Population Survey.

For each intervention, we calculate cost per student per year for both treatment and control, the age at which the intervention starts, treatment effects in math and reading, and the year the individual enters the labor market. Below we describe these interventions and calculations in greater detail, and present the resulting IRRs in Appendix Table 6. For ease of comparison, the table presents the sum of math and reading treatment effects, wherever applicable, and presents the cost numbers in 2014 dollars; however, IRR's are calculated based on cost numbers from the relevant year of the intervention. In the description below, all cost numbers are described from the year of the intervention.

Management Training

For our management experiment, we spent approximately $445,000 over the two years. This includes the cost of materials used in training, the technology systems used to manage student data, the salary of the Chief Management Officer, and the cost of preparing interim assessments for both treatment and control schools. With 24,000 students and 31,000 students in treatment and control schools each year, respectively, this brings the cost per student in treatment per year to $9.61 and the cost per student in control to $0.35. For a student who is 11 years old at the beginning of the implementation and enters the labor market at age 2021, spending an average of 1.5 years in treatment, we calculate IRRs for the overall experiment (79%), for the schools in which we predict high implementation (96%), and for the schools in which we predicted the principal staying (94%). The associated treatment effects on the sum of math and reading are 0.060σ (0.01), 0.120σ (0.015), and 0.113σ (0.014), respectively. The results are presented in Appendix Table 6. All cost numbers are in 2014 dollars.

Financial Incentives

Coschocton Incentive Program: Bettinger (2012) evaluated a pay-for-performance program for students in grades three through six in Coschocton, Ohio from 2004-2007. Eligible students received cash payments for improving achievement in standardized tests for five core

subjects: math, reading, writing, science and social studies. He reports a 0.133σ (0.0485) increase in math scores, a 0.01σ (0.0454) increase in reading scores, a 0.23σ (0.041) increase in social studies scores and a 0.048σ (0.039) decrease in science scores. Writing scores are excluded from the analysis because in any given year, different grades took different writing tests that were not comparable. We include the social studies and science effects in the calculation for the Coschocton Incentive Program's IRR because the experiment provided incentives for improving test scores across all subjects. Pooling these effects therefore gives a more comprehensive view of treatment effects. In this experiment, randomization was done at the grade-school level each year of the experiment, yielding a total of 1,615 students in the experimental sample, with 801 students being eligible for treatment overall (source: personal communication). As the program cost $52,000 in incentives and administrative costs across three years, the cost per student per year is approximately $65. Using an average initial age during intervention as 11, we get an IRR of 49%. All cost numbers are presented in 2006 dollars.

NYC, Dallas, and Chicago: Fryer (2011) summarizes the results of financial incentives on student achievements in New York, Chicago, and Dallas. In Dallas, second grade students were paid to read books. In New York, students were rewarded for performance on interim assessments. In Chicago, students were paid for classroom grades. The incremental cost per student per year in Dallas was $62.21 which included $13.81 paid on average in incentives and $86,000 in administrative costs. In New York, the incremental cost per student per year was $377.04 for 7th graders and $339.25 for 4th graders, including average incentives paid and administrative costs, but excluding $500 spent per school to collect surveys. The incremental cost per student per year in Chicago was $373.76 which included incentive payments and $85,000 in administrative costs. The weighted average cost for an extra student in the three experiments was $323.47. The estimated treatment effect on math and reading scores is zero when pooled across all three cities. As a result, we are unable calculate an IRR, because the discount rate would have to be a very large negative number to bring the net present value of costs equal to zero. All cost numbers are presented in 2009 dollars.

Teacher Certification

Teach for America: Teach for America is a non-profit organization that recruits recent college graduates to teach for two years in low-income communities. Glazerman et al. (2006) report findings from a national randomized evaluation of the impact of TFA on student outcomes. The experiment involved approximately 100 elementary classrooms, grades 1 through 5, from 17 schools across Baltimore, Chicago, Compton, Houston, New Orleans, and the Mississippi Delta. Students were stratified by grade and school and randomly assigned to either a TFA or non-TFA teacher. Glazerman et al. (2006) report that students assigned to a TFA teacher score about 0.15σ (0.04) higher in math and 0.03σ (0.04) higher in reading than students assigned to non-TFA teachers. In an interview, the national spokesperson for TFA, Takirra Winfield, claims that TFA spent around $16,400 to recruit and select each new teacher, $7,000 to train them, and $14,000 per year on stipends for the two years of the program (Cohen 2015). Thus, we get a total cost of $51,400 per TFA recruit per year. This study had a total of 44 TFA teachers teaching 785 students, giving a per student cost of $2,881. Using an average initial age during intervention of 9, we get an IRR of 11.73%. All cost numbers are presented in 2003 dollars.

Early Childhood Interventions

Head Start Impact Study: Head Start is a preschool program funded by federal grants, and is designed to serve 3- to 5-year-old children living at or below the federal poverty line. Puma et al. (2010) evaluate Head Start by studying randomized admission into the program. They investigate the impact on two different cohorts, a 3-year-old cohort, which is exposed to the program for two years, and a 4-year-old cohort which is exposed to the program for just one year. Puma et al. (2010) report that winning a lottery to attend Head Start resulted in an increase of 0.135σ (0.071) in math test scores and 0.188σ (0.064) in reading test scores. According to a National Institute for Early Education Research report, the average spending per child in Head Start was $9,198 in 2010. However, this is not necessarily the marginal cost of Head Start because as Puma et al. note, approximately 60 percent of the control group children in their study participated in child care or other early education programs. Based on the same report, average spending per child on other pre-K programs was $4,831. Using these cost calculations and an average initial age of 4, we get an IRR of 9.19%. All cost numbers are presented in 2010 dollars.

Class Size

Tennessee STAR experiment: Project STAR was an experiment carried out in 79 Tennessee schools from 1985 to 1989 where 11,600 students in kindergarten to third grade were randomly assigned to small classes (13-17 students), regular classes (22- 25 students), or regular classes with a full-time aide. Krueger (1999) estimates the impact of reduced class size on test scores using a student's initial assignment to one of the three groups. He reports that students in smaller classes had a 0.133σ (0.033) increase in reading test scores and a 0.107σ (0.033) increase in math test scores, compared to students assigned to a regular class without an aide. In conducting a cost benefit analysis, Krueger (2003) assumes that since, class size reduced from about 22 to about 15 students, funds are allocated to create 7/15 = 47% more classes. Accordingly, the marginal cost per student for each year a student is in a small class is $3,501, or 47% of the nationwide total expenditure per student in 1997-1998. The average number of years spent in a small class was 2.3 years. Using this and an average initial age of 7, we get an IRR of 9.75%. All cost numbers are presented in 1998 dollars.

Charter Schools

Harlem Children's Zone: The Harlem Children's Zone (HCZ) is a 97-block area in central Harlem, New York that combines reform-minded charter schools with a web of community services designed to ensure that the social environment outside of school is positive and supportive for children from birth to college graduation. Dobbie and Fryer (2011) estimate the causal impact of attending the Promise Academy in the HCZ by exploiting the fact that HCZ charter schools are required to select students by lottery when the number of applicants exceeds the number of available slots for admission. In this scenario, the treatment group is composed of students who are lottery winners and the control group consists of students who are lottery losers. The two-stage-least-squares (2SLS) estimates for attending these charter schools during middle school are 0.229σ (0.037) in math scores and a 0.047σ (0.033) in reading scores. Similarly, the 2SLS estimates for elementary school imply

that attending Promise Academy charter schools for one year increases reading scores by 0.114σ (0.095) and math scores by 0.191σ (0.116) relative to the control group. Dobbie and Fryer (2011) state that the New York Department of Education provided every charter school, including the Promise Academy, $12,443 per student in 2008-2009. HCZ estimates add another $4,657 per student for in-school costs and approximately $2,172 per pupil for after-school and "wrap-around" programs. This implies that HCZ spends $19,272 per student per year. Using this number and adjusting for average number of years spent in treatment (1.24 years for middle school and 0.834 years for elementary school), we get an IRR of 10.84% and 11.92% for elementary and middle school, respectively. All cost numbers are presented in 2009 dollars.

Injecting Best Practices: Fryer (2014) examines the impact on student achievement of implementing a bundle of best practices from high-performing charter schools into low-performing, traditional public schools in Houston, Texas. Fryer uses a school-level randomized field experiment and quasi-experimental comparisons. Treatment schools implemented the following five practices: increased instructional time; replacement of principals and teachers who failed to adequately increase student achievement; implementation of daily high-dosage mathematics tutoring for fourth graders; use of data-driven curricula; and fostering a culture of high expectations. The intervention was done in 8 elementary schools and 9 middle and high schools. Fryer reported a yearly increase of 0.072σ (0.039) in reading test scores and an increase of 0.184σ (0.06) in math test scores for elementary school students over an average of 1.34 years spent in treatment. For middle and high schools, Fryer reports a yearly decrease of 0.012σ (0.022) in reading scores and an increase of 0.146σ (0.031) in math test scores, over an average of 1.31 years in treatment. The reported costs per student per year were $355 for elementary school students and $1,837 for secondary school students. Using an average initial age of 10 for elementary school and 14 for secondary school, we have an IRR of 35.11% and 18.41% for elementary and secondary schools, respectively. All cost numbers are presented in 2013 dollars.

SEED: SEED schools are five-day-a-week urban boarding schools that have an extended school day, provide extensive after-school tutoring, utilize data-driven curricula, and maintain a culture of high expectations. Curto and Fryer (2014) utilize the fact that when a SEED school is oversubscribed, it determines admission via a random lottery. Thus, the treatment group is composed of lottery winners and the control group consists of lottery losers. Curto and Fryer (2014) report that winning the lottery increases math achievement by 0.218σ (0.082) and reading achievement by 0.201σ (0.086). Using data from District of Columbia Public Schools (DCPS) and SEED schools' financial reports, Curto and Fryer report that SEED's cost per student per year in 2008-09 were $39,275. According to the National Center for Education statistics, the total expenditure per student in DCPS was $20,523 for the same year, giving us an incremental cost of attending a SEED school of around $18,752 per student per year. Using an average initial age of 13 and an average 2.33 years of being enrolled in SEED, we get an IRR of 8.64%. All cost numbers are presented in 2008 dollars.

Managed Professional Development

Success for All: Success for All is a school-level elementary school intervention that focuses on improving literacy outcomes for all students in order to improve overall student achievement. In 2007, it was used in 1,200 schools across the country (Borman et al., 2007). The program is designed to identify and address deficiencies in reading skills at a young age using a variety of instruction strategies, ranging from cooperative learning to data-driven instruction. Borman et al. (2007) use a cluster randomized trial design to evaluate the impacts of the Success for All model on student achievement. Thirty-five schools from eleven states volunteered and were randomly assigned to either the treatment or control group for a 3-year longitudinal study. Control schools implemented Success for All in grades 3-5, while treatment schools implemented Success for All in grades K-2. Comparisons were then made between the treated K-2 students and the untreated K-2 students. Borman et al. report a 0.09σ (0.06) increase in reading test scores. Implementing Success for All would cost schools $75,000 the first year, $35,000 the second year, and $25,000 the third year, for a total of $135,000. For the purpose of this evaluation, all participating schools received Success for All but in different grades. However, for a more realistic cost of implementing this program, we only consider the incremental cost for the treatment schools, which is roughly $746 per student per year, using 18 treatment schools and 1,085 treatment students across the 3 years. Using an initial age of 6, we get an IRR of 14.15%. All cost numbers are presented in 2007 dollars.

Curriculum

Enhanced Reading Opportunities: The US Department of Education initiated the Enhanced Reading Opportunities (ERO) study to evaluate supplemental literacy programs targeted at 9th graders whose reading levels were between two and five years below grade level. As part of the study, two cohorts of ninth grade students from 34 high schools and 10 school districts implemented one of two reading interventions: Reading Apprenticeship Academic Literacy (RAAL) and Xtreme Reading. Students were selected based on being two to five years below grade level on reading comprehension test scores, and were randomly assigned to enroll in an ERO class or not. Experienced English and social studies teachers volunteered to teach the ERO class for two years, and were provided training and technical assistance by the program's developers (Somers et al., 2010). Somers et al. (2010) find an increase of 0.11σ (0.037) in reading test scores and a 0.07σ (0.035) increase in math test scores as a result of the program. The average annual cost per student of implementing the programs was $1,931. Using an initial age of 15, we get an IRR of 22.04%. All cost numbers are presented in 2010 dollars.

Teacher Incentives

Talent Transfer Incentives: Glazerman et al. (2013) use a randomized experiment in 10 districts across the nation to investigate the impact of filling vacancies with high-achieving teachers through the Talent Transfer Initiative (TTI). In each district, the TTI offered teachers with consistently high value-added (ranking in the top 20 percent within their subject and grade) $20,000, paid over two years, to teach at low-achieving schools randomly assigned to treatment. Across the 10 districts included in the study, 165 teacher teams from 114 schools were randomly assigned to treatment or control spanning grades 3 through 8. The initiative began in 2009 with 7 districts (cohort 1) and 3 additional districts were added in 2010

(cohort 2). Each team consisted of focal teachers, who were the teachers that filled the vacancies, and non-focal teachers who constituted the rest of the team. Glazerman et al. (2013) report positive impacts on test scores for elementary school students as a result of the TTI. The cumulative effect of focal teachers in elementary school on cohort 1 is a 0.22σ (0.06) increase in math scores and a 0.25σ (0.05) increase in reading scores, which we divide by two to get the yearly effect. The sample of treatment students in cohort 1 is roughly 2,451, which is half of the sample size reported for grades 3 through 8 with unique student-focal teacher combinations. Glazerman et al. (2013) estimate the cost of implementing TTI was $36,382 per team, the majority of which included transfer stipends and retention stipends over the two years. Half of this cost multiplied by 87 teams gives a per student per year cost of $645.70. Using an average initial age of 10, we get an IRR of 27.80%. All cost numbers are presented in 2013 dollars.

High Dosage Tutoring

Experience Corps: This program trains older adults, aged 55 and above, to tutor and mentor elementary school children who are at risk of academic failure. Volunteers receive training focused on literacy and relationship building, as well as a stipend based on number of hours worked. Volunteers work with students one-on-one for about 15 hours a week. Morrow-Howell et al. (2009) use a randomized experiment across 23 schools in Boston, New York City, and Port Arthur, Texas to evaluate the effectiveness of this program. At the beginning of the school year in 2006, all students in need of reading assistance were referred to the Experience Corps program. All referred students were then randomly assigned to the treatment or control group. The EC program tutored 430 students in total, with 451 students in the control group. Morrow-Howell et al. report an average increase of 0.075σ (0.067) on reading test scores. To calculate cost per student per year, we first calculated average cost per tutor. Based on its IRS 990 form, Experience Corps had a total cost of $1,343,936 in 2009, when the program had 2,000 tutors (Morrow-Howell et al., 2009). This gives us a per tutor cost of $671. With 505 tutors in the evaluation and 430 students tutored, we have a per student per year cost of $788. Using an average initial age of 8, we get an IRR of 13.81%. All cost numbers are presented in 2009 dollars, corresponding to Morrow-Howell et al., (2009) and the year of the IRS 990 form.

Appendix D: Survey Instruments

I. Principal Survey (Summer 2015)

HISD-EdLabs Principal Survey 2015

The following questions are for a research study and are designed to help us understand your thoughts about leading a school and managing your staff, as well as to gather information about your beliefs, attitudes, time use, and skills. These survey results are very important to help us understand how management practices may impact teaching and learning. We expect this survey will require about ninety minutes to complete, and your participation is sincerely appreciated but not required. As a token of our appreciation, you will receive a $50 Amazon gift card for completing the entire survey. The gift card will be delivered to the email account that you specify in the survey, and will be delivered within two weeks of completion of the survey. You may choose not to answer one or more questions. If you choose to stop at any time, this will involve no penalty or loss of benefits to which you are otherwise entitled. Failure to complete at least 75% of the survey will make you ineligible for the $50 Amazon gift card, however. Information collected will follow these guidelines and procedures. All data collected will be used anonymously. No teachers, school administrators, or district administrators will be able to see your answers, and results from this survey will only be presented in aggregate form. No one outside of the research team will know whether you have completed this survey. Data containing identifiable information will be destroyed within one year of the conclusion of the study. Data collected will be handled in accordance with EdLabs Data Security Policy, which includes transferring the information from Qualtrics to a computer with no network access in a secure data room via an encrypted portable storage device. If you have questions regarding this survey, you may contact [Edlabs Project Manager] at Harvard EdLabs at XX@edlabs.harvard.edu. The address of the Harvard EdLabs office is XX, and the phone number is XX. You should reach out to Harvard EdLabs: if you have questions, concerns, or complaints, if you would like to talk to the research team, if you think the research has hurt you, or if you wish to withdraw from the study. This research has been reviewed by the Committee on the Use of Human Subjects in Research at Harvard University. They can be reached at XX for any of the following: if your questions, concerns, or complaints are not being answered by the research team, if you cannot reach the research team, if you want to talk to someone besides the research team, or if you have questions about your rights as a research participant. By completing the following survey, you give permission to include your responses in this research. Please print a copy of this form for your records.

Name

First name:

Last name:

What school did you lead during the 2014-15 school year?

- ❍ XX
- ❍ YY
- ❍ Etc.

Please tell us the number of years you have spent in each of the following categories.

Years in current school:

Years as a principal:

Years in HISD:

Years working in education:

Please complete the following questions about your own education. Leave blank any non-applicable lines.

Undergraduate Degree (e.g., BA, BS):

Major/Concentration:

College/University:

Please complete the following questions about your first graduate degree. Leave blank any non-applicable lines.

Graduate Degree #1:

Major/Concentration:

College/University:

Please complete the following questions about your second graduate degree. Leave blank any non-applicable lines.

Graduate Degree #2:

Major/Concentration:

College/University:

To the best of your memory, please tell us your score on the SAT and/or ACT from when you took it in high school.

SAT Math

SAT Verbal

ACT English

ACT Math

ACT Reading

ACT Science

Please answer each of the following questions as thoroughly as possible.

When problems occur within your school, how do they typically get exposed and fixed? Describe the process for a recent problem that you faced. Who within the school gets involved in changing or

improving process? How do the different staff groups get involved in this? Does the staff ever suggest process improvements?

What kind of main indicators do you use to track school performance? What sources of information are used to inform this tracking? How frequently are these measured? Who gets to see this performance data? If someone were to walk through your school, how could he tell how it was doing against these main indicators?

How often do you review school performance --formally or informally-- with teachers and staff? Describe the steps you go through in a process review. Who is involved in these meetings? Who gets to see the results of this review? What sort of follow-up plan would you leave these meetings with? Is there an individual performance plan?

What types of targets are set for the school to improve student outcomes? Which staff levels are held accountable to achieve these stated goals? How much are these targets determined by external factors? Describe the goals that are not externally set for the school.

What kind of time scale are you looking at with your targets? Which goals receive the most emphasis? Are the long-term and short-term goals set independently? Could you meet all your short-run goals but miss your long-run goals?

How tough are your targets? How pushed are you by the targets? On average, how often would you say that you and your school meet its targets? How are your targets benchmarked? Do you feel that on targets all departments/ areas receive the same degree of difficulty? Do some departments/ areas get easier targets?

If someone asked one of your staff members directly about individual targets, what would they say? Does anyone complain that the targets are too complex? Could every staff member employed by the school tell me what they are responsible for and how it will be assessed? How do people know about their own performance compared to other people's performance?

If you had a teacher who was struggling or who could not do his/ her job, what would you do? Please describe a recent example (please do not name the teacher). How long is under-performance tolerated? How difficult is it to terminate a teacher? Do you find staff members/ teachers who lead a sort of charmed life? Do some individuals always just manage to avoid being fired?

Describe your career progression/ promotion system. How do you identify and develop your star performers? What types of professional development opportunities are provided? How are these opportunities personalized to meet individual teacher needs? How do you make decisions about promotion/ progression and additional opportunities within the school, such as performance, tenure, other? Are better performers likely to be promoted faster, or are promotions given on the basis of tenure/ seniority?

If you had a top performing teacher who wanted to leave, what would the school do? Please give an example of a star performer being persuaded to stay after wanting to leave. Please give an example of a star performer who left the school without anyone trying to keep him. (Please do not name the teachers in either case.)

Think about your day on Tuesday, May 19th. In the "Morning" list below, please indicate when you arrived on your campus by selecting "BEGIN" from the drop-down menu at the appropriate time. From that point on, please think about your time spent on campus and for each 15-minute increment, indicate what activity you were doing by choosing the appropriate activity from the drop-down menu for each time. Feel free to reference your calendar or any other resource that will help you accurately recall your schedule that day. If you were off-campus or were hosting district personnel for the majority of the day on Tuesday, May 19th, please pick another school day that week during which you were on campus the majority of the time and complete this task for that day. [Principals see this list with each time slot from 5:00am to 8:45pm in 15-minute increments]

- Meeting with central HISD employees (e.g., SSO, TDS, etc.)
- Meeting with school administrative/leadership team
- Meeting with teacher (one-on-one)
- Meeting with teachers (group)
- Meeting with other school-based staff (e.g., counselor, social worker, etc.)
- Meeting with parents (in-person)
- Phone calls with parents
- Observing whole-school transitions and activities (e.g., breakfast, lunch, dismissal, etc.)
- Leading whole-school events (e.g., assemblies, morning meetings, etc.)
- Observing classroom instruction
- Completing administrative paperwork
- Reviewing teacher lesson plans and/or curriculum plans
- Teaching a class
- Leading professional development for teachers
- Attending professional development for self
- Handling facilities issues
- Examining student performance data
- Reading and/or responding to emails
- Attending student extracurricular programming
- Dealing with school technology systems
- Addressing student discipline problems
- Planning whole-school events
- Updating bulletin boards/announcements/newsletters
- Dealing with vendors (e.g., food services, transportation, etc.)
- Lunch break
- Other

For the following questions, please consider your experiences as a school leader during the 2014-15 school year.

How would you describe the effort of the average teacher in your school during the 2014-15 school year?

- ❍ Significantly less effort than the 2013-14 school year
- ❍ Slightly less effort than the 2013-14 school year
- ❍ About the same effort as the 2013-14 school year
- ❍ Slightly more effort than the 2013-14 school year
- ❍ Significantly more effort than the 2013-14 school year

How would you describe the effectiveness of the average teacher in your school during the 2014-15 school year?

- ❍ Significantly less effective than the 2013-14 school year
- ❍ Slightly less effective than the 2013-14 school year
- ❍ About as effective as the 2013-14 school year
- ❍ Slightly more effective than the 2013-14 school year
- ❍ Significantly more effective than the 2013-14 school year

To the best of your memory, how many times did you (or another member of your leadership team) observe each of the teachers listed below while they delivered instruction during the 2014-15 school year? [Each principal sees the names of four teachers employed at their school]

Teacher 1
Teacher 2
Teacher 3
Teacher 4

Approximately what percentage of teachers on your campus submitted weekly lesson plans during the 2014-15 school year?

How frequently did you (or another member of your leadership team) discuss student performance data with each teacher on your campus this year?

- ❍ Daily
- ❍ At least weekly
- ❍ At least monthly
- ❍ After each snapshot assessment/DLA
- ❍ After some snapshot assessments/DLA, but not all
- ❍ Once
- ❍ Never

If you had the opportunity to build your staff from scratch, what percentage of the teachers that currently work at your school would you re-hire?

- ❍ 0%
- ❍ 1-25%
- ❍ 26-50%
- ❍ 51-75%
- ❍ 76-99%
- ❍ 100%

How effective was any training you received for the 2014-15 school year compared to any training you received for the 2013-14 school year?

- ❍ Significantly less effective than the 2013-14 school year
- ❍ Slightly less effective than the 2013-14 school year
- ❍ About as effective as the 2013-14 school year
- ❍ Slightly more effective than the 2013-14 school year
- ❍ Significantly more effective than the 2013-14 school year

From your perspective as a school leader, please rate the performance of your School Support Officer (SSO) or Lead Principal during the 2014-15 school year on a scale of 1 to 10, with 1 being very poor and 10 being excellent.

- ❍ 1
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7
- ❍ 8
- ❍ 9
- ❍ 10

From your perspective as a school leader, please rate the performance of the Chief School Leadership Officer during the 2014-15 school year on a scale of 1 to 10, with 1 being very poor and 10 being excellent.

- ❍ 1
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7
- ❍ 8
- ❍ 9
- ❍ 10

Please describe any supports that you needed during the 2014-15 school year that you did not receive.

For each pair of statements in the following 4 questions, please select the answer that indicates which statement is closer to your opinion and whether the statement is slightly closer or much closer to your opinion. In some cases you may find that you believe both statements; in other cases you may believe neither one. Even when you feel this way about a pair of statements, select the one statement which is more nearly true in your opinion. Try to consider each pair of statements separately when making your choices; do not be influenced by your previous choices.

A. What happens to me is my own doing. B. Sometimes I feel that I don't have enough control over the direction my life is taking.

- ❍ Statement A is much closer to my opinion.
- ❍ Statement A is slightly closer to my opinion.
- ❍ Statement B is slightly closer to my opinion.
- ❍ Statement B is much closer to my opinion.

A. When I make plans, I am almost certain that I can make them work. B. It is not always wise to plan too far ahead, because many things turn out to be a matter of good or bad fortune anyway.

- ❍ Statement A is much closer to my opinion.
- ❍ Statement A is slightly closer to my opinion.
- ❍ Statement B is slightly closer to my opinion.
- ❍ Statement B is much closer to my opinion.

A. In my case, getting what I want has little or nothing to do with luck. B. Many times we might just as well decide what to do by flipping a coin.

- ❍ Statement A is much closer to my opinion.
- ❍ Statement A is slightly closer to my opinion.
- ❍ Statement B is slightly closer to my opinion.
- ❍ Statement B is much closer to my opinion.

A. Many times I feel that I have little influence over the things that happen to me. B. It is impossible for me to believe that chance or luck plays an important role in my life.

- ❍ Statement A is much closer to my opinion.
- ❍ Statement A is slightly closer to my opinion.
- ❍ Statement B is slightly closer to my opinion.
- ❍ Statement B is much closer to my opinion.

The following 12 questions contain statements that may or may not apply to you. When responding, think of how you compare to most people - not just the people you know well, but most people in the world. There are no right or wrong answers, so just answer honestly.
New ideas and projects sometimes distract me from previous ones.

- ❍ Very much like me
- ❍ Mostly like me
- ❍ Somewhat like me
- ❍ Not much like me
- ❍ Not like me at all

Setbacks don't discourage me.

- ❍ Very much like me
- ❍ Mostly like me
- ❍ Somewhat like me
- ❍ Not much like me
- ❍ Not like me at all

I have been obsessed with a certain idea or project for a short time but later lost interest.

- ❍ Very much like me
- ❍ Mostly like me
- ❍ Somewhat like me
- ❍ Not much like me
- ❍ Not like me at all

I am a hard worker.

- ❍ Very much like me
- ❍ Mostly like me
- ❍ Somewhat like me
- ❍ Not much like me
- ❍ Not like me at all

I often set a goal but later choose to pursue a different one.

- ❍ Very much like me
- ❍ Mostly like me
- ❍ Somewhat like me
- ❍ Not much like me
- ❍ Not like me at all

I have difficulty maintaining my focus on projects that take more than a few months to complete.

- ❍ Very much like me
- ❍ Mostly like me
- ❍ Somewhat like me
- ❍ Not much like me
- ❍ Not like me at all

I finish whatever I begin.
- ❍ Very much like me
- ❍ Mostly like me
- ❍ Somewhat like me
- ❍ Not much like me
- ❍ Not like me at all

I am diligent.
- ❍ Very much like me
- ❍ Mostly like me
- ❍ Somewhat like me
- ❍ Not much like me
- ❍ Not like me at all

I have overcome setbacks to conquer an important challenge.
- ❍ Very much like me
- ❍ Mostly like me
- ❍ Somewhat like me
- ❍ Not much like me
- ❍ Not like me at all

My interests change from year to year.
- ❍ Very much like me
- ❍ Mostly like me
- ❍ Somewhat like me
- ❍ Not much like me
- ❍ Not like me at all

I have achieved a goal that took years of work.
- ❍ Very much like me
- ❍ Mostly like me
- ❍ Somewhat like me
- ❍ Not much like me
- ❍ Not like me at all

I become interested in new pursuits every few months.
- ❍ Very much like me
- ❍ Mostly like me
- ❍ Somewhat like me
- ❍ Not much like me
- ❍ Not like me at all

For the following 13 questions, solve each problem and select the answer choice that is best from the choices given. You may use pen and paper to help you solve any questions.

The result when a number is divided by 2 is equal to the result when that same number is divided by 4. What is that number?

- -4
- -2
- 0
- 2
- 4

If $10+x$ is 5 more than 10, what is the value of $2x$?

- -5
- 5
- 10
- 25
- 50

In a certain store, the regular price of a refrigerator is $600. How much money is saved by buying this refrigerator at 20 percent off the regular price rather than buying it on sale at 10 percent off the regular price with an addition discount of 10 percent off the sale price?

- $6
- $12
- $24
- $54
- $60

A total of 120,000 votes were cast for two opposing candidates, Garcia and Perez. If Garcia won by a ratio of 5 to 3, what was the number of votes cast for Perez?

- 15,000
- 30,000
- 45,000
- 75,000
- 80,000

If a positive integer n is picked at random from the positive integers less than or equal to 10, what is the probability that $5n + 3 \leq 14$?

- 0
- 1/10
- 1/5
- 3/10
- 2/5

If j, k, and n are consecutive integers such that $0 < j < k < n$ and the units (ones) digit of the product jn is 9, what is the units digit of k?

- 0
- 1
- 2
- 3
- 4

The average (arithmetic mean) of t and y is 15, and the average of w and x is 15. What is the average of t, w, x, and y?

- 7.5
- 15
- 22.5
- 30
- 60

"All of Kay's brothers can swim." If the statement above is true, which of the following must also be true?"

- If Fred cannot swim, then he is not Kay's brother.
- If Dave can swim, then he is not Kay's brother.
- If Walt can swim, then he is Kay's brother.
- If Pete is Kay's brother, then he cannot swim.
- If Mark is not Kay's brother, then he cannot swim.

Each of the following is equivalent to $\frac{a}{b}(bc + k)$ EXCEPT

- $a(\frac{c+k}{b})$
- $a(c + \frac{k}{b})$
- $\frac{a}{b}(k + bc)$
- $ac + \frac{ak}{b}$
- $\frac{abc+ak}{b}$

On Wednesday Heather ran 3 miles in 30 minutes. If she ran for 45 minutes at this rate on Thursday, how far did Heather run on Thursday?

- 3.5 miles
- 4 miles
- 4.5 miles
- 5 miles
- 5.5 miles

Q70 Let F(x) be defined as $x+1/x$ for all nonzero integers. If F(x) = t where t is an integer, which of the following is a possible value of t?

- 1
- 0
- -1
- -2
- -3

Q71 10, 18, 4, 15, 3, 21, x If x is the median of the 7 numbers listed above, which of the following could be the value of x?

- 5
- 8
- 9
- 14
- 16

If x and y are integers, $7 < y < 16$, and $x/y=2/5$ how many possible values are there for x?

- One
- Two
- Three
- Four
- Five

For the following 2 questions, solve each problem and enter your solution into the text box provided. You may use pen and paper to help you solve any questions.

What is the product of the smallest prime number that is greater than 50 and the greatest prime number that is less than 50?

Three more than twice a number is equal to 4. What is the number?

Thank you for completing the EdLabs Principal Survey! Please enter your email address below to receive your $50 Amazon gift card as a token of our appreciation. The gift card will be sent to you via email with a validation code for use. Your email address will not be shared with any outside vendors or your school district and will ONLY be used to send you the Amazon gift card. You will not receive any other emails or spam as a result of entering your email address. You will receive your gift card via email within 1-2 weeks of completing the survey.If you have not received your $50 Amazon gift card within the next 2 weeks, please email [Edlabs Project Manager] at

XX@edlabs.harvard.edu so that we can make sure you receive your gift card as soon as possible. Thanks again!

Email address (e.g., "principal@gmail.com" or "lastname1@houstonisd.org")

II. Principal Shadowing Materials (Summer 2016)

Principal Shadowing Instructions

An important part of the shadowing program is to understand how principals use their time on a typical day. There are many activities that a principal does each day that are not readily visible to a casual observer. Therefore, throughout your observations, you are expected to keep a record of a principal's activities. You will keep a record beginning when the principal arrives on campus at the start of the school day and concluding when the principal leaves campus at the end of the school day.

Throughout the day, you are asked to keep a running diary of activities. The diary is meant to capture two important pieces of information: (1) the principal's activity at the time; and (2) any resources/references (broadly defined) that the principal is using at the time.

This information is important for two reasons. One, it will provide you a clear and thorough record of the daily activity of a single leader. It will be an important reference document for you as you go back over your experience after the fact. Secondly, the information collected from all aspiring leaders during their shadowing experiences will be collated and shared – anonymously – with the full cohort in order to allow for a rich insight into the daily roles and responsibilities of

The goal of this exercise is to compile as complete a picture of a principal's time use as possible while being minimally invasive so that you can remain as focused as possible during your observation.

Logistics

You may choose to collect this time use information in one of two ways: (1) paper and pen, or (2) Google Sheets iPad application. You will be provided with a paper template; if you wish to use Google Sheets instead, you may request access to it by emailing [Edlabs Project Manager at XX@edlabs.harvard.edu]. The method of collection is up to you and your own comfort. Regardless of your means of collection, you will be asked to share the information you collect with the Office of School Leadership for the benefit of the cohort, so if you choose to hand-write your records, please make sure your writing is legible enough!

Activity

When recording the activity of the principal at a given time, record enough information about the activity that you will later be able to recall what the activity was. At the end of the day of shadowing, you will need to select from a provided list of activities.

Note that within the activities listed, there are small but important distinctions that you should pay attention to and be sure to accurately capture. For instance, when a principal is meeting with teachers, it is important to note if the principal is meeting with a group of teachers together or one-on-one with a single teacher.

Resource Used

We define "resource" broadly for these purposes. This would include the more traditional student performance data (e.g., snapshot assessment data, student grades) as well as other information that might not be typically thought of as data (e.g., student behavior reports or teacher lesson plans). This should mostly be thought of as any information that the principal refers to in the course of an activity. If you are unsure whether to list something as "Resource Used", it is better to list it than to not. In discussions with the aspiring leaders cohort after the shadowing is completed, you will have the opportunity to decide whether to ultimately include something you initially listed in your data collection.

Time Use Diary Activity Codes

Activity	**Code**
Meeting with central HISD employees (e.g., SSO, TDS, etc.)	1
Meeting with school administrative/leadership team	2
Meeting with teachers (one-on-one)	3
Meeting with teachers (groups)	4
Meeting with other school-based staff (e.g., counselor, social worker, etc.)	5
Meeting with parents (in-person)	6
Phone calls with parents	7
Observing whole-school transitions and activities (e.g., breakfast, lunch, dismissal)	8
Leading whole-school events (assemblies, morning meetings, etc.)	9
Observing classroom instruction	10
Completing administrative paperwork	11
Reviewing teacher lesson plans and/or curriculum plans	12
Teaching a class	13
Leading professional development for teachers	14
Attending professional development for self	15
Handling facilities issues	16
Examining student performance data	17
Reading and/or responding to emails	18
Attending student extracurricular programming	19
Dealing with school technology systems	20
Addressing student discipline problems	21
Planning whole-school events	22
Updating bulletin boards/announcements/newsletters	23
Dealing with vendors (e.g., food services, transportation)	24
Lunch break	25
Other (specify):_____	26
Other (specify):_____	27
Other (specify):_____	28

APPENDIX REFERENCES

Bambrick-Santoyo, Paul (2012), Leverage Leadership: A Practical Guide to Building Exceptional Schools. San Francisco: Jossey-Bass.

Bettinger, Eric (2012), "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores." *Review of Economics and Statistics* 94(3): 686-698.

Bloom, Nicholas, Christos Genakos, Raffaelaa Sadun, and John Van Reenen (2012), "Management Practices Across Firms and Countries," *Academy of Management Perspectives* 26(1): 12-33.

Borman, Geoffrey, Robert Slavin, Alan Cheung, Anne Chamberlain, Nancy Madden, and Bette Chambers (2007), "Final Reading Outcomes of the National Randomized Field Trial of Success for All." *American Education Research Journal*, 44(3): 701-731.

Chetty, Raj, John Friedman, and Jonah Rockoff (2014) "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review* 104(9): 2633-2679.

Cohen, Rachel M. (2015), "The True Cost of Teach For America's Impact on Urban Schools," *The American Prospect*, January 5, 2015. Accessed November 1, 2016. http://prospect.org/article/true-cost-teach-americas-impact-urban-schools

Curto, Vilsa, and Roland Fryer (2014), "The Potential of Urban Boarding Schools for the Poor." *Journal of Labor Economics*, 32(1): 65-93.

Dobbie, Will and Roland G. Fryer (2011), "Are High Quality Schools Enough to Increase Achievement Among the Poor? Evidence From the Harlem Children's Zone", *American Economic Journal: Applied Economics* 3(3): 158-187.

Duckworth, Angela, Christopher Peterson, Michael D. Matthews and Dennis R. Kelly (2007), "Grit: Perseverance and Passion for Long-Term Goals," *Journal of Personality and Social Psychology* 92(6): 1087-1101.

Fryer, Roland G. (2011), "Financial Incentives and Student Achievement: Evidence from Randomized Trials," *Quarterly Journal of Economics* 126(4): 1755-1798.

Fryer, Roland G. (2014), "Injecting Charter School Best Practices Into Traditional Public Schools: Evidence from Field Experiments," *The Quarterly Journal of Economics* 129(3): 1355-1407.

Fryer, Roland G. *(forthcoming)*, "The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments." In: Handbook of Field Experiments.

Glazerman, Steven, Danial Mayer, and Paul Decker (2006), "Alternative Routes to Teaching: The Impacts of Teach for America on Student Achievement and Other Outcomes," *Journal of Policy Analysis and Management* 25(1): 75-96.

Glazerman, Steven, Ali Protik, Bing-ru The, Julie Bruch, Jeffrey Max (2013), "Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment." Washington, D.C.: National Center for Education Evaluation and Regional Assistance.

Krueger, Alan B. (1999), "Experimental Estimates of Education Production Functions," *The Quarterly Journal of Economics* 114(2): 497-532.

Krueger, Alan B. (2003), "Economic Considerations and Class Size," *The Economic Journal* 113(485): F34-F63.

Morrow-Howell, Nancy, Melissa Jonson-Reid, Stacey McCrary, YungSoo Lee, and Ed Spitznagel (2009), "Evaluation of Experience Corps," Washington University in St. Louis: Center for Social Development.

Puma, Michael, Stephen Bell, Ronna Cook, and Camilla Heid (2010), "Head Start Impact Study Final Report." Washington, D.C.: U.S. Department of Health and Human Services, Administration for Children and Families.

Rotter, Julian B. (1966) "Generalized Expectancies for Internal Versus External Control of Reinforcement," *Psychological Monographs: General and Applied* 80(1): 1-28.

Somers, Marie-Andree, William Corrin James J. Kemple, Elizabeth Nelson, Susan Sepanik, et al. (2010), "The Enhanced Reading Opportunities Study Final Report", U.S. Department of Education, Institute of Education Sciences, Washington, DC.

Valecha, Gopal K. and Ostrom, Thomas M. (1974), "An Abbreviated Measure of Internal-External Locus of Control," *Journal of Personality Assessment* 38(4): 369-376.