Online Appendix for

The 'Pupil' Factory: Specialization and the Production of Human Capital in Schools

By Roland G. Fryer

[NOT FOR PUBLICATION]

Appendix Table of Contents:

Appendix Table 1: The Effect of Treatment on Proof of Treatment Survey Outcomes (Lee Bounds)

	ITT	Lee Bound
	(1)	(2)
Teacher does not teach both math and reading	0.345***	0.345***
	(0.042)	(0.042)
Observations	711	711
Teacher teaches 3 or fewer subjects	0.332***	0.323***
	(0.040)	(0.038)
Observations	666	663
Teacher teaches 2 or fewer subjects	0.259***	0.249***
	(0.037)	(0.035)
Observations	666	663
Teacher self-reports being departmentalized	0.360***	0.333***
	(0.049)	(0.046)
Observations	649	641
Number of subjects taught	-0.770***	-0.739***
	(0.100)	(0.093)
Observations	666	663
Number of grades taught	-0.176	-0.176
	(0.129)	(0.129)
Observations	693	689

Notes: This table presents bounded estimates to provide a conservative bound on the true treatment effects under the assumption that there are differential rates of response to the survey. Column (1) presents the ITT estimates from Table 6. Column (2) provides the Lee bounded estimates. Standard errors, clustered by school, are located in parentheses. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Appendix Table 2: The Effect of Treatment on High-Stakes Math and Reading Test Scores

	ITT			2SLS (Ever)			2SLS (Years)		
	2014	2015	Pooled	2014	2015	Pooled	2014	2015	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Math	-0.056**	-0.058*	-0.057**	-0.062**	-0.067**	-0.064***	-0.067**	-0.037**	-0.047***
	(0.027)	(0.029)	(0.022)	(0.030)	(0.033)	(0.025)	(0.032)	(0.018)	(0.018)
Reading	-0.067***	-0.031	-0.049**	-0.075***	-0.036	-0.056**	-0.080***	-0.020	-0.041**
	(0.024)	(0.027)	(0.020)	(0.026)	(0.030)	(0.022)	(0.028)	(0.017)	(0.017)
Observations	10,462	10,360	20,822	10,462	10,360	20,822	10,462	10,360	20,822
First Stage Coefficient				0.896***	0.858***	0.877***	0.833***	1.552***	1.191***
				(0.009)	(0.012)	(0.008)	(0.009)	(0.025)	(0.042)

Notes: This table presents estimates of being enrolled in or attending a treatment school on STAAR math and reading test scores. Here treatment is defined as attending a treatment school as the last school in 2012-13. The sample is restricted each year to those students who are attending grades 3 through 5 and have both valid math and reading test scores. Columns (1), (2), and (3) report Intent-to-Treat (ITT) estimates. Columns (4), (5), and (6) report 2SLS estimates and use treatment assignment as an instrument for having ever attended a treatment school during years of treatment. Columns (7), (8), and (9) report 2SLS estimates and use treatment assignemtn to instrument for the number of years spent in a treatment school. Columns (1), (4), and (7) use 2013-2014 scores as the outcome variable. Columns (2), (5), and (8) use 2014-2015 scores as the outcome variable. Columns (3), (6), and (9) use scores from both 2013-2014 and 2014-2015 as the outcome variable. The depended variable in all specifications is state test score, standardized to have a mean of zero and standard deviation one by grade and year. All specifications adjust for the student-level demographic variables summarized in Table 2, student-level math and reading scores (3 years prior to 2013-2014) and their squares, and indicators for taking a Spanish baseline test. All specifications have grade-by-year and matched-pair fixed effects. The last row provides the first stage coefficient of instrumenting the 2SLS *Ever* or *Years* variable with ITT treatment assignment. This number can be used to scale the ITT estimate into other estimates. Standard errors, reported in parentheses, are clustered at the school-year level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively.

Appendix Table 3: The Effect of Treatment on Low-Stakes Math and Reading Test Scores

	ITT			2SLS (Ever)			2SLS (Years)		
	2014	2015	Pooled	2014	2015	Pooled	2014	2015	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: All Grades									
Math	-0.044	-0.032	-0.038*	-0.049*	-0.038	-0.044*	-0.053*	-0.021	-0.033*
	(0.027)	(0.034)	(0.023)	(0.030)	(0.040)	(0.026)	(0.032)	(0.022)	(0.020)
Reading	-0.065**	-0.058*	-0.062***	-0.073***	-0.068*	-0.071***	-0.079***	-0.038*	-0.053***
	(0.025)	(0.032)	(0.021)	(0.027)	(0.037)	(0.024)	(0.030)	(0.021)	(0.018)
Observations	18618	16849	35467	18618	16849	35467	18618	16849	35467
Panel B: Grades 3 - 5									
Math	-0.057**	-0.039	-0.047**	-0.063**	-0.046	-0.053**	-0.068**	-0.025	-0.040**
	(0.023)	(0.029)	(0.020)	(0.026)	(0.033)	(0.023)	(0.028)	(0.018)	(0.017)
Reading	-0.071**	-0.060**	-0.066***	-0.080***	-0.069**	-0.075***	-0.086***	-0.038**	-0.055***
	(0.028)	(0.029)	(0.021)	(0.030)	(0.032)	(0.024)	(0.033)	(0.018)	(0.018)
Observations	10974	10281	21255	10974	10281	21255	10974	10281	21255
Panel C: Grades 1 - 2									
Math	-0.039	-0.017	-0.031	-0.044	-0.020	-0.036	-0.048	-0.011	-0.011
	(0.046)	(0.050)	(0.037)	(0.051)	(0.059)	(0.043)	(0.056)	(0.033)	(0.033)
Reading	-0.071*	-0.056	-0.065*	-0.081**	-0.066	-0.075**	-0.088**	-0.037	-0.037
	(0.036)	(0.051)	(0.033)	(0.041)	(0.059)	(0.038)	(0.044)	(0.034)	(0.034)
Observations	7644	14212	14,212	7644	14212	14,212	7644	14212	14,212

Notes: This table presents estimates of being enrolled in a treatment school on low-stakes math and reading test scores. In all panels the dependent variable is a nationally-normed low-stakes test score, standardized to have a mean of zero and standard deviation one by grade and year. In the 2013-2014 school year, HISD administered the Stanford 10 test. In the 2014-2015 school year, HISD administered the Iowa Test of Basic Skills. The sample and specification is identical to that used in Table 4. All specifications adjust for the student level demographic variables summarized in Table 2, student level math and reading scores (3 years prior to 2013-2014) and their squares, and indicator variables for taking a Spanish baseline test. All specifications have grade-by-year and matched pair fixed effects. Standard errors, reported in parentheses, are clustered at the school-year level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively.

Appendix Table 4: The Effect of Treatment on Student Test Scores (2SLS - Degree of Specialization)

	High-Stakes		Low-Stakes	
	2014	2015	2014	2015
	(1)	(2)	(3)	(4)
Full Sample	-0.976***	-1.026	-0.441**	-0.809*
	(0.367)	(0.724)	(0.185)	(0.422)
N	10,133	10,033	17,778	16,063
Grade 1	—		-0.004	-0.829**
			(0.230)	(0.405)
N			3,643	2,778
Grade 2	—		-0.424**	-0.004
			(0.192)	(1.027)
N			3,504	3,342
Grade 3	-0.853*	-0.925	-0.995*	-0.667
	(0.488)	(0.759)	(0.551)	(0.687)
N	3,397	3,136	3,515	3,116
Grade 4	-0.469	-2.073*	-0.674**	-2.254**
	(0.326)	(1.095)	(0.327)	(1.044)
N	3,546	3,512	3,741	3,481
Grade 5	-1.819	0.082	-1.213	-0.451
	(1.449)	(1.288)	(0.862)	(1.268)
N	3,190	3,385	3,375	3,346

Notes: This table reports estimates of the effect of increasing specialization on high- and low-stakes test scores. This table presents estimates of increasing the degree of specialization on high- and low-stakes test scores. High-stakes test scores are summed math and reading STAAR scores and low-stakes scores are summed math and reading Stanford 10 scores (in year 1) or ITBS scores (in year 2). Here treatment is defined as attending a treatment school as the last school in 2012-13. The sample is restricted each year to those students who are attending grades 3 through 5 (for high-stakes exams) and grades 1 through 5 (for low-stakes exams) and have both valid math and reading test scores.Here treatment is defined as attending a treatment school as the last school in 2012-13. All specifications use treatment assignment as an instrument for the percent of potential specialization being utilized by a school. Columns (1) and (3) use 2013-2014 scores as the outcome variable. Columns (2) and (4) use 2014-2015 scores as the outcome variable. The dependent variable in all specifications is the sum of standardized math and reading test scores, (standardized across the district to have a mean of zero and standard deviation one by grade and year). All specifications adjust for the student-level demographic variables summarized in Table 2, student-level math and reading scores (3 years prior to 2013-2014) and their squares, and indicators for taking a Spanish baseline test. All specifications have grade, year, and matched-pair fixed effects. Standard errors, reported in parentheses, are clustered at the school-year level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively.

Appendix Table 5: The Effect of Treatment on Student Attendance and Behavior

	ITT			2SLS (Ever)			2SLS (Years)		
	2014	2015	Pooled	2014	2015	Pooled	2014	2015	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Attendance Rate	-0.002***	-0.002**	-0.002***	-0.002***	-0.003**	-0.003***	-0.003***	-0.002**	-0.002***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Control Mean	0.968	0.969	0.968						
Disciplinary Action	0.007	0.006	0.007*	0.008	0.007	0.008*	0.009	0.004	0.006*
	(0.005)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.006)	(0.003)	(0.004)
Control Mean	0.059	0.045	0.053						
Observations	20,032	18,109	38,141	20,032	18,109	38,141	20,032	18,109	38,141
First Stage Coefficient				0.884***	0.843***	0.865***	0.797***	1.496***	1.129***
				(0.009)	(0.012)	(0.008)	(0.010)	(0.027)	(0.041)

Notes: This table presents estimates of being enrolled in or attending a treatment school on student attendance and behavior. Here treatment is defined as attending a treatment school as the last school in 2012-13. The sample is restricted each year to those students who are attending grades 1 through 5. Columns (1), (2), and (3) report Intent-to-Treat (ITT) estimates. Columns (4), (5), and (6) report 2SLS estimates and use treatment assignment as an instrument for having ever attended a treatment school during years of treatment. Columns (7), (8), and (9) report 2SLS estimates and use treatment assignment to instrument for the number of years spent in a treatment school. Columns (1), (4), and (7) use 2013-2014 measures as the outcome variable. Columns (2), (5), and (8) use 2014-2015 measures as the outcome variable. Columns (3), (6), and (9) use measures from both 2013-2014 and 2014-2015 as the outcome variable. The dependent variable is either student attendance rates (between 0 and 1) or an indicator variable for committing a behavioral infraction that led to at least a suspension. All specifications adjust for the student-level demographic variables summarized in Table 2, student-level math and reading scores (3 years prior to 2013-2014) and their squares, and indicators for taking a Spanish baseline test, as well as either student attendance or behavior in the year before treatment. All specifications have grade-by-year and matched-pair fixed effects. The final row provides the first stage coefficient of instrumenting the 2SLS *Ever* or *Years* variable with ITT treatment assignment. This number can be used to scale the ITT estimate into other estimates. Standard errors, reported in parentheses, are clustered at the school-year level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively.

Appendix Table 6: Sensitivity Analysis or Extension of the Basic Model, 2SLS (Ever)

	High-Stakes	*p-value*	Obs	Low-Stakes	*p-value*	Obs
	(1)	(2)	(3)	(4)	(5)	(6)
Full Sample (pooled)	-0.121***		20,822	-0.115**		35,467
	(0.043)			(0.047)		
Panel A: Demographics						
Special Education: Yes	-0.332***	0.035	688	-0.248***	0.087	1,424
	(0.100)			(0.077)		
Special Education: No	-0.114***		19,346	-0.109**		32,633
	(0.044)			(0.049)		
Gifted: Yes	-0.043	0.023	3,603	-0.073	0.204	4,220
	(0.066)			(0.057)		
Gifted: No	-0.153***		16,431	-0.131***		29,837
	(0.043)			(0.050)		
LEP: Yes	-0.085	0.542	8,856	-0.046	0.124	15,406
	(0.076)			(0.081)		
LEP: No	-0.133***		11,178	-0.167***		18,651
	(0.039)			(0.041)		
LEP: Yes (Grade 1-3)	-0.119	0.709	3,166	-0.044	0.108	9,570
	(0.118)			(0.102)		
LEP: No (Grade 1-3)	-0.166**		3,736	-0.203***		11,064
	(0.070)			(0.055)		
LEP: Yes (Grade 4-5)	-0.081	0.720	5,690	-0.101	0.874	5,836
	(0.071)			(0.077)		
LEP: No (Grade 4-5)	-0.107**		7,442	-0.114**		7,587
	(0.047)			(0.047)		
Panel B: Grade Levels						
Grade 1 Cohort	—			-0.128	0.588	6,887
				(0.085)		
Grade 2 Cohort	—			-0.101		7,325
				(0.084)		
Grade 3 Cohort	-0.153**	0.040	7,189	-0.157**		7,312
	(0.068)			(0.069)		
Grade 4 Cohort	-0.142**		7,058	-0.169***		7,222
	(0.063)			(0.066)		
Grade 5 Cohort	-0.077		6,575	-0.067		6,721
	(0.055)			(0.051)		
Panel C: Teacher Characteristics						
Math:						
Teacher experience $<$ 3 years	-0.175***		6,435	-0.106***		9,762
	(0.030)			(0.028)		
Teacher experience $>=$ 3 years	0.003	0.000	10,401	-0.002	0.003	18,816
	(0.035)			(0.033)		
Reading:						
Teacher experience $<$ 3 years	-0.057		6,419	-0.107***		9,098
	(0.040)			(0.035)		
Teacher experience $>=$ 3 years	-0.061**	0.938	10,330	-0.055*	0.190	19,445
	(0.025)			(0.029)		
Math:						
Above-Median Potential Gains Spec.	-0.056		3,010	-0.049		7,835
	(0.056)			(0.050)		
Below-Median Potential Gains Spec.	-0.044		6,941	0.002		9,298
	(0.033)			(0.039)		

Missing Potential Gains Spec.	-0.098***	0.328	7,563	-0.073**	0.189	12,610
	(0.028)			(0.031)		
Reading:						
Above-Median Potential Gains Spec.	-0.093		2,999	-0.049**		5,615
	(0.041)			(0.050)		
Below-Median Potential Gains Spec.	-0.063***		7,370	0.002*		12,401
	(0.034)			(0.039)		
Missing Potential Gains Spec.	-0.078**	0.413	7,222	-0.073***	0.775	11,796
	(0.028)			(0.031)		
Math:						
Teacher-Student Same Race	-0.024		5,558	-0.034		10,363
	(0.050)			(0.042)		
Teacher-Student Different Race	-0.121***	0.025	3,643	-0.083**	0.171	5,899
	(0.041)			(0.034)		
Read:						
Teacher-Student Same Race	-0.090**		5,506	-0.093***		10,438
	(0.037)			(0.031)		
Teacher-Student Different Race	-0.065***	0.431	3,892	-0.070**	0.500	5,936
	(0.025)			(0.033)		
Math:						
Has Student-Teacher Link	-0.068***	0.156	17,599	-0.044*	0.245	29,921
	(0.026)			(0.027)		
Missing Student-Teacher Link	0.062		3,319	0.061		5,785
	(0.087)			(0.089)		
Read:						
Has Student-Teacher Link	-0.056**	0.345	17,784	-0.071***	0.771	29,895
	(0.022)			(0.024)		
Missing Student-Teacher Link	0.022		3,256	-0.043		5,710
	(0.080)			(0.092)		

Notes: This table presents 2SLS estimates of the effect of attending treatment school on high- and low-stakes test scores in different subgroups of the sample. In Panels A, B, and C, high-stakes test scores are summed math and reading STAAR scores and low-stakes scores are summed math and reading Stanford 10 scores (in year 1) or ITBS scores (in year 2). In Panel D (teacher characteristics), math and reading scores are reported separately since students (may) have different math and reading teachers. Teachers' potential gains from specialization is defined as the difference between their TVA in the subject they teach the student in and their average TVA in both math and reading. For details on all variables used to subset the sample, see the Online Appendix. Here treatment is defined as attending a treatment school as the last school in 2012-13. This treatment assignment is used to instrument for having ever attended a treatment school during years of treatment. The sample is restricted each year to those students who are attending grades 3 through 5 (for high-stakes exams) and grades 1 through 5 (for low-stakes exams) and have both valid math and reading test scores. All columns report 2SLS estimates and follow the pooled specification from Table 4. The dependent variable in all specifications are standardized math and reading test scores, (standardized across the district to have a mean of zero and standard deviation one by grade and year). In Panels A, B, and C, the sum of math and reading is used. In Panel D, math and reading are reported separately. All specifications adjust for the student-level demographic variables summarized in Table 2, student-level math and reading scores (3 years prior to 2013-2014) and their squares, and indicators for taking a Spanish baseline test. All specifications have grade-by-year and matched-pair fixed effects. Standard errors, reported in parentheses, are clustered at the school-year level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively.

Appendix Table 7: Sensitivity Analysis or Extension of the Basic Model, 2SLS (Years)

	High-Stakes	*p-value*	Obs	Low-Stakes	*p-value*	Obs
	(1)	(2)	(3)	(4)	(5)	(6)
Full Sample (pooled)	-0.089***		20,822	-0.086**		35,467
	(0.032)			(0.036)		
Panel A: Demographics						
Special Education: Yes	-0.224***	0.050	688	-0.196***	0.072	1,424
	(0.069)			(0.062)		
Special Education: No	-0.084***		19,346	-0.082**		32,633
	(0.032)			(0.037)		
Gifted: Yes	-0.031	0.021	3,603	-0.056	0.220	4,220
	(0.048)			(0.043)		
Gifted: No	-0.112***		16,431	-0.098***		29,837
	(0.032)			(0.037)		
LEP: Yes	-0.061	0.482	8,856	-0.033	0.100	15,406
	(0.054)			(0.058)		
LEP: No	-0.100***		11,178	-0.129***		18,651
	(0.030)			(0.033)		
LEP: Yes (Grade 1-3)	-0.086	0.667	3,166	-0.032	0.090	9,570
	(0.085)			(0.075)		
LEP: No (Grade 1-3)	-0.126**		3,736	-0.158***		11,064
	(0.054)			(0.043)		
LEP: Yes (Grade 4-5)	-0.057	0.657	5,690	-0.072	0.798	5,836
	(0.050)			(0.054)		
LEP: No (Grade 4-5)	-0.080**		7,442	-0.087**		7,587
	(0.035)			(0.036)		
Panel B: Grade Levels						
Grade 1	—			-0.100	0.560	6,887
				(0.066)		
Grade 2	—			-0.076		7,325
				(0.064)		
Grade 3	-0.114**	0.040	7,189	-0.118**		7,312
	(0.051)			(0.052)		
Grade 4	-0.105**		7,058	-0.126***		7,222
	(0.046)			(0.048)		
Grade 5	-0.056		6,575	-0.049		6,721
	(0.040)			(0.038)		
Panel C: Teacher Characteristics						
Math:						
Teacher experience $<$ 3 years	-0.132***		6,435	-0.083***		9,762
	(0.023)			(0.022)		
Teacher experience $>=$ 3 years	0.002	0.000	10,401	-0.001	0.002	18,816
	(0.025)			(0.024)		
Reading:						
Teacher experience $<$ 3 years	-0.043		6,419	-0.082***		9,098
	(0.031)			(0.028)		
Teacher experience $>=$ 3 years	-0.045**	0.964	10,330	-0.041*	0.170	19,445
	(0.018)			(0.022)		
Math:						
Above-Median Potential Gains Spec.	-0.039		3,010	-0.038		7,835
	(0.040)			(0.038)		
Below-Median Potential Gains Spec.	-0.034		6,941	0.001		9,298
	(0.026)			(0.030)		

Missing Potential Gains Spec.	-0.068***	0.435	7,563	-0.052**	0.223	12,610
	(0.020)			(0.022)		
Reading:						
Above-Median Potential Gains Spec.	-0.073		2,999	-0.038**		5,615
	(0.032)			(0.038)		
Below-Median Potential Gains Spec.	-0.049***		7,370	0.001*		12,401
	(0.026)			(0.030)		
Missing Potential Gains Spec.	-0.054**	0.377	7,222	-0.052***	0.783	11,796
	(0.019)			(0.022)		
Math:						
Teacher-Student Same Race	-0.025		5,558	-0.036		10,363
	(0.053)			(0.045)		
Teacher-Student Different Race	-0.130***	0.025	3,643	-0.089**	0.169	5,899
	(0.044)			(0.037)		
Read:						
Teacher-Student Same Race	-0.096**		5,506	-0.100***		10,438
	(0.039)			(0.033)		
Teacher-Student Different Race	-0.069***	0.428	3,892	-0.075**	0.502	5,936
	(0.026)			(0.035)		
Math:						
Has Student-Teacher Link	-0.050***	0.217	17,599	-0.033*	0.302	29,921
	(0.019)			(0.020)		
Missing Student-Teacher Link	0.064		3,319	0.064		5,785
	(0.089)			(0.094)		
Read:						
Has Student-Teacher Link	-0.041**	0.447	17,784	-0.052***	0.940	29,895
	(0.017)			(0.018)		
Missing Student-Teacher Link	0.022		3,256	-0.045		5,710
	(0.082)			(0.096)		

Notes: This table presents 2SLS estimates of the effect of attending treatment school on high- and low-stakes test scores in different subgroups of the sample. In Panels A, B, and C, high-stakes test scores are summed math and reading STAAR scores and low-stakes scores are summed math and reading Stanford 10 scores (in year 1) or ITBS scores (in year 2). In Panel D (teacher characteristics), math and reading scores are reported separately since students (may) have different math and reading teachers. Teachers' potential gains from specialization is defined as the difference between their TVA in the subject they teach the student in and their average TVA in both math and reading. For details on all variables used to subset the sample, see the Online Appendix. Here treatment is defined as attending a treatment school as the last school in 2012-13. This treatment assignment is used to instrument for the number of years spent in a treatment school. The sample is restricted each year to those students who are attending grades 3 through 5 (for high-stakes exams) and grades 1 through 5 (for low-stakes exams) and have both valid math and reading test scores. All columns report 2SLS estimates and follow the pooled specification from Table 4. The dependent variable in all specifications are standardized math and reading test scores, (standardized across the district to have a mean of zero and standard deviation one by grade and year). In Panels A, B, and C, the sum of math and reading is used. In Panel D, math and reading are reported separately. All specifications adjust for the student-level demographic variables summarized in Table 2, student-level math and reading scores (3 years prior to 2013-2014) and their squares, and indicators for taking a Spanish baseline test. All specifications have grade-by-year and matched-pair fixed effects. Standard errors, reported in parentheses, are clustered at the school-year level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively.

Appendix Table 8: Attrition due to Modified, Accommodated or Missing Staar Test Scores

	Missing Score		*Modified*		*L Score*	
	Control Mean	Treatment Effect	Control Mean	Treatment Effect	Control Mean	Treatment Effect
	(1)	(2)	(3)	(4)	(5)	(6)
Math	0.061	-0.001	0.020	-0.001	0.012	-0.007***
		(0.003)		(0.004)		0.002
Observations		22,846		11,770		22,846
Reading	0.061	-0.002	0.022	-0.002	—	—
		(0.003)		(0.004)		
Observations		22,846		11,770		

Notes: This table presents estimates of the effects of being enrolled in a treatment school on three measures of attrition. Students can exit our sample in one of four ways: taking the Modified STAAR exam offered to students with an individualized education program, taking the STAAR L exam offered to students with limited English proficiency, or missing the exam entirely. Beginning in 2014-2015, the Modified STAAR exam was no longer offered. Columns (1), (3), and (5) report the mean levels of attrition in the control group. The treatment effects in columns (2), (4), and (6) follow the ITT specification described in Section IV. Columns (2) and (6) are pooled across years and column (4) uses data only from 2013-14. Standard errors, reported in parentheses, are clustered at the school-year level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively.

Appendix Table 9: The Effect of Treatment on Outcomes, School-Level Regressions

	Matched-Pair Fixed Effects			School-Level Demographic and Testing Controls		
	2014	2015	Pooled	2014	2015	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Student Achievement						
High-Stakes Math	-0.068	-0.042	-0.055	-0.132	-0.042	-0.133
	(0.049)	(0.043)	(0.040)	(0.087)	(0.054)	(0.092)
High-Stakes Reading	-0.072*	-0.023	-0.049	-0.106	0.008	-0.082
	(0.039)	(0.042)	(0.039)	(0.078)	(0.060)	(0.078)
Low-Stakes Math	-0.068	-0.034	-0.052	-0.086	-0.063	-0.107
	(0.045)	(0.040)	(0.039)	(0.077)	(0.060)	(0.076)
Low-Stakes Reading	-0.080*	-0.049	-0.067	-0.065	-0.036	-0.074
	(0.046)	(0.046)	(0.044)	(0.090)	(0.046)	(0.073)
Panel B: Alternate Outcomes						
Attendance (in years)	-0.003	-0.002	-0.003	-0.002	-0.001	-0.003
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
Behavioral Incidents	0.012			-0.004		
	(0.010)	–	–	(0.033)	–	–
Teacher Retention	-0.018	0.043	0.012	-0.018	0.038	-0.020
	(0.028)	(0.036)	(0.027)	(0.067)	(0.046)	(0.048)

Notes: This table presents the estimates of being enrolled in a treatment school on school level average of subject test scores in both treatment years. The specifications follow the main OLS specification from Table 3 at the school level rather than the individual level. All regressions presented have a sample size of 46. Panel A presents ITT estimates on math and reading scores. Panel B presents ITT estimates on average attendance rates (measured in years), average behavioral incidents and teacher retention. Teacher retention is calculated as the fraction of teachers retained between 2012-2013 and 2013-2014 per school. See Online Appendix B for a detailed construction of all variables. Column (1) uses outcomes from 2013-2014, column (2) uses outcomes for 2014-2015, and column (3) uses outcomes from both years. Columns (1)-(3) include matched-pair fixed effects as controls. Columns (4)-(6) include the mean of demographic controls taken at the school level and the school mean of students' 2010-2011, 2011-2012, and 2012-2013 test scores. School level baseline means of attendance rates and behavioral incidents are included when the outcome variable is attendance rate and behavioral incidence, respectively. These means are taken only over students included in the main analysis in Table (4), those with valid STAAR math and reading test scores in the outcome year and in grades 3-5. Standard errors, robust to heteroskedasticity, are located in parentheses. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels respectively.

Appendix Table 10: The Effect of Treatment on Science and Social Studies Test Scores

	ITT			2SLS (Ever)			2SLS (Years)		
	2014	2015	Pooled	2014	2015	Pooled	2014	2015	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: All Grades									
Science	-0.077***	-0.069**	-0.071***	-0.086***	-0.081**	-0.082***	-0.093***	-0.045**	-0.057***
	(0.020)	(0.029)	(0.022)	(0.022)	(0.034)	(0.025)	(0.024)	(0.019)	(0.017)
Observations	10,948	16,797	27,745	10,948	16,797	27,745	10,948	16,797	27,745
Social Studies	-0.039	-0.084**	-0.066**	-0.044	-0.098**	-0.076**	-0.047	-0.055**	-0.053**
	(0.029)	(0.038)	(0.029)	(0.032)	(0.044)	(0.033)	(0.034)	(0.025)	(0.023)
Observations	10,945	16,801	27,746	10,945	16,801	27,746	10,945	16,801	27,746
Panel B: Grades 3 - 5									
Science	-0.077***	-0.067**	-0.072***	-0.086***	-0.077**	-0.082***	-0.093***	-0.043**	-0.043**
	(0.020)	(0.027)	(0.020)	(0.022)	(0.031)	(0.023)	(0.024)	(0.017)	(0.017)
Observations	10,948	10,261	21,209	10,948	10,261	21,209	10,948	10,261	10,261
Social Studies	-0.039	-0.047	-0.043*	-0.044	-0.055	-0.049*	-0.047	-0.030	-0.030
	(0.029)	(0.033)	(0.024)	(0.032)	(0.038)	(0.027)	(0.034)	(0.021)	(0.021)
Observations	10,945	10,258	21,203	10,945	10,258	21,203	10,945	10,258	10,258
Panel C: Grades 1 - 2									
Science	—	-0.059	—	—	-0.071	—	—	-0.040	—
		(0.042)			(0.050)			(0.028)	
Observations		7,026			7,026			7,026	
Social Studies	—	-0.131**	—	—	-0.156**	—	—	-0.088**	—
		(0.053)			(0.062)			(0.035)	
Observations		7,034			7,034			7,034	

Notes: This table presents estimates of being enrolled in a treatment school on Stanford 10 science and social studies test scores. Specifications are identical to those in Table 4 and samples are subsets of those in Table 4 with non-missing science or social studies scores. In all panels the dependent variable is a nationally-normed low-stakes test score, standardized to have a mean of zero and standard deviation one by grade and year. In the 2013-2014 school year, HISD administered the Stanford 10 test. In the 2014-2015 school year, HISD administered the Iowa Test of Basic Skills. Stanford 10 only administers science and social studies tests to those in 3-5th grade, and so scores for younger students are only available in the second year of treatment. All specifications adjust for the student level demographic variables summarized in Table 2, student level math and reading scores (3 years prior to 2013-2014) and their squares, and indicator variables for taking a Spanish baseline test. All specifications have grade-by-year and matched pair fixed effects. Standard errors, reported in parentheses, are clustered at the school-year level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively.

Appendix Table 11: Main ITT Estimates Accounting for Multiple Hypothesis Testing

	ITT Estimate	Uncorrected p-value	Holm Corrected p-value
	(1)	(2)	(3)
Main High-Stakes Effect 2013-14	-0.122	0.010	0.051
	(0.046)		
Main High-Stakes Effect 2014-15	-0.089	0.092	0.183
	(0.051)		
Main High-Stakes Effect, Pooled	-0.106	0.006	0.039
	(0.038)		
Special Ed High-Stakes Effect	-0.296	0.002	0.023
	(0.094)		
Non-Special Ed High-Stakes Effect	-0.101	0.011	0.092
	(0.039)		
Main Low-Stakes Effect 2013-14	-0.109	0.025	0.074
	(0.047)		
Main Low-Stakes Effect 2014-15	-0.090	0.171	0.183
	(0.065)		
Main Low-Stakes Effect, Pooled	-0.100	0.018	0.073
	(0.042)		
Special Ed Low-Stakes Effect	-0.219	0.002	0.023
	(0.069)		
Non-Special Ed Low-Stakes Effect	-0.096	0.028	0.196
	(0.043)		
Less Exp. Math Teachers High-Stakes Math Effect	-0.176	0.000	0.004
	(0.045)		
More Exp. Math Teachers High-Stakes Math Effect	-0.025	0.556	1.000
	(0.041)		
Less Exp. Math Teachers Low-Stakes Math Effect	-0.123	0.005	0.045
	(0.041)		
More Exp. Math Teachers Low-Stakes Math Effect	-0.000	0.994	1.000
	(0.047)		
Less Exp. Reading Teachers High-Stakes Reading Effect	0.040	0.450	1.000
	(0.053)		
More Exp. Reading Teachers High-Stakes Reading Effect	-0.075	0.030	0.196
	(0.034)		
Less Exp. Reading Teachers Low-Stakes Reading Effect	-0.069	0.117	0.574
	(0.043)		
More Exp. Reading Teachers Low-Stakes Reading Effect	-0.072	0.115	0.574
	(0.045)		

Notes: This table reports ITT estimates correcting for multiple hypothesis testing. Column (1) reports ITT estimates following the specifications described in Tables 4 and 5. Column (2) reports the unadjusted p-value. Column (3) reports p-values controlling for the Familywise Error Rate, the probability of at least one false rejection, using the Holm Stepdown method described in Romano, Shaikh and Wolf (2010). Standard errors, reported in parentheses, are clustered at the school-year level.

Appendix Table 12: The Effect of Treatment on Survey Outcomes (Lee Bounds)

	Full Sample		Inexperienced Teachers		Experienced Teachers	
	ITT	Lee Bound	ITT	Lee Bound	ITT	Lee Bound
	(1)	(2)	(3)	(4)	(5)	(6)
Know students	-0.045*	-0.086*	-0.114***	-0.139***	-0.013	-0.060
	(0.026)	(0.044)	(0.040)	(0.047)	(0.030)	(0.037)
Observations	663	637	227	222	409	385
Gives individual attention	-0.089***	-0.136***	-0.191***	-0.218***	-0.048	-0.090**
	(0.025)	(0.031)	(0.064)	(0.064)	(0.037)	(0.039)
Observations	667	643	229	226	412	390
Rules enforced	-0.041	-0.070	-0.084	-0.116	-0.002	-0.048
	(0.061)	(0.064)	(0.088)	(0.087)	(0.062)	(0.064)
Observations	670	648	229	226	414	393
Enthusiasm for teaching math	-0.026	-0.043	-0.032	-0.051	-0.017	-0.031
	(0.025)	(0.026)	(0.034)	(0.035)	(0.030)	(0.030)
Observations	695	690	234	231	434	431
Enthusiasm for teaching reading	-0.013	-0.022	-0.033	-0.033	0.005	-0.001
	(0.020)	(0.021)	(0.039)	(0.039)	(0.022)	(0.021)
Observations	698	694	234	231	436	434
Lesson differentiation	2.137	0.251	-0.999	-2.411	2.638	0.012
	(1.942)	(1.971)	(3.419)	(3.140)	(2.636)	(2.495)
Observations	684	670	232	228	424	412
Above-median change in job satisfaction	-0.170***	-0.542***	-0.154**	-0.470***	-0.248***	-0.599***
	(0.047)	(0.047)	(0.062)	(0.067)	(0.059)	(0.054)
Observations	484	342	172	121	299	215
Above-median change in job performance	-0.159***	-0.513***	-0.216***	-0.507***	-0.175***	-0.554***
	(0.052)	(0.048)	(0.066)	(0.070)	(0.058)	(0.054)
Observations	484	344	172	121	299	216

Notes: This table presents bounded estimates to provide a conservative bound on the true treatment effects under the assumption that there are differential rates of response to the survey. Odd columns present the ITT estimates from Table 6. Even columns provide the Lee bounded estimates. Standard errors, clustered by school, are reported in parentheses. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Appendix Table 13: The Effect of Treatment on Student High-Stakes Test Scores Using Highest of Multiple Test Scores

	ITT			2SLS (Ever)			2SLS (Years)		
	2014	2015	Pooled	2014	2015	Pooled	2014	2015	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Math	-0.056**	-0.061**	-0.058**	-0.062**	-0.071**	-0.066***	-0.067**	-0.040**	-0.049***
	(0.027)	(0.029)	(0.022)	(0.030)	(0.033)	(0.025)	(0.032)	(0.018)	(0.019)
Reading	-0.064**	-0.031	-0.048**	-0.071***	-0.036	-0.055**	-0.077***	-0.020	-0.040**
	(0.025)	(0.028)	(0.020)	(0.027)	(0.032)	(0.023)	(0.029)	(0.018)	(0.017)
Observations	10,462	10,159	20,621	10,462	10,159	20,621	10,462	10,159	20,621
First Stage Coefficient				0.896***	0.858***	0.877***	0.833***	1.551***	1.187***
				(0.009)	(0.012)	(0.008)	(0.009)	(0.025)	(0.042)

Notes: This table presents the identical specifications to Appendix Table 2 using a student's highest test score as the outcome variable rather than following the procedure detailed in the Online Appendix to choose between a student's multiple test scores. Standard errors, reported in parentheses, are clustered at the school-year level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively.

Appendix Table 14: The Effect of Treatment on Student Test Scores, Without Matched Pairs with Highly Specialized Control Schools

	ITT			2SLS (Ever)			2SLS (Years)		
	2014	2015	Pooled	2014	2015	Pooled	2014	2015	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
High-Stakes	-0.172**	-0.090	-0.124**	-0.192***	-0.104	-0.141**	-0.206***	-0.058	-0.104**
	(0.064)	(0.078)	(0.054)	(0.069)	(0.088)	(0.060)	(0.074)	(0.049)	(0.045)
Observations	6,315	6,295	12,610	6,315	6,295	12,610	6,315	6,295	12,610
First Stage Coefficient				0.894***	0.865***	0.879***	0.832***	1.553***	1.193***
				(0.013)	(0.016)	(0.011)	(0.014)	(0.035)	(0.055)
Low-Stakes	-0.157**	-0.140	-0.148**	-0.177**	-0.164	-0.170**	-0.191**	-0.092	-0.127**
	(0.068)	(0.101)	(0.061)	(0.075)	(0.116)	(0.070)	(0.081)	(0.065)	(0.053)
Observations	11,116	10,130	21,246	11,116	10,130	21,246	11,116	10,130	21,246
First Stage Coefficient				0.886***	0.854***	0.871***	0.822***	1.527***	1.159***
				(0.013)	(0.017)	(0.011)	(0.013)	(0.038)	(0.054)

Notes: This table presents the identical specifcations to Table 4 but for the subset of matched pairs in which the control schools have fewer than 50 percent of teachers specialized, as measured by the administrative data presented in Panel B of Tables (1) and (2). Standard errors, reported in parentheses, are clustered at the school-year level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively.

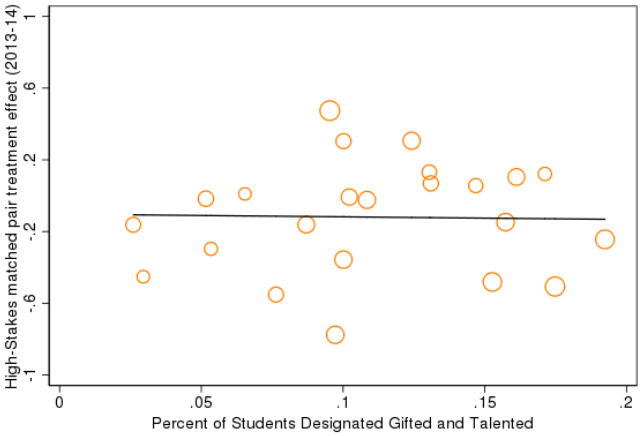
Appendix Table 15: The Effect of Treatment on Student Test Scores (Raw Regressions)

	ITT			2SLS (Ever)			2SLS (Years)		
	2014	2015	Pooled	2014	2015	Pooled	2014	2015	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
High-Stakes	-0.139**	-0.091*	-0.115***	-0.155**	-0.106*	-0.131***	-0.167**	-0.059*	-0.097***
	(0.059)	(0.053)	(0.042)	(0.064)	(0.060)	(0.048)	(0.069)	(0.033)	(0.035)
Observations	10,462	10,360	20,822	10,462	10,360	20,822	10,462	10,360	20,822
First Stage Coefficient				0.897***	0.859***	0.878***	0.834***	1.553***	1.192***
				(0.009)	(0.012)	(0.008)	(0.009)	(0.025)	(0.042)
Low-Stakes	-0.113*	-0.089	-0.102**	-0.127**	-0.105	-0.117**	-0.137**	-0.058	-0.088**
	(0.056)	(0.068)	(0.046)	(0.063)	(0.078)	(0.053)	(0.068)	(0.044)	(0.040)
Observations	18,618	16,849	35,467	18,618	16,849	35,467	18,618	16,849	35,467
First Stage Coefficient				0.890***	0.855***	0.873***	0.824***	1.535***	1.162***
				(0.009)	(0.012)	(0.008)	(0.010)	(0.026)	(0.042)

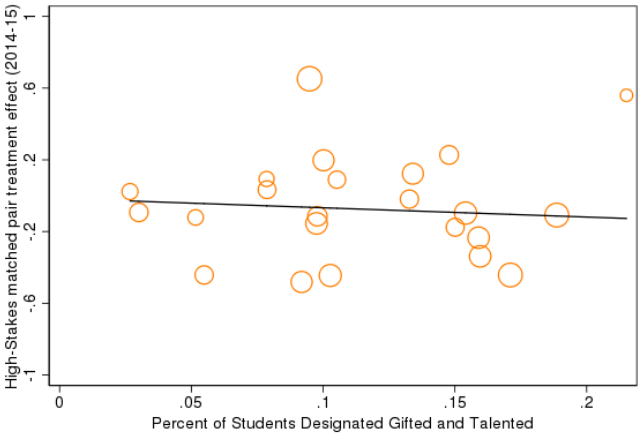
Notes: This table presents the identical specifications to Table 4 without any student demographic or pre-treatment test controls. The specifications include grade-by-year and matched pair fixed effects. Standard errors, reported in parentheses, are clustered at the school-year level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively.

Appendix Figure 1: Heterogeneous Treatment Effects

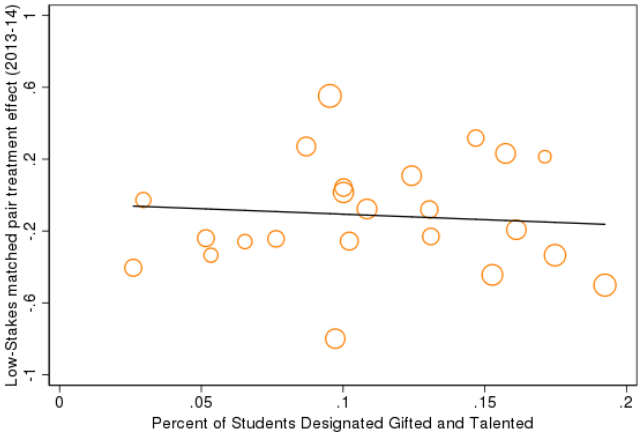
Panel A: School Characteristics



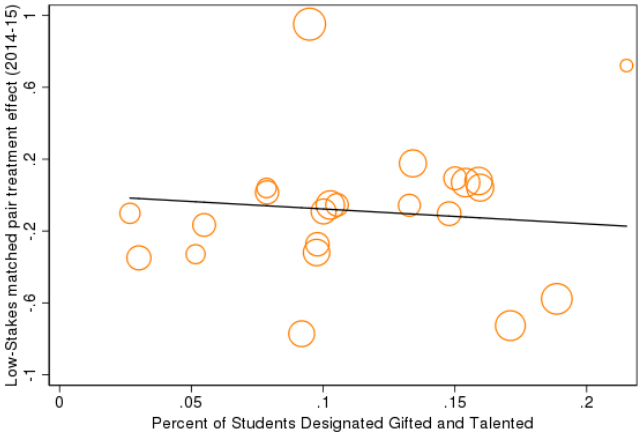
Slope: -0.147, p-value: 0.9085, N: 23



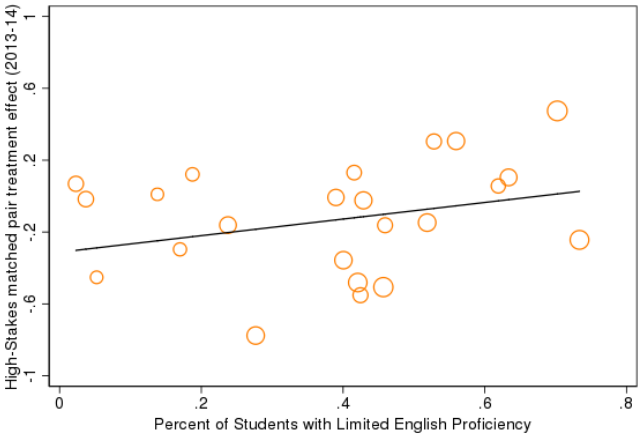
Slope: -0.517, p-value: 0.6812, N: 23



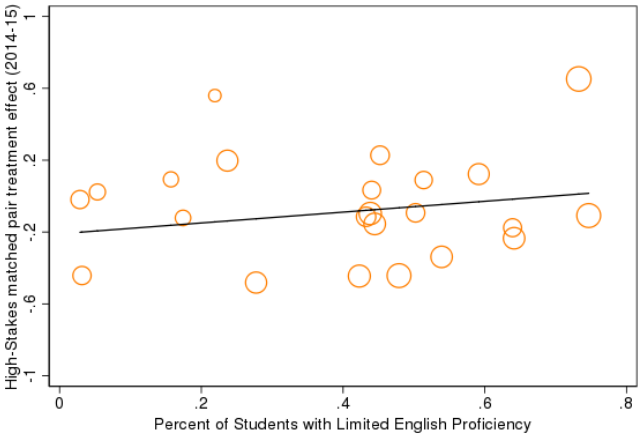
Slope: -0.609, p-value: 0.6885, N: 23



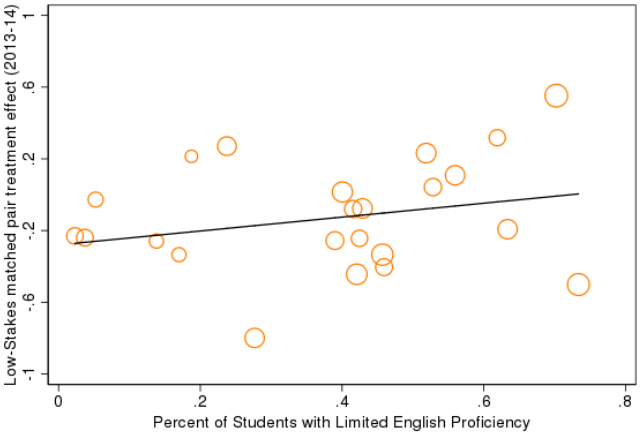
Slope: -0.830, p-value: 0.6873, N: 23



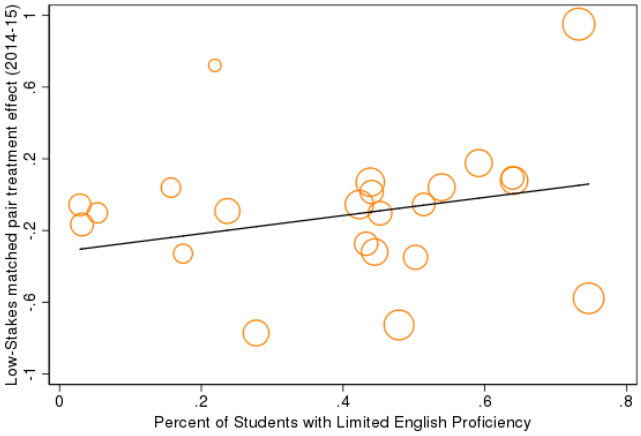
Slope: 0.462, p-value: 0.2201, N: 23



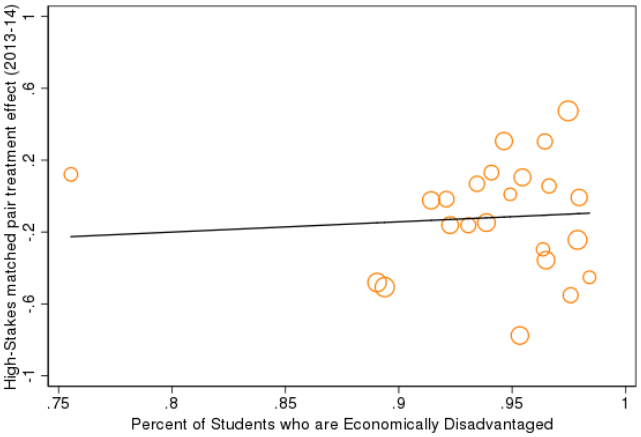
Slope: 0.302, p-value: 0.4405, N: 23



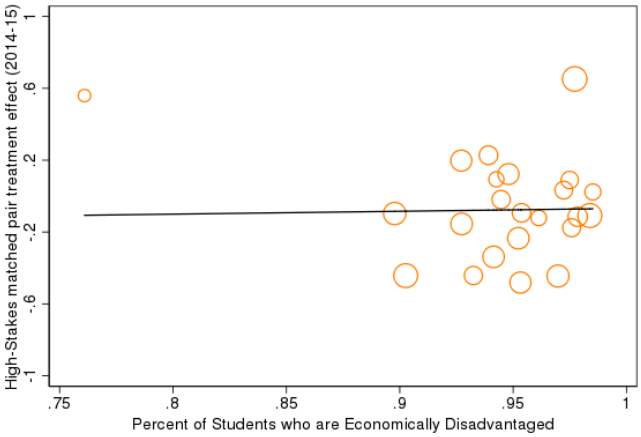
Slope: 0.387, p-value: 0.3903, N: 23



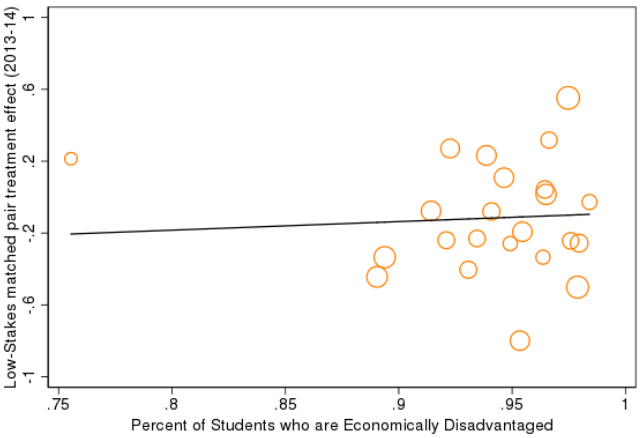
Slope: 0.506, p-value: 0.4130, N: 23



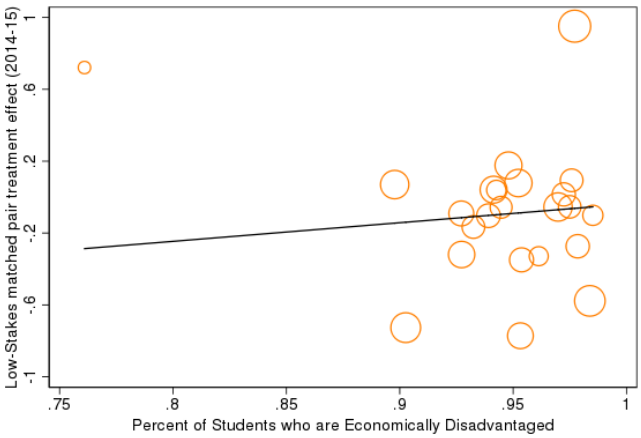
Slope: 0.571, p-value: 0.7410, N: 23



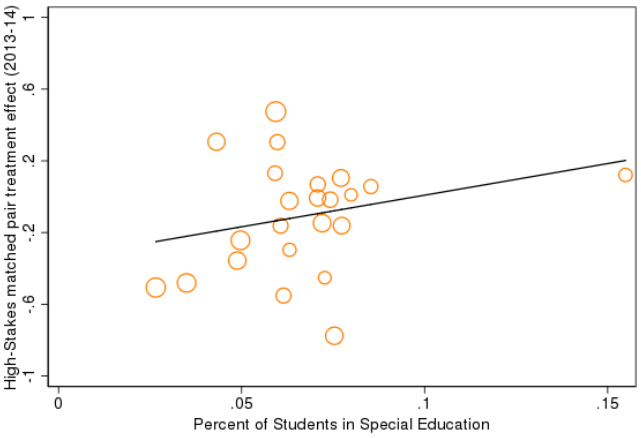
Slope: 0.159, p-value: 0.9470, N: 23



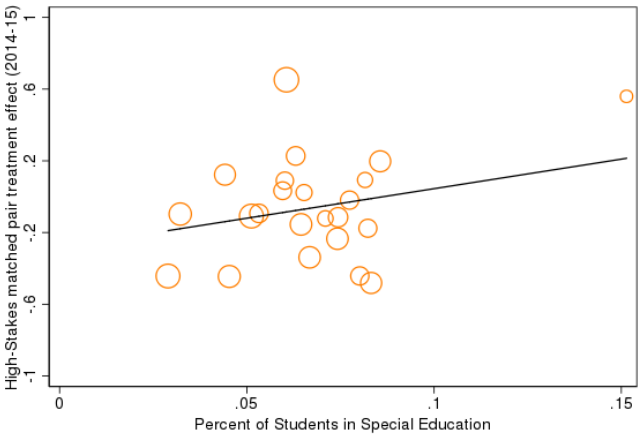
Slope: 0.478, p-value: 0.8058, N: 23



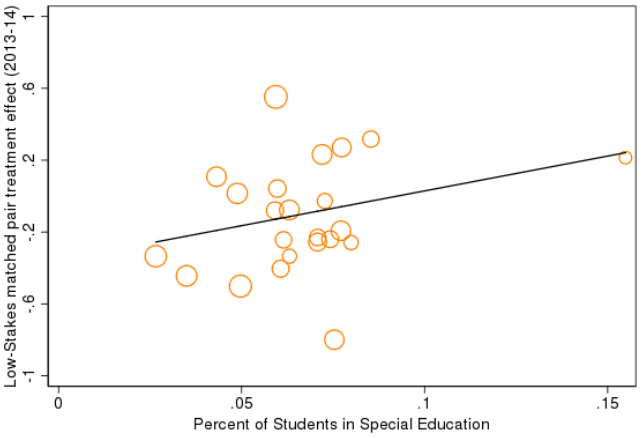
Slope: 1.036, p-value: 0.7843, N: 23



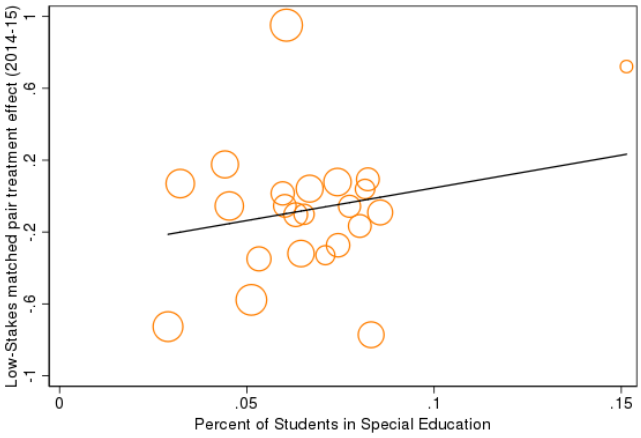
Slope: 3.525, p-value: 0.1372, N: 23



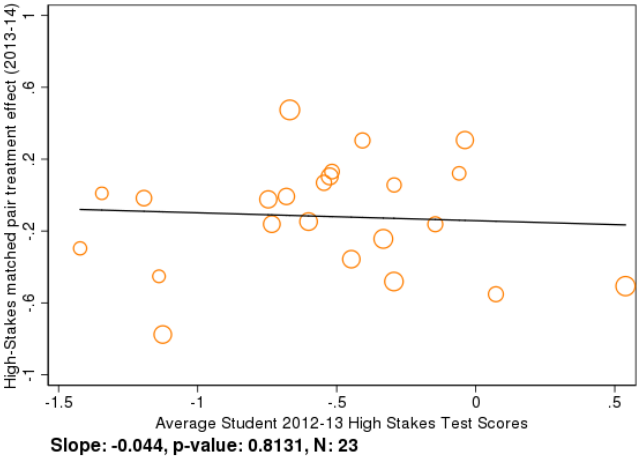
Slope: 3.293, p-value: 0.2444, N: 23



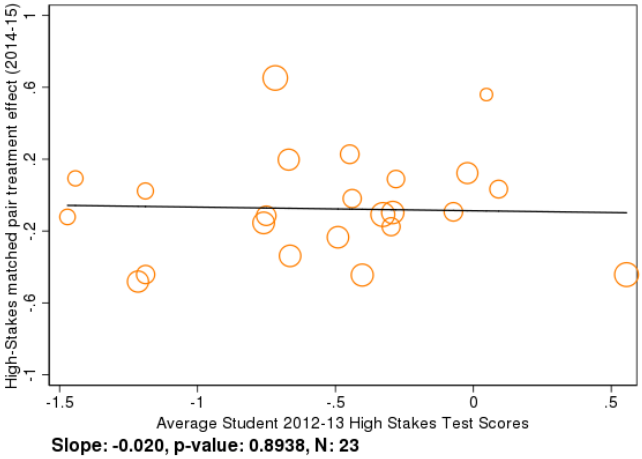
Slope: 3.872, p-value: 0.1042, N: 23



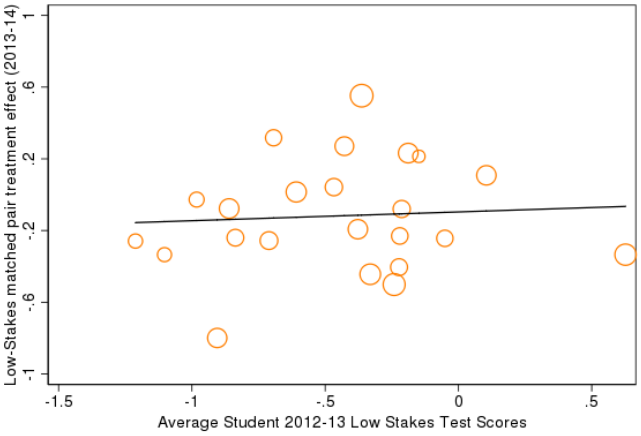
Slope: 3.638, p-value: 0.4420, N: 23



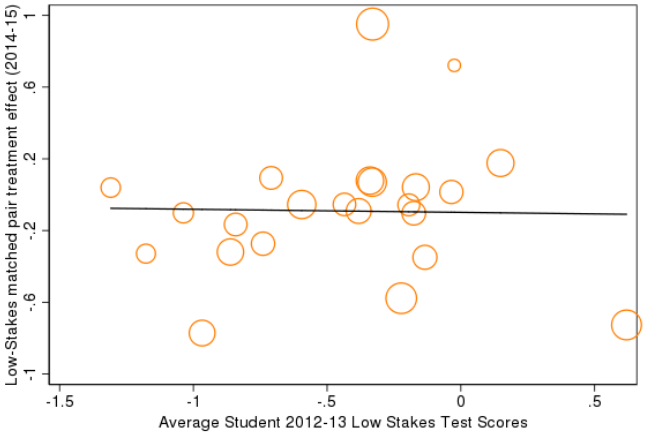
Slope: -0.044, p-value: 0.8131, N: 23



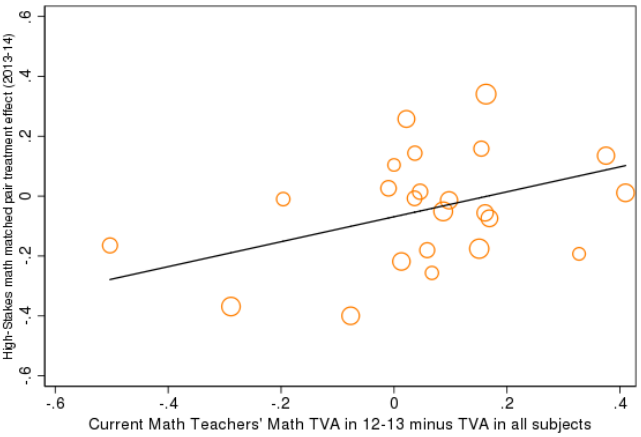
Slope: -0.020, p-value: 0.8938, N: 23



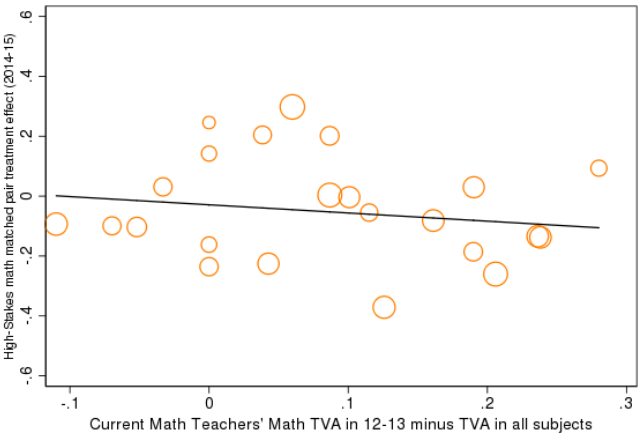
Slope: 0.049, p-value: 0.7625, N: 23



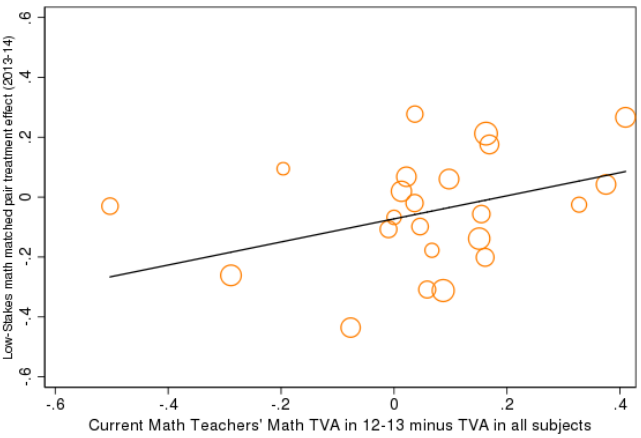
Slope: -0.017, p-value: 0.9465, N: 23



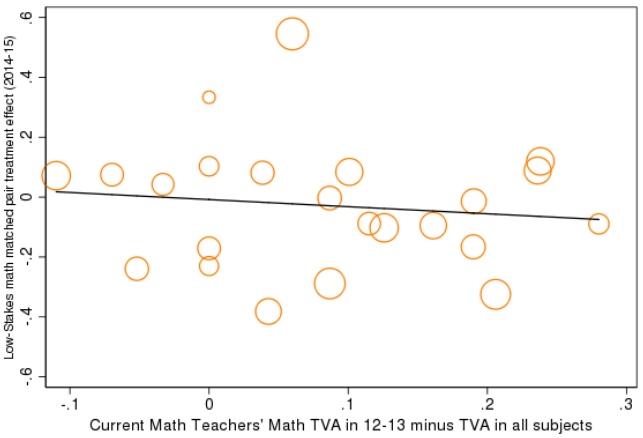
Slope: 0.417, p-value: 0.0241, N: 23



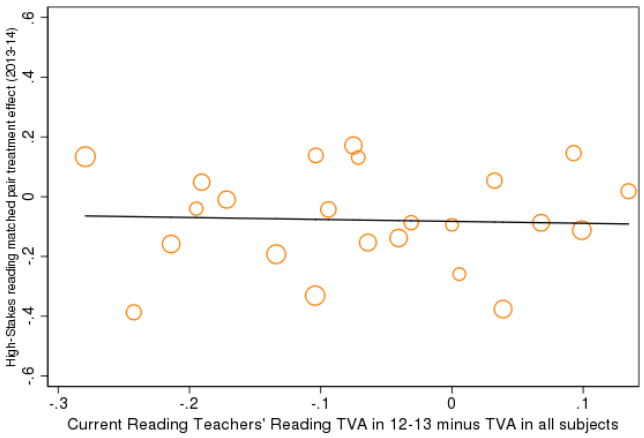
Slope: -0.274, p-value: 0.2997, N: 23



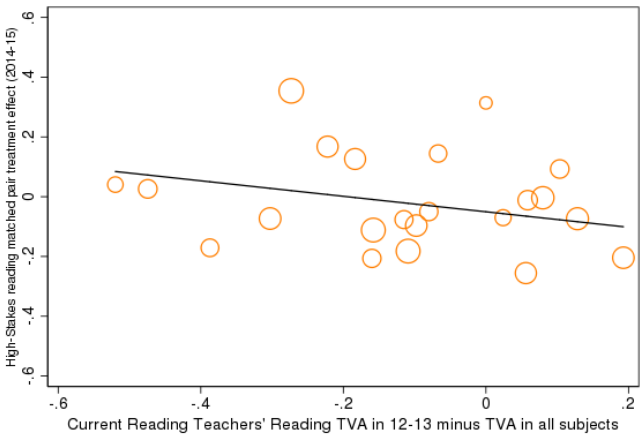
Slope: 0.385, p-value: 0.0623, N: 23



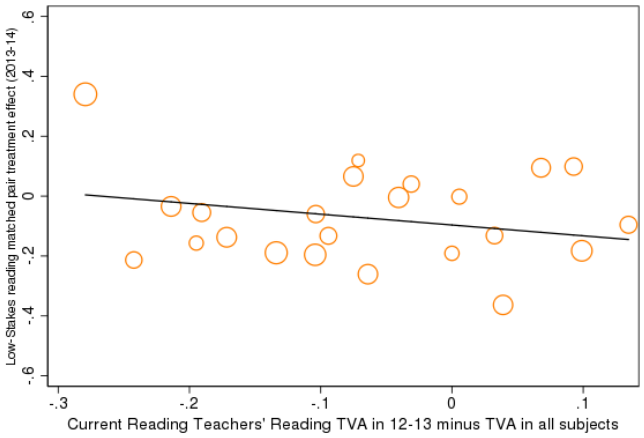
Slope: -0.237, p-value: 0.5012, N: 23



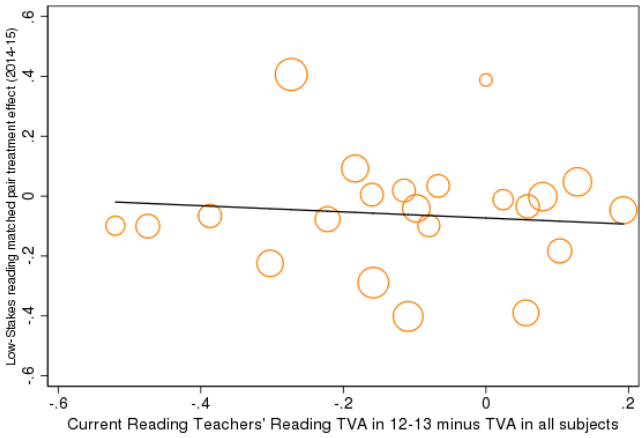
Slope: -0.065, p-value: 0.8494, N: 23



Slope: -0.259, p-value: 0.1630, N: 22

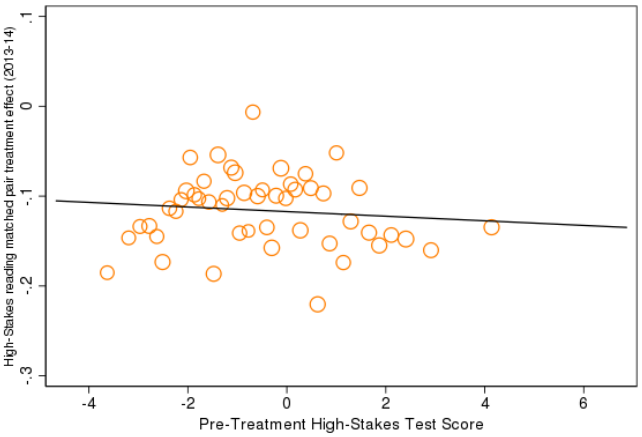


Slope: -0.360, p-value: 0.4199, N: 23

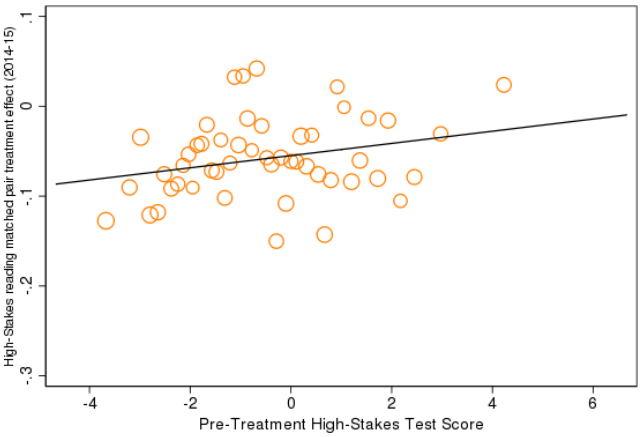


Slope: -0.103, p-value: 0.6521, N: 22

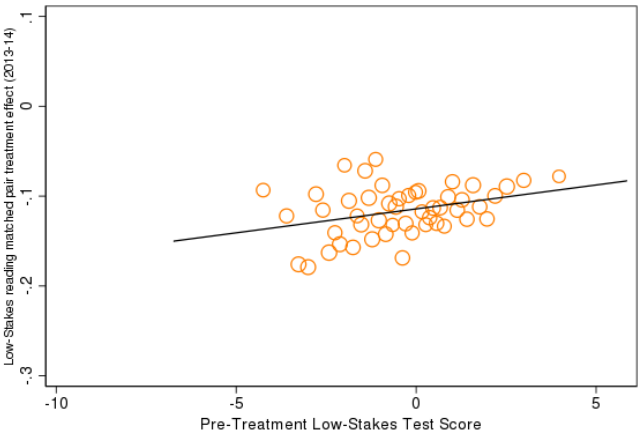
Panel B: Student Characteristics



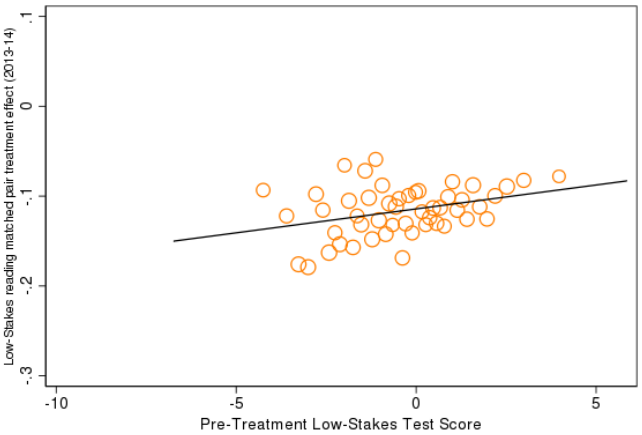
Slope: -0.003, p-value: 0.4271, N: 3686



Slope: 0.007, p-value: 0.0944, N: 1886

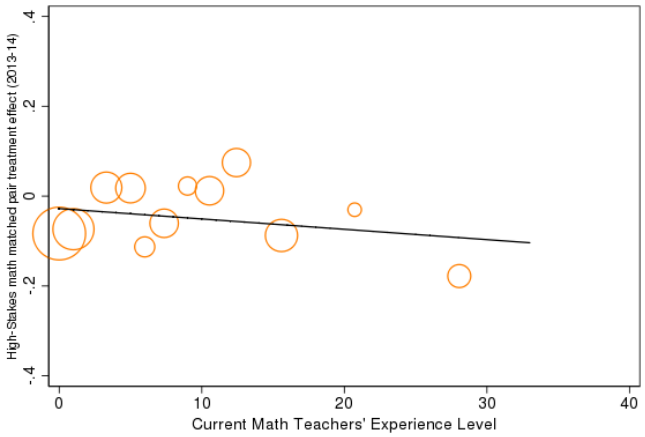


Slope: 0.005, p-value: 0.0053, N: 9586

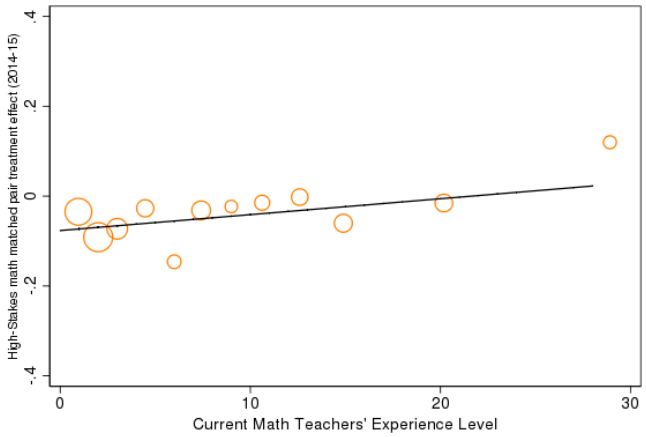


Slope: 0.005, p-value: 0.0053, N: 9586

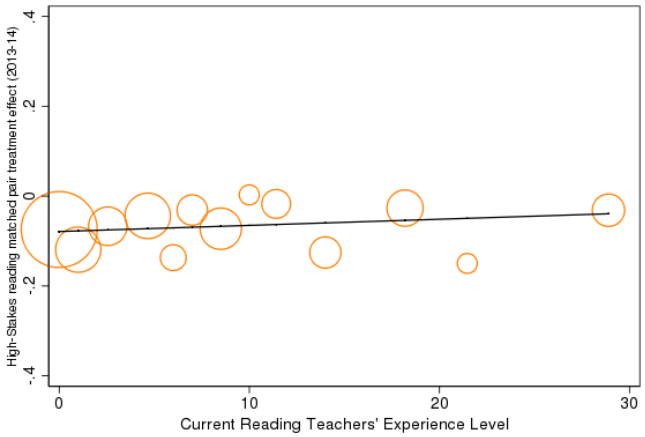
Panel C: Teacher Characteristics



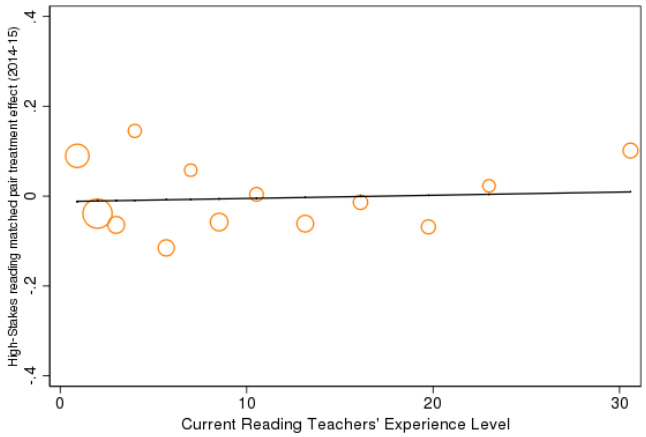
Slope: -0.002, p-value: 0.3361, N: 137



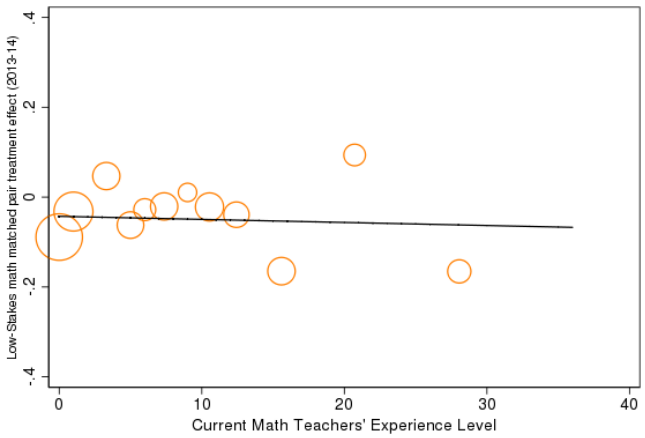
Slope: 0.004, p-value: 0.1560, N: 137



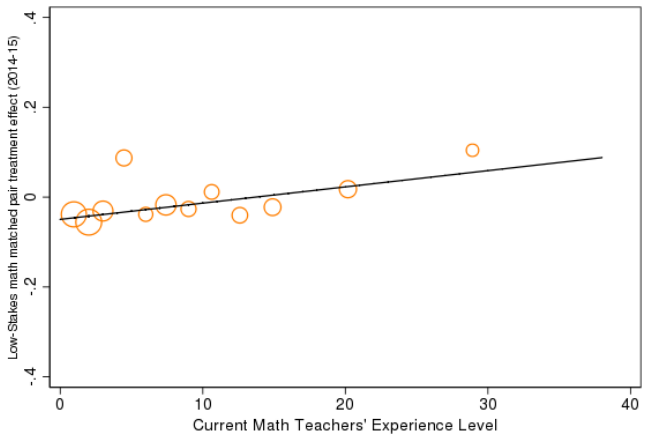
Slope: 0.001, p-value: 0.0003, N: 148



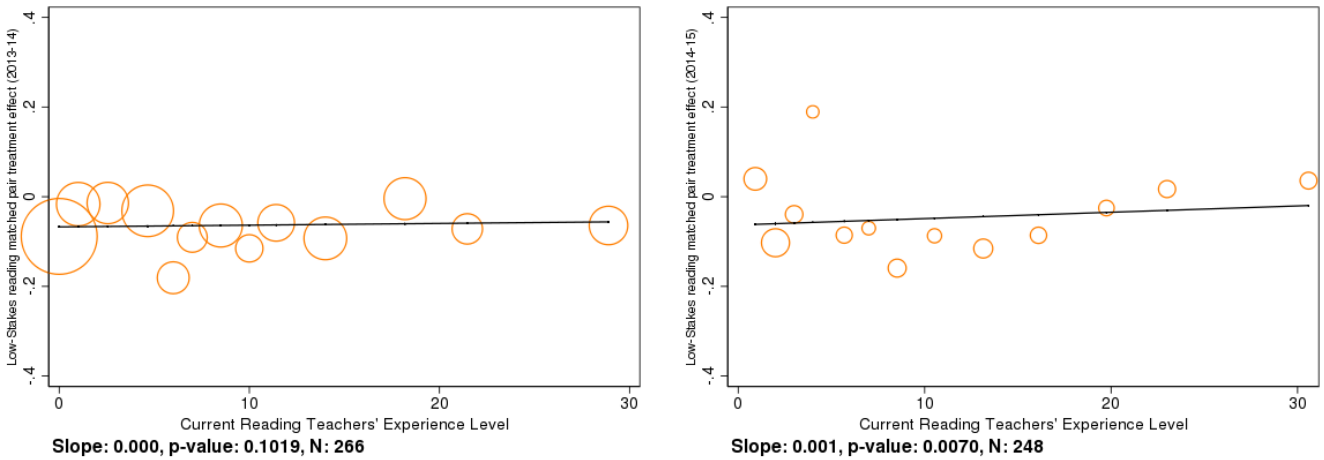
Slope: 0.001, p-value: 0.4436, N: 133



Slope: -0.001, p-value: 0.7174, N: 255

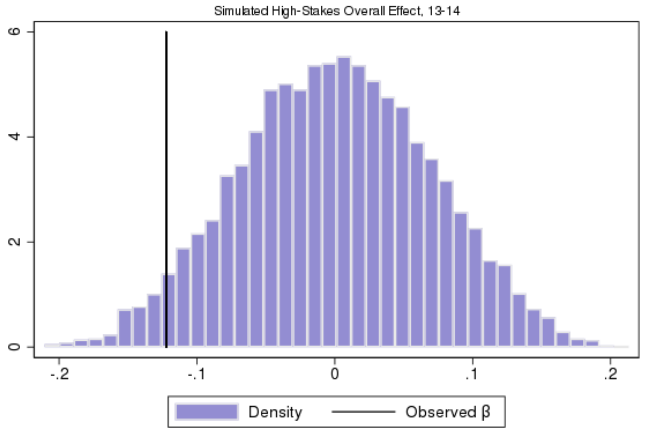


Slope: 0.004, p-value: 0.1637, N: 252

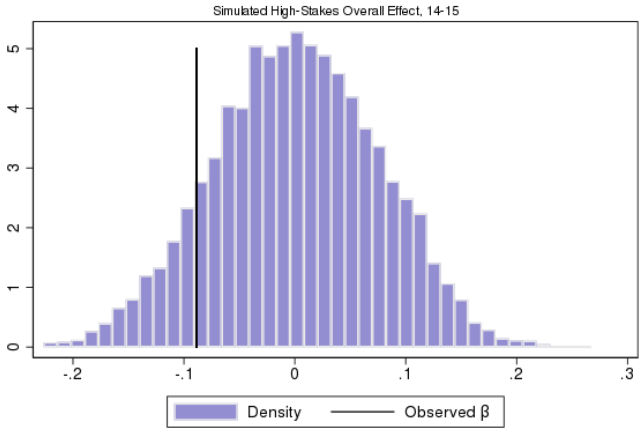


Notes: Each figure plots matched pair treatment effects against school, student, and teacher characteristics. When the variable on the horizontal axis is subject-dependent (i.e. math teachers' math TVA), the y-axis variable is the matched pair treatment effect in that subject. Elsewhere the y-axis variable is the matched pair treatment effect on summed math and reading high- and low-stakes scores. The matched pair treatment coefficient is calculated by regressing students' test scores on an indicator for treatment and controls for the student-level demographic variables summarized in Table 2, student-level math and reading scores (3 years prior to 2013-2014) and their squares, and indicators for taking a Spanish baseline test as well as grade fixed effects separately for each matched pair. The plots in Panels B and C bin variables into 50 and 15 bins, respectively, but the regression results and the line of best fit are calculated from the unbinned data.

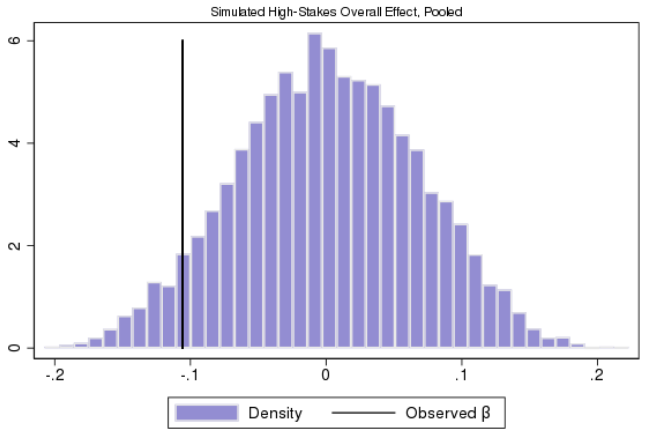
Appendix Figure 2: Permutation Tests



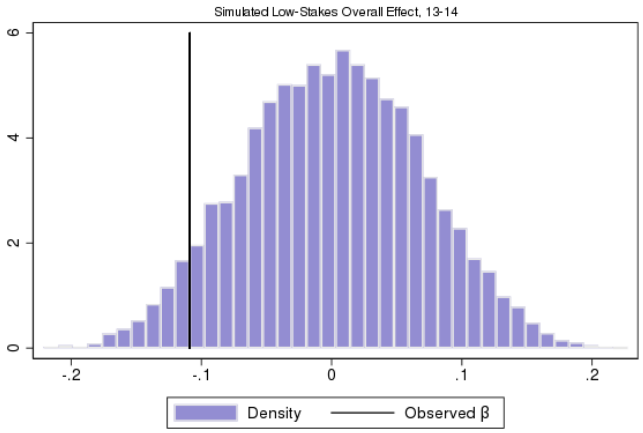
N = 10000 Replications. **P-value** (two-tailed) = 0.078



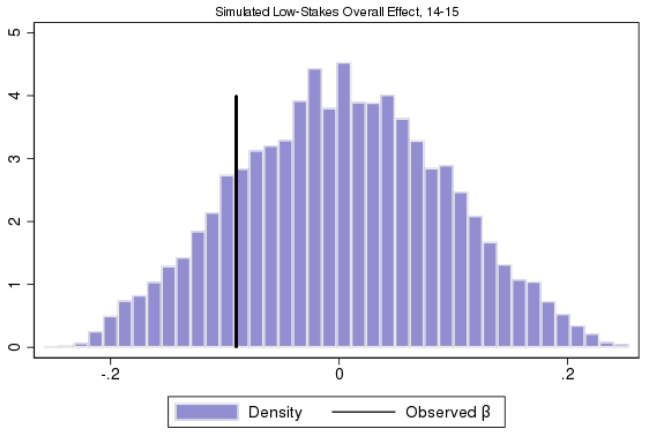
N = 10000 Replications. **P-value** (two-tailed) = 0.247



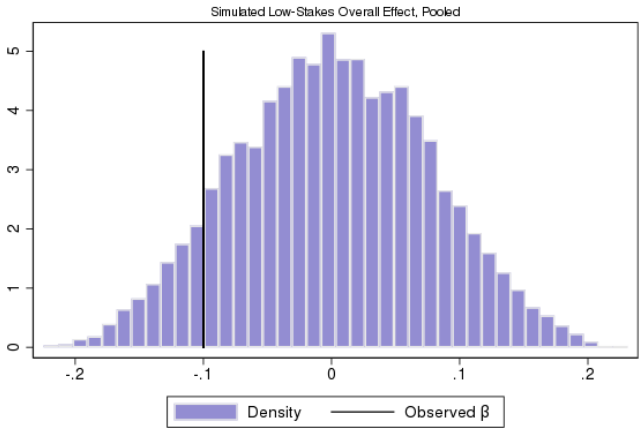
N = 10000 Replications. **P-value** (two-tailed) = 0.120



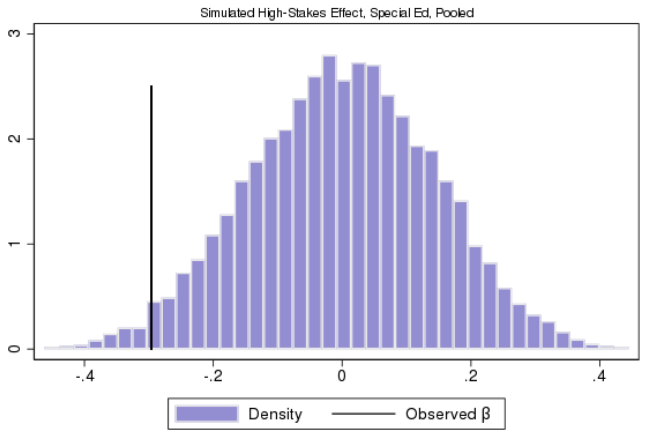
N = 10000 Replications. **P-value** (two-tailed) = 0.115



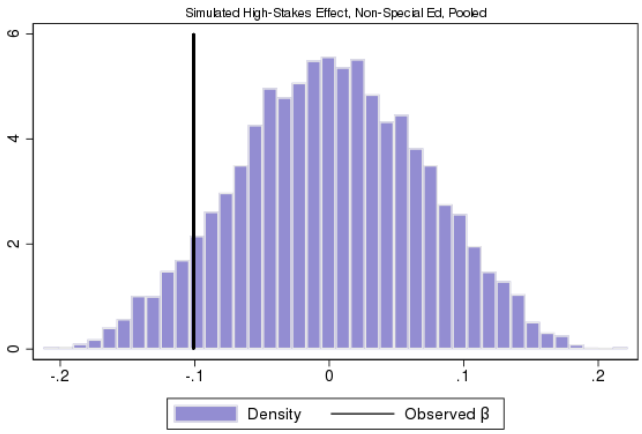
N = 10000 Replications. **P-value** (two-tailed) = 0.348



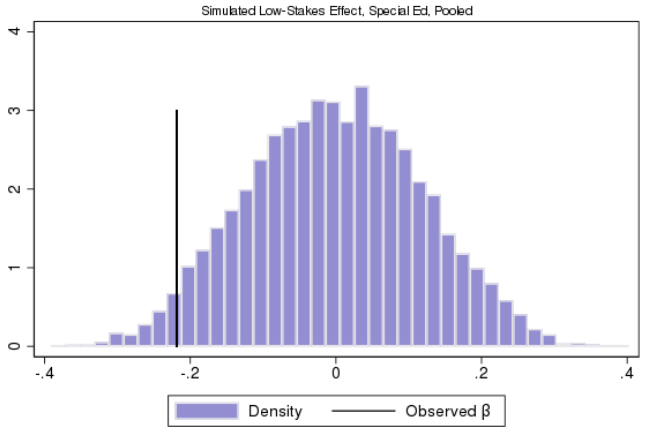
N = 10000 Replications. **P-value** (two-tailed) = 0.197



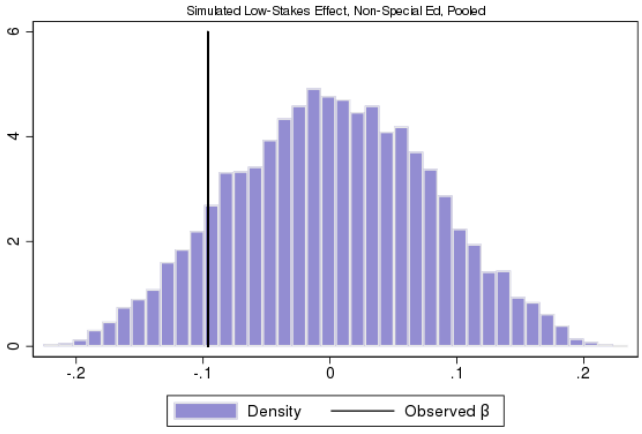
N = 10000 Replications. **P-value** (two-tailed) = 0.038



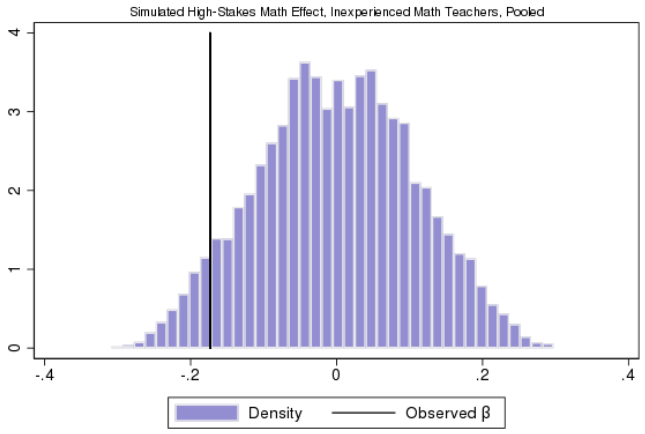
N = 10000 Replications. **P-value** (two-tailed) = 0.157



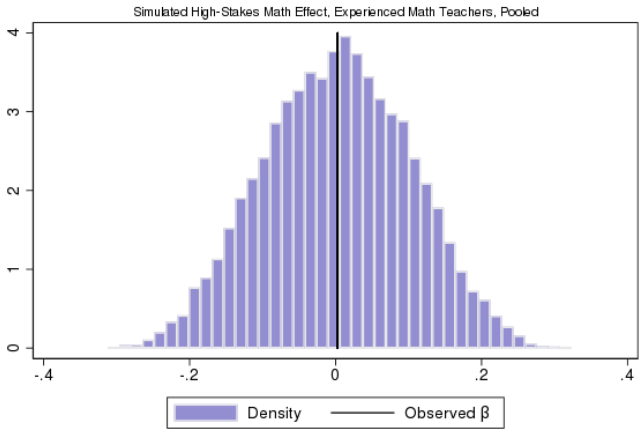
N = 10000 Replications. **P-value** (two-tailed) = 0.064



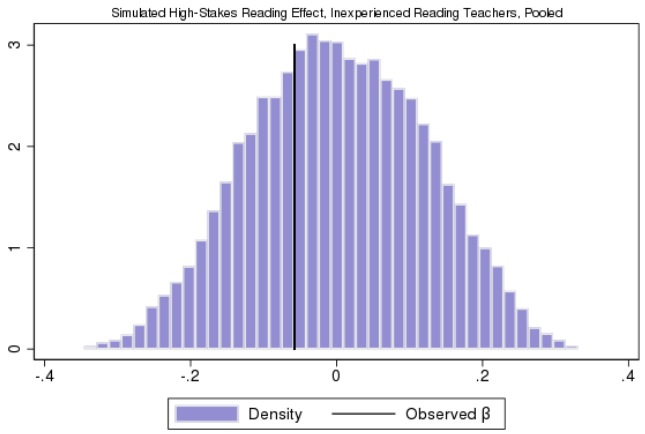
N = 10000 Replications. **P-value** (two-tailed) = 0.233



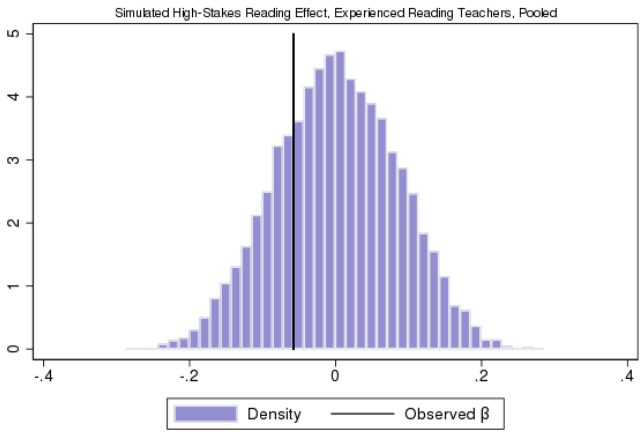
N = 10000 Replications. **P-value** (two-tailed) = 0.116



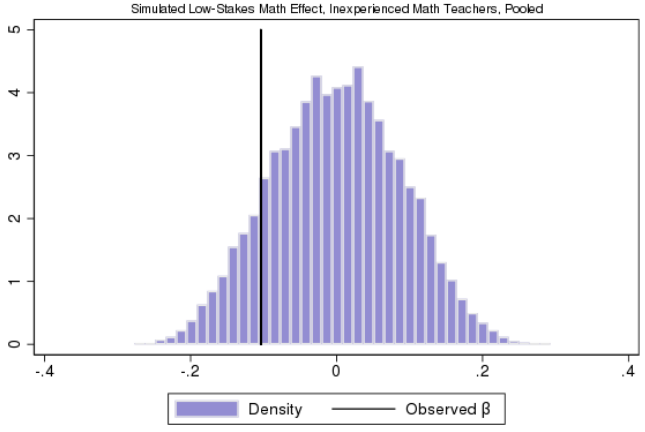
N = 10000 Replications. **P-value** (two-tailed) = 0.978



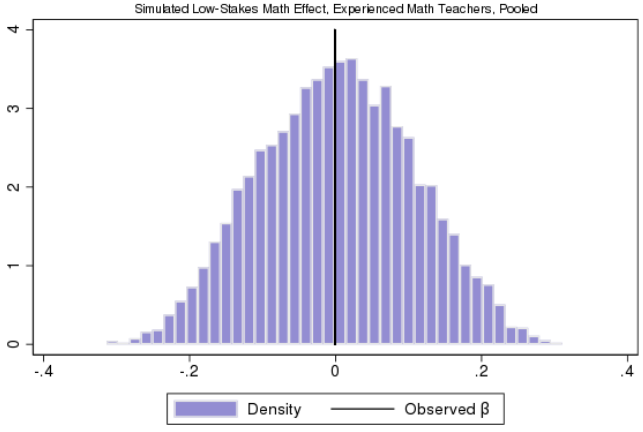
N = 10000 Replications. **P-value** (two-tailed) = 0.656



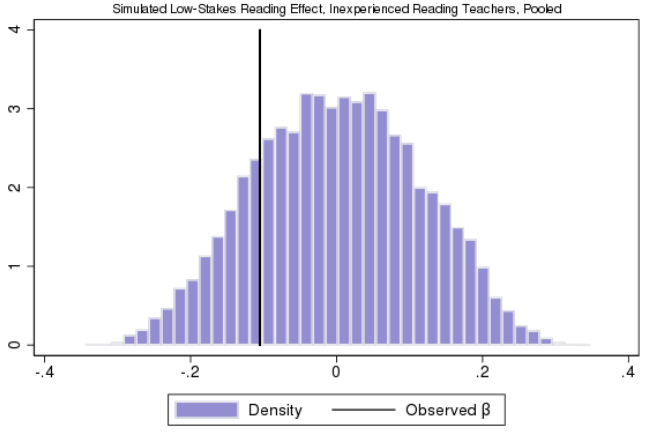
N = 10000 Replications. **P-value** (two-tailed) = 0.511



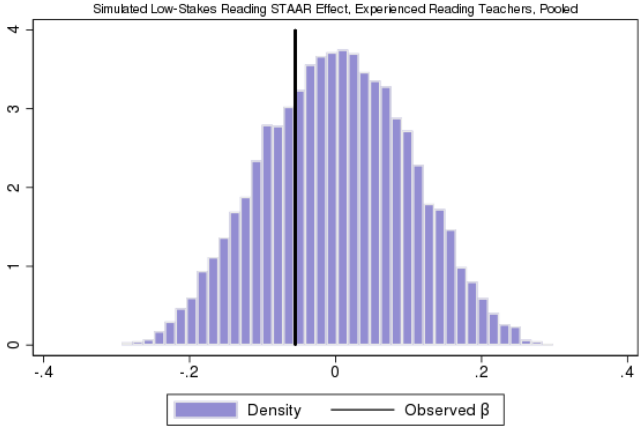
N = 10000 Replications. **P-value** (two-tailed) = 0.259



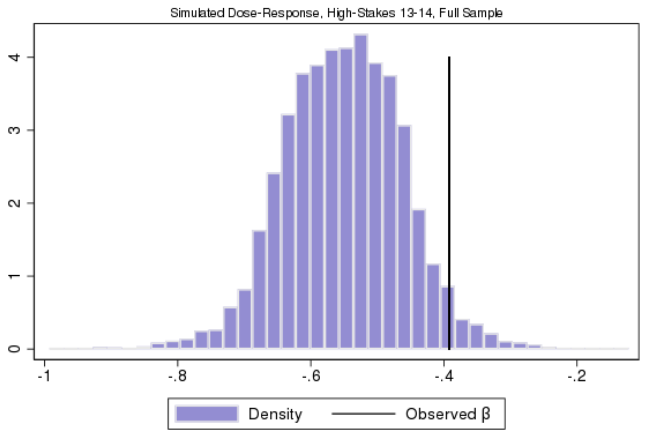
N = 10000 Replications. **P-value** (two-tailed) = 0.992



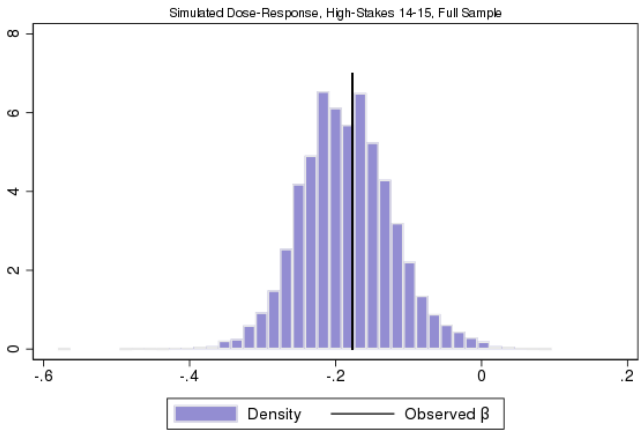
N = 10000 Replications. **P-value** (two-tailed) = 0.385



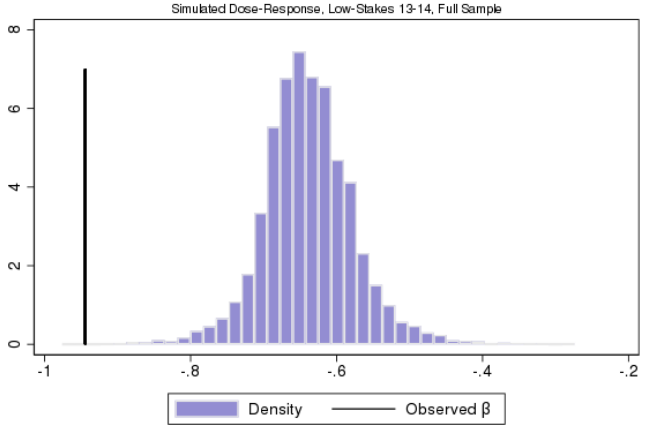
N = 10000 Replications. **P-value** (two-tailed) = 0.603



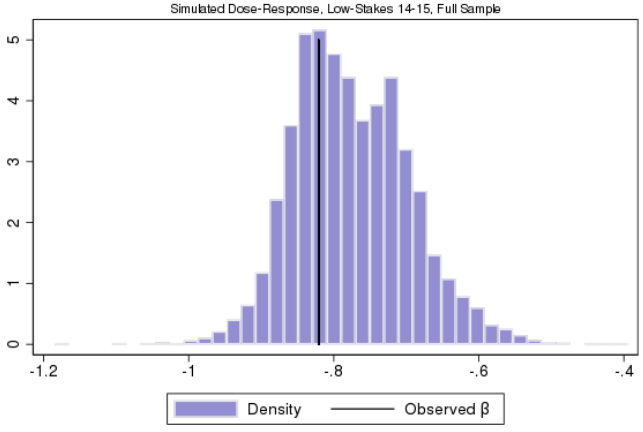
N = 10000 Replications. **P-value** (two-tailed) = 0.965



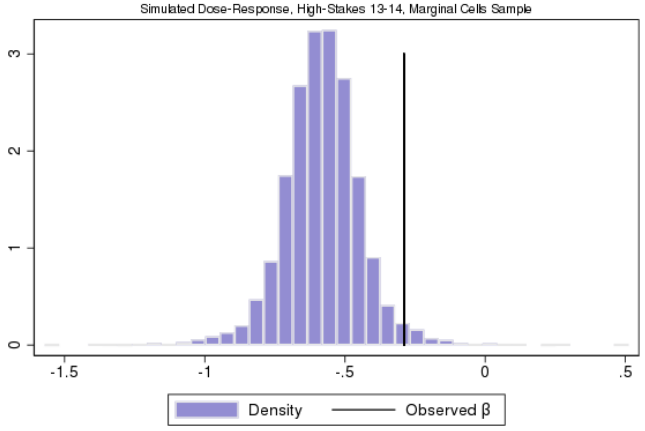
N = 10000 Replications. **P-value** (two-tailed) = 0.560



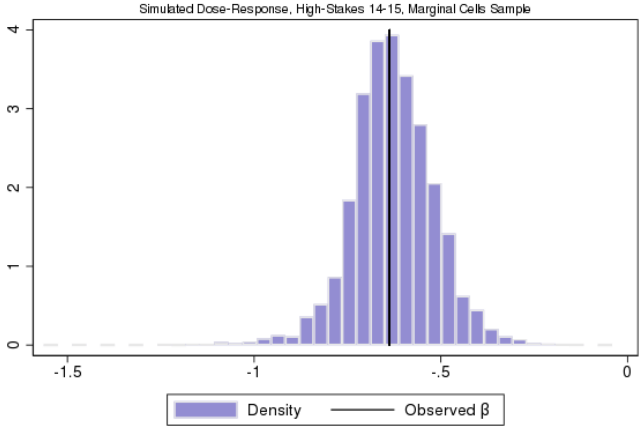
N = 10000 Replications. **P-value** (two-tailed) = 0.000



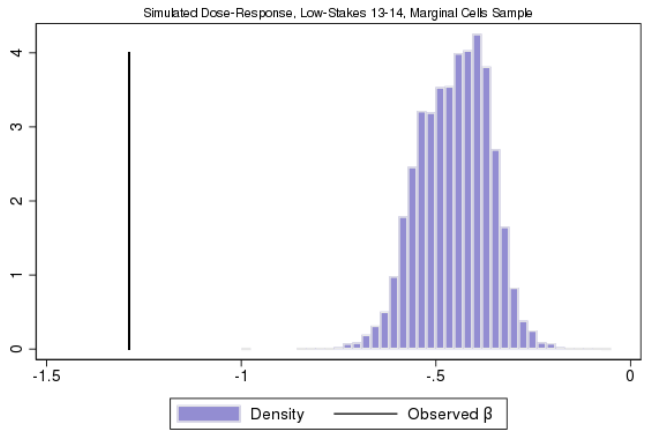
N = 10000 Replications. **P-value** (two-tailed) = 0.317



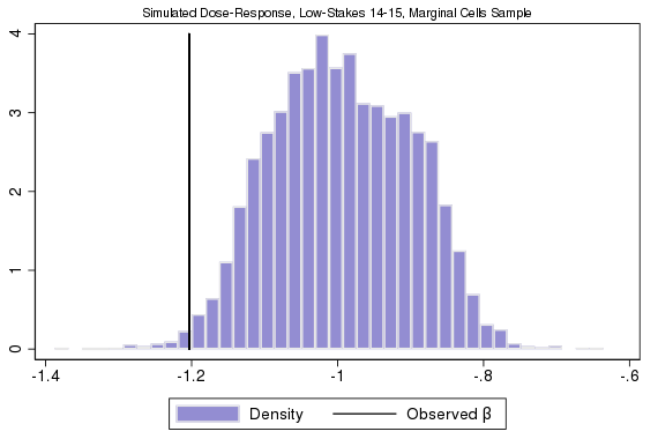
N = 10000 Replications. **P-value** (two-tailed) = 0.978



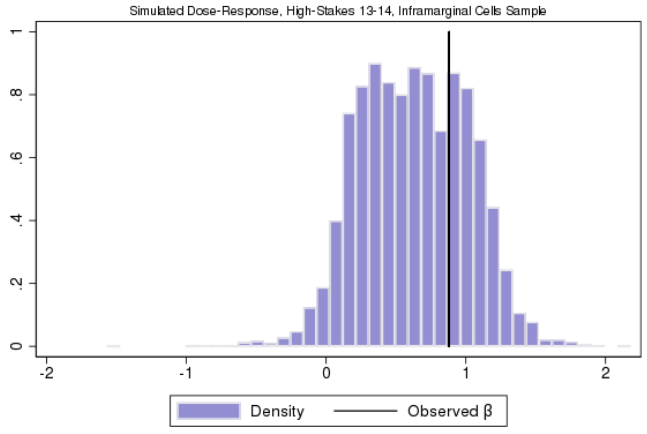
N = 10000 Replications. **P-value** (two-tailed) = 0.477



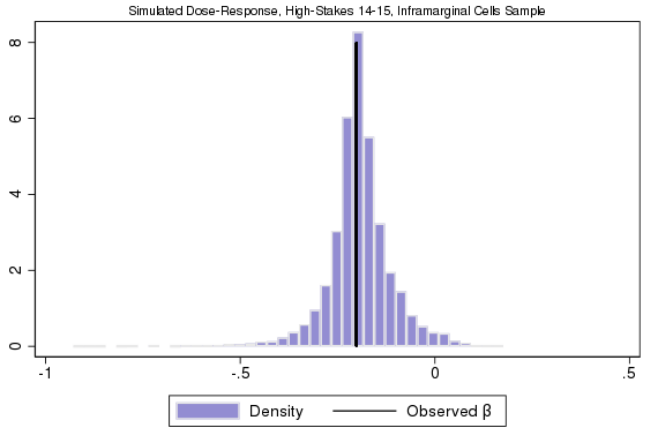
N = 10000 Replications. **P-value** (two-tailed) = 0.000



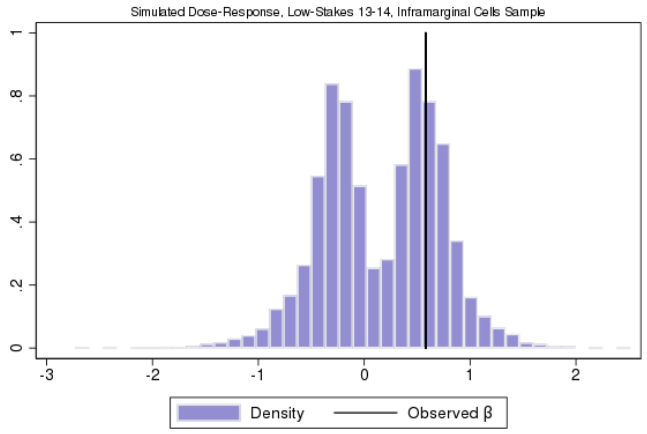
N = 10000 Replications. **P-value** (two-tailed) = 0.009



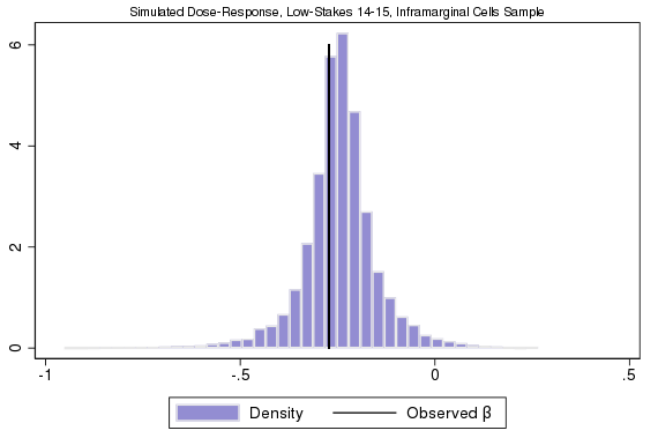
N = 10000 Replications. **P-value** (two-tailed) = 0.296



N = 10000 Replications. **P-value** (two-tailed) = 0.454



N = 10000 Replications. **P-value** (two-tailed) = 0.332



N = 10000 Replications. **P-value** (two-tailed) = 0.333

Notes: These figures plot the distribution of treatment coefficients obtained by conducting OLS regressions on 10000 re-randomized samples. Re-randomization is done by randomly assigning one school to treatment and one school to control within each matched pair 10000 times. The main specifications are re-run using this simulated treatment assigment and the simulated betas are stored. The exact two-sided p-value is the number of simulated betas that are greater than the observed beta in absolute value.

Appendix A: Technical Appendix

Proposition 1. *With teacher specialization, total student achievement increases if teacher's knowledge increases such that* $\sum_{j=1}^{N} f(e_j, \theta_j, H_s) - \sum_{j=1}^{N} f(e_j, \theta_j, H_{ns}) > N\left(\mu_\alpha - \sum_{j=1}^{N} \frac{\alpha_j}{N}\right)^2$.

Proof – Without teacher specialization:

$$E(\alpha_j | S) = \alpha_j$$

$$x_{ns}^* = \sum_{j=1}^{N} \frac{\alpha_j}{N}$$

$$Y_{ns}^* = \sum_{j=1}^{N} f(e_j, \theta_j, H_{ns}) - \left(\alpha_j - \sum_{j=1}^{N} \frac{\alpha_j}{N}\right)^2$$

With teacher specialization:

$$E(\alpha_j | S) = \mu_\alpha$$

$$x^* = \mu_\alpha$$

$$Y_s^* = \sum_{j=1}^{N} f(e_j, \theta_j, H_s) - (\alpha_j - \mu_\alpha)^2$$

The change in total student achievement is greater than 0 if

$$Y_s^* - Y_{ns}^* > 0$$

$$\sum_{j=1}^{N} f(e_j, \theta_j, H_s) - (\alpha_j - \mu_\alpha)^2 - \sum_{j=1}^{N} f(e_j, \theta_j, H_{ns}) - \left(\alpha_j - \sum_{j=1}^{N} \frac{\alpha_j}{N}\right)^2 > 0$$

Rewriting,

$$\sum_{j=1}^{N} f(e_j, \theta_j, H_s) - \sum_{j=1}^{N} f(e_j, \theta_j, H_{ns}) > N\left(\mu_\alpha - \sum_{j=1}^{N} \frac{\alpha_j}{N}\right)^2,$$

which was the desired result. QED.

Appendix B: Data Appendix

A. Administrative Data

Attendance Rates

Attendance rate is days present in class divided by days, or number of years, enrolled in Houston Independent School District (HISD). For 2013-2014, attendance rate is calculated as the number of years attended in 2013-2014. As expected, this variable ranges from 0 to 1. For 2014-2015, attendance rate is calculated as the number of years attended between 2013 to 2015. This variable ranges from 0 to 2.

Behavioral Infractions

A student is flagged as committing a behavioral incident if they show up in the HISD file cataloguing incidents resulting in a serious disciplinary action. At this time the file is only available for the 2013-2014 school year.

Economically Disadvantaged

We consider a student economically disadvantaged if he is eligible for free or reduced price lunch, or if he satisfies one or more of the following criteria:

- Family income at or below the official federal poverty line
- Eligible for Temporary Assistance to Needy Families (TANF) or other public assistance
- Received a Pell Grant or comparable state program of need-based financial assistance
- Eligible for programs assisted under Title II of the Job Training Partnership Act (JTPA)
- Eligible for benefits under the Food Stamp Act of 1977

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 is involved in either of these programs.

Special Education and Limited English Proficiency

These statuses are determined by a student's designation in the official Houston Enrollment file; they enter into our regressions as indicator variables. Special education indicates students participating in a special education instructional and related services program or a general education program using special education support services, supplementary aids, or other special arrangements. Limited English proficiency is a designation identified by the Language Proficiency Assessment Committee. We do not consider students who have recently transitioned out of LEP status to be of limited English proficiency.

Race/Ethnicity

We code the race variables such that the five categories – white, black, Hispanic, Asian and other – are complete and mutually exclusive. Hispanic ethnicity is an absorbing state. Hence "white" implies non-Hispanic white, "black" non-Hispanic black, and so on.

Race and gender variables are assigned using the most recently available data and are filled in from past years of data if unavailable in the most recent data. Designations that change over time (LEP, gifted, special education, and economically disadvantaged status) are assigned from enrollment data in the year previous to treatment.

Test Scores

We observe results from the State of Texas Assessments of Academic Readiness (STAAR), the Stanford 10, and the Iowa Test of Basic Skills. For ease of interpretation, we normalize all scores to have mean zero and standard deviation one by grade, subject, and year. Scores are normalized *after* cases in which students have multiple test scores have been resolved using the following procedures:

Fifth graders must meet certain standards on their state tests to advance to the next grade, and those who fail on their first attempt are allowed to take a retest approximately one month later. When selecting a score for students who take the retest, we select the first score where it exists and only take the retest score where the first is missing, though our results do not change if we instead choose the retest score, the mean of the two scores, or the higher score.

In addition to retest scores, students occasionally have multiple test scores if they took various versions of a test. Multiple STAAR test scores were eliminated using the following procedure: If a student took a test in both English and Spanish, the score from the English version was used. If a student took a test with accommodations and without, the score from the test without accommodations was used. If a student took an on-grade level test (i.e. the testing grade was the same as their grade in the enrollment file) and an off-grade level test, the on-grade level score was used. If the student tested in two grades neither of which matched the enrollment file, the one closest to the enrollment grade was used and if the grades were equidistant to the enrolled grade, the score from the lowest grade tested was used. If a student still had more than one score with all the same testing conditions at the end of this procedure, the mean of those scores was used. In the case of Stanford 10/ITBS scores, precedence was given to on-grade level test scores or the lowest grade tested if there was no on-grade level test. Scores from tests administered in English were given precedence over tests administered in Spanish. Our results do not change if we instead use the highest of all of a student's test scores.

Treatment

Treatment is defined as being enrolled in a treatment school as the last school attended in 2012-2013. Treatment takes on a value of 0 for students enrolled in a control school as the last school attended in 2012-2013. The EVER and YEARS variables used in the 2SLS regressions are calculated from administrative reports of the number of days a student attended each school in each year.

In the 2014 specification, EVER is equal to one if a student attended a treatment school for at least one day in the 2013-14 school year and zero otherwise. In the 2015 and pooled specifications, EVER is equal to one if a student attended a treatment school for at least one day in either year of treatment.

In the 2014 specification, YEARS ranges continuously from zero to one. In the 2015 and pooled specifications, YEARS ranges continuously from zero to two.

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 2013-14 and 2014-15.

Teacher Value-Added

HISD officials provided us with 2012-2013 value-added data for 4,581 teachers. HISD calculates Teacher Value-Added using the Education Value-Added Assessment System. We use the gain indices. We normalize these indices such that the average teacher in each subject has mean zero and the sample standard deviation is one.

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 2012-13 on test scores in 2011-12 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 2012-13 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. Among teachers with non-missing values of both measures, the correlation between the official measure of TVA and our calculated teacher effect is 0.8 in math and 0.6 in reading in the experimental sample. Throughout the paper, we use this author-calculated measure of teacher effects for any teacher who is missing the official TVA measures.

Student-Teacher Linkage

I 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). 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.

Teacher Retention

HISD provided teacher certification files for 2012-2013, 2013-2014, and 2014-15. Teacher retention for every school is calculated in each year as how many teachers from the previous year stayed in the same school in the given year.

Teacher Specialization

Administrative data-based measures of teacher specialization were constructed using the teacher certification files and files that link students to the teacher they have in each subject. In the linkage file, teachers were tagged for every subject and grade that they taught. These values were then summed to create continuous measures of specialization; other measures reported in the text are binary indicators of specialization that are created from the continuous measures. The average number of students per grade is calculated by summing the number of students in the student-teacher linkage file per subject per grade, and the average is taken over all subjects and grades in the school that the teacher is assigned to in the certification file. Since teachers occasionally showed up in multiple schools (likely due to teachers switching schools) teachers were only included in the analysis in the school that matched the certification file. Teachers were removed from the analysis if their description in the certification file indicated that they were a Special Education teacher.

Figure 1 uses the measure of departmentalization in the first row of Table 2 (where specialization is defined as not teaching both math and reading). The fraction of specialized teachers for every school is the total number of teachers teaching not teaching math and reading divided by the total number of teachers in the school who are present in the student-teacher linkage file, who are linked to the school in the certification file, and who are not Special Education teachers.

I additionally calculate a measure of teacher specialization that takes into account both the "quality" and "quantity" of specialization – the number of subjects a teacher teaches, the combination of subjects a teacher teaches, and the rightness of the match of the teacher to the correct subjects. Each of these components is conditional on the number of teachers in a teachers' grade-language cell, and therefore takes into account the constraints facing schools as they attempt to specialize their teachers.

First, one must decide what "ideal" specialization looks like for a cell of each size. In experimental schools, cells range in size from 1 to 8 teachers. Assuming that school schedules are completely flexible and all teachers can teach any of the four subjects, I determined the following to be "ideal" –

- 1 teacher per cell: 1 teacher teaches all four subjects
- 2 teachers per cell: 1 teacher teaches math and science, 1 teacher teaches reading and social studies
- 3 teachers per cell: 1 teacher teaches math, 1 teacher teaches reading, 1 teacher teaches science and social studies
- 4 teachers per cell: 2 teams of 2 teachers, each team is assigned the same way as a 2-teacher cell
- 5 teachers per cell: 1 team of 2 teachers and 1 team of 3 teachers, each team assigned the same way as a 2- or 3-teacher cell
- 6 teachers per cell: 2 teams of 3 teachers, each assigned the same way as a 3-teacher cell
- 7 teachers per cell: 2 teams of 2 teachers and 1 team of 3 teachers, each is assigned the same way as a 2- or 3-teacher cell

- 8 teachers per cell: 2 teams of 3 teachers and 1 team of 2 teachers, each is assigned the same way as a 2- or 3-teacher cell.

Note that if teachers are split into teams of different sizes, I assume they are responsible for teaching the appropriate proportion of students – i.e. in a 5-teacher cell, the 3-teacher team instructs three-fifths of the students in the cell and the 2-teacher team instructs two-fifths of the students in the cell.

Next, I assign teachers to the appropriate subjects in the ideal case using their (pre-treatment) TVA. HISD provides TVA only for a small subset of teachers; I calculate a measure of teacher effectiveness (controlling for student demographics and previous year test scores) in order to supplement this data. To maximize my ability to determine an ideal teacher assignment, I use my self-calculated measure for any teacher who does not have official HISD TVA. Both measures are separately standardized over the entire district to have a mean of zero and standard deviation one.

If you are missing TVA in all subjects, I assume you are a new teacher. Consistent with my previous experience in HISD, I assume that a new, untested teacher is better suited to teach math than an experienced teacher who is bad at teaching math (where "bad" teachers are those with TVA in a given subject that is below the average subject TVA in their school). However, if you are missing TVA in only one subject, I assume that you only have experience in the subject that you have TVA in. I assume that this is for some unobservable reason, and therefore, if you are a "reading only" teacher you will not be assigned to teach math over a bad math teacher.

Note that in assigning teachers to subjects, one must determine the extent to which we care about teachers' ability in each subject. Given that (i) students are not tested in science and social studies on state exams until 5th or 8th grade, respectively, (ii) instructional time is more heavily focused on math and reading, and (iii) many would argue that in elementary school, math and reading teach more fundamental skills that form a foundation for later learning, I chose to assign teachers to subjects in a way that maximizes student achievement in math and reading, rather than in all four subjects.

I calculate total student achievement under each possible teacher assignment. Note that – given the ideal subject assignments described above – in a 2-teacher cell, there are 2 possible teacher assignments. 3- and 4-teacher cells each have 6 possible assignments. 5-teacher cells have 120 possible assignments, 6-teacher cells have 90 possible assignments, 7-teacher cells have 1260 possible assignments, and 8-teacher cells have 5040 possible assignments. I choose the teacher assignment that maximizes total student achievement. In the case that there are multiple maximizing assignments, I record all of them (this occurs in the case where teachers are missing TVA).

Now, armed with my "ideal" teacher assignments in each cell, I calculate a number that measures the degree of specialization (for a given teacher in a cell, or averaged over a cell), which takes into account both the number of subjects a teacher teaches and how well a teacher is matched to the subjects that he or she is teaching. I calculate this degree of specialization in both the ideal scenario and the actual scenario. The measure of the percent of potential specialization utilized by a given cell is the actual degree of specialization divided by the ideal degree of specialization.

First, calculate the percent of the four main subjects that a teacher teaches. I scale this number so that a zero is perfectly unspecialized (teaches 4 subjects) and a 1 is perfectly specialized (teaches 1 subject). Second, calculate the rightness of match to subjects. I measure this as the distance between the vector of subjects you actually teach and the vector of subjects you teach in the ideal case. I scale this distance to range from 0 to 1 and subtract it from one to get a measure of "rightness" of match such that a 0 is perfectly wrongly matched and a 1 is perfectly rightly matched. In the "ideal" case, this value is always 1. In cases where there are multiple ideal assignments that lead to maximized student achievement, I use the assignment that minimizes the distance between the actual and ideal assignments – i.e., I give schools the benefit of the doubt when I am missing information. However, I then weight the rightness of match variable by the percent of TVA information that is non-missing for each teacher. The degree of specialization of each teacher is the mean of these two components that range from 0 to 1 where 0 is unspecialized and 1 is

specialized. I also take the mean value over each cell, this is the degree of specialization of the cell in either the ideal or actual scenario.

B. Survey Data

Survey outcomes are from a survey administered to teachers in treatment and control schools at the end of 2013-2014 school year.

Know students

Teachers could answer strongly disagree, disagree, somewhat disagree, neutral, somewhat agree, agree, or strongly agree to the following statement - "I have personal relationships with each and every one of my students." An indicator for this variable is equal to 1 when the teacher agrees to any extent, and is 0 otherwise.

Gives individual attention

Teachers could answer strongly disagree, disagree, somewhat disagree, neutral, somewhat agree, agree, or strongly agree to the following statement - "I feel that I have the time to give each student I teach the individual attention they need." An indicator for this variable is equal to 1 when the teacher agrees to any extent, and is 0 otherwise.

Rules enforced

Teachers could answer strongly disagree, disagree, somewhat disagree, neutral, somewhat agree, agree, or strongly agree to the following statement - "Rules for student behavior are consistently enforced by teachers in this school, even for students who are not in their classes." An indicator for this variable is equal to 1 when the teacher agrees to any extent, and is 0 otherwise.

Enthusiasm for teaching math

Teachers could answer strongly disagree, disagree, somewhat disagree, neutral, somewhat agree, agree, or strongly agree to the following statement – "I am enthusiastic about teaching Math." An indicator for this variable is equal to 1 when the teacher agrees to any extent, and is 0 otherwise.

Enthusiasm for teaching reading

Teachers could answer strongly disagree, disagree, somewhat disagree, neutral, somewhat agree, agree, or strongly agree to the following statement – "I am enthusiastic about teaching ELA/Reading." An indicator for this variable is equal to 1 when the teacher agrees to any extent, and is 0 otherwise.

Lesson differentiation

Teachers were asked "Of the total time you spend planning your lessons per week, what percentage of that time is spent differentiating your lessons for different student abilities?" This variable is the percentage that the teacher reported.

Job performance and satisfaction

Teachers could respond that they felt significantly decreased, slightly decreased, neutral, slightly increased, or significantly increased job satisfaction relative to the previous year. An indicator for this variable is 1 if the teacher responds above the median answer, and 0 otherwise.

Survey Measures of Teacher Specialization

Teachers were asked to mention all the subjects they taught. The first measure of specialization is whether or not a teacher teaches both math and reading (non-specialized) or not (specialized). According to teacher's reports of the subjects they taught, we calculated the total number of subjects each teacher taught in a school. Two additional measures of specialization are based on the number of subjects a teacher teaches. For each measure, if a teacher mentioned teaching less than or equal to 3 (or 2) subjects, we flagged him as a specialized teacher. If a teacher mentioned teaching more than 3 (or 2) subjects, we flagged him as a non-specialized teacher. Teachers also described how their classrooms at school were organized. The options were Departmentalized Instruction, Self-Contained Classes, or Team Teaching (with an explanation of each organizational structure), or Other. Teachers were flagged as specialized if they responded that Departmentalized Instruction best described their classes and not specialized if they chose any of the other responses. Teachers also self-reported which grades they taught, from which we calculated the total number of grades a taught.

Figure 1 uses the measure of departmentalization in the first row of Table 2 (where specialization is defined as not teaching both math and reading). The fraction of specialized teachers for every school is the total number of teachers teaching not teaching math and reading divided by the total number of teachers who responded to the question.