Abstract Title Page

Title: Missing Data and Mixed Results: The Effects of Teach For America on Student Achievement Revisited

Authors and Affiliations: Emily K. Penner, School of Education, University of California, Irvine

Background / Context:

This paper revisits existing experimental work on Teach For America (TFA) and extends it by examining treatment effects across the distribution of student achievement. TFA is a rapidly expanding teacher preparation program that currently serves over half a million students in low income districts across the country. Previously, Glazerman, Mayer, and Decker (2006) found positive effects of TFA in elementary school in math, but not reading, echoing quasi-experimental findings from several other studies (Kane, Rockoff, and Staiger, 2008; Noell and Gansle, 2009; and Xu, Hannaway, and Taylor, 2011). These results did not have notable variation by subgroup. However, these estimates were inaccurate due to the treatment of a non-response code as a valid response value. Revised estimates confirm positive effects for math and not reading, but show that TFA teachers were especially effective for African American students, but not Hispanics, and for females, but not males.

In addition to examining differences across subgroup, others have argued that a distributional approach is important for thoroughly investigating policy interventions because examinations focused solely on mean impacts might obscure large countervailing differences in program impacts that offset one another (Bitler, Gelbach, and Hoynes 2006). Thus, to deepen our understanding of the effect of TFA on student achievement, this study investigates distributional as well as mean impacts. New distributional results reveal TFA teachers are especially beneficial for the upper-middle of the distribution of math achievement, but not the upper and lower tails.

Purpose / Objective / Research Question / Focus of Study:

This study examines whether the effect of TFA varies across student subgroups and the distribution of student achievement. Previous work using these data did not find any notable subgroup differences. However, in light of coding errors which are described below, the first research question examines whether there are overall and subgroup differences after correcting for coding errors.

The second research question examines whether there are distributional differences in the effect of TFA, which has not yet been examined in the existing research on TFA. Several competing hypotheses are tested. While there is little research documenting the TFA training and coaching process, materials developed by TFA staff suggest that TFA focuses on setting goals and monitoring progress for all learners, and overcoming student struggles (Farr 2010). This commitment might lead TFA corps members to have especially large impacts on the lower portion of the distribution because these students are overlooked by other teachers.

Alternatively, it could be that TFA teachers have an affinity towards students who are especially focused and academically advanced. Perhaps they see these students as the most likely to complete their K-12 education and attend college, so they give them extra support. In this case, TFA teachers would have a larger impact on the top of the distribution than the bottom.

Finally, it may be the case that the tails of the distribution, the very lowest achievers and the very highest achievers are not exceptionally impacted by having TFA teachers (at the lower tail because they are in need of special education and behavioral assistance or at the upper tail because they are exceptionally gifted and will do well regardless of teacher support). In this case, the middle of the distribution would benefit from the training and effort of a TFA teacher, but not the tails.

Setting:

This study uses data collected by Mathematica Policy Research (MPR) from low-income schools in 6 TFA regions throughout the country, including Baltimore, Chicago, Los Angeles, Houston, New Orleans, and the Mississippi Delta.

Population / Participants / Subjects:

To understand whether the effect of TFA varies across subgroups and the distribution of student achievement, this study estimates the effect of being randomly assigned to a TFA teacher or a non-TFA teacher. It uses experimental data, collected during the 2001-2002 and 2002-2003 school years. The study was restricted to teachers in grades 1 to 5. The final sample included 100 classrooms at 17 schools, for a total of nearly 2,000 students. Descriptive statistics for the sample are presented in Table 1.

[Insert Table 1 Here]

Intervention / Program / Practice:

This study evaluates the effect of TFA on student achievement in elementary school. Teach For America is a national program that recruits recent college graduates and professionals to commit to work in low-income schools for two years. It provides pre-service training to its corps members in the summer prior to their first teaching assignment and ongoing coaching and mentoring throughout their two-year stint in the program. In 2012 alone, more than 10,000 current TFA corps members serve 750,000 students in 46 regions across 36 states and the District of Columbia. TFA has a presence in 18 of the 20 most populous metropolitan areas in the US, as well as many suburban and rural regions. TFA has also inspired similar programs in districts and states across the country and has launched partner programs under the Teach For All network in 25 countries in five continents. Given its important role in the education policy landscape, it is important to understand the impact it has on the diverse groups of students it serves.

Research Design:

Data for this project were collected using a randomized control trial design. 6 TFA regions were selected by stratifying on urbanicity and student race. Within each region, schools were randomly selected from those that had the staffing needed to support the design. Within such schools students in grades 1-5 were randomly assigned to TFA or non-TFA classrooms at each grade level for which there was at least one TFA and one non-TFA teacher. “Treatment teachers” were current and former TFA corps members and “Control teachers” were any other teachers at the same grade levels. Student and teacher demographic data and student test scores (using the Iowa Test of Basic Skills, ITBS) were collected at the start of the school year. Post-test scores were collected at the end of the same school year (Glazerman et al. 2006).

Data Collection and Analysis:

This project builds on the work of Glazerman et al. (2006), who estimated average effects of assignment to a TFA teacher using these data. They found average effects of roughly three percentage points (equivalent to one additional month of instruction) in math, but not reading. This paper revises the work of Glazerman and colleagues, who treated a non-response test score value as a valid score. 18 percent of students were found to have a reading raw score of 99, when the next highest raw score was approximately 40. These students had corresponding percentile and normal curve equivalent scores of 0. Verification with the ITBS publishers confirms that a raw score of 99 is not a valid score (http://www.riversidepublishing.com/products/itbs/details.html). It is therefore likely that these values represent a non-response category coded by NPR. However, the sample sizes, means, and standard deviations presented by Glazerman and colleagues suggest that these invalid scores were incorporated in their estimates. Table 2 shows a comparison of sample sizes with and without the included non-response values, overall and by subgroup, for pre-tests in math and reading. This non-response code was widespread, but especially pervasive in reading in first grade. Group differences in the occurrence of the non-response score are significant for most subgroup comparisons.

[Insert Table 2 Here]

This paper estimates average treatment effects using OLS regression with fixed effects for block randomization (which are grade specific, and thus account for grade differences). It adjusts for the same list of covariates used by Glazerman et al., including pre-test score. Rather than estimate a model with students nested within blocks, this paper adjusts for similarities in students at the school level by clustering at the block level.[1]

This paper also estimates quantile treatment effects following the assumptions and procedures outlined in Bitler, Hoynes, and Gelbach (2008). Briefly, instead of comparing the mean of test score differences, this paper examines how the shapes of the distribution of the treatment changes relative to that of the control in ways that are not captured by the mean. To make such a comparison in an experimental setting, rather than comparing the average of the treatment relative to the control, quantile treatment effects (QTE) are estimated by calculating the difference in the two marginal distributions (cumulative distribution functions, or CDFs) under the potential outcomes framework. From these CDFs, I examine the difference between these two distributions at various quantiles of the outcome variable. For example, I can estimate the QTE at the 0.50 quantile by subtracting the control group's sample median from the treatment group's sample median. Graphically, QTE estimates are the differences in the inverse CDFs of the outcome for the treatment and control groups.

As an example, Figures 1 and 2 show the inverse CDFs and QTE for the baseline math scores. Figure 1 shows the inverse CDF for the baseline math scores in the treatment and control groups. The vertical distance between these inverse CDFs at each point in the distribution is the quantile treatment effect at that point or quantile. Figure 2 shows the corresponding QTE for the inverse CDFs shown in Figure 1 for the baseline math percentile scores (solid red line), along with 90% confidence intervals (dashed lines), calculated by stratifying on block and treatment status and bootstrapping. Figure 2 shows that the bulk of the QTE point estimates are zero or close to zero for the baseline scores. The exception is the upper portion of the distribution near the 90th percentile. This suggests some distributional imbalance in random assignment.

Findings / Results:

Re-estimations of the mean treatment effects affirm previous conclusions that TFA has a positive effect on math, but not reading. The magnitudes of these estimations are reduced when the invalid scores are removed, but the substantive conclusions remain unchanged. However, some new subgroup conclusions are revealed by these differences. Two notable findings emerge.

[1] Analyses using the alternative method yield substantively similar results.

Whereas, previous estimates found no evidence of positive effects for any particular racial subgroup in math, revised estimates indicate that TFA has a significant positive effect for black students, but not for Hispanics. In earlier estimates, both boys and girls had significant treatment effects in math. However, revised estimates reveal that there is only a significant treatment effect for girls, but not boys. There are additionally some positive and significant effects for reading at some grades (1, 2, and 5), but not others. As the non-response issues remove most of the first grade sample, this estimate is not definitive. Overall, this suggests some significant and positive effect of TFA in reading is masked by analyzing the grades together. These results are presented in Table 3.

[Insert Table 3 Here]

The QTE results are presented in Figures 3 and 4. The QTE results suggest that TFA has impacts that vary across the distribution of math, but not reading. In math, TFA has a significant, positive effect for the upper-middle of the achievement distribution. In contrast, there is no effect, positive or negative at the upper and lower tail of the distribution. This difference is shown in Figure 3, where the confidence interval is above the zero line for several quantiles near the 60^{th} and 80^{th} quantiles. In reading, there are no quantiles at which the confidence intervals cross the zero line, suggesting no significant distributional differences.

[Insert Figures 3 and 4 Here]

Conclusions:

This paper makes two primary contributions to the literature on TFA, which provide evidence of treatment heterogeneity in its effect on student achievement. First, it corrects coding errors in previous work to reveal important subgroup differences by gender and race. Second, it identifies distributional differences in the effects of TFA in math. Together these findings suggest particular strengths and weaknesses of the TFA program in elementary school.

The subgroup findings highlight several connections to research on other types of policy interventions. Other intervention policies have identified particular effects for blacks, but not those of other races, such as the voucher interventions examined by Howell, Wolf, Campbell, and Peterson (2002), so perhaps this finding is not surprising. Although girls typically underperform relative to boys in mathematics beginning in elementary school (Rathburn et al. 2004), and boys have been more responsive to previous elementary-level math interventions (Arnold et al. 2002), this study finds that girls benefitted from having a TFA teacher, but boys did not.

The distributional findings suggest that TFA teachers are particularly beneficial to students in the upper-middle of the distribution. The effect of TFA teachers does not differ from that of non-TFA teachers at either the upper or lower tail of the distribution. Given that nearly all students that TFA serves score below the national average, this suggests that TFA teachers are more effective than non-TFA teachers for above-average students in their own classrooms, but which would typically be classified as low performing elsewhere.

Taken together, these findings provide more detailed information about which types of students TFA best serves. They also underscore that there are no subgroups or points along the distribution for which TFA teachers are significantly less effective than non-TFA teachers. These experimental results are suggestive of a pattern that will be tested in future work using state-wide data to test the same research questions in a general equilibrium context.

Appendix A. References

Arnold, D. H., Fisher, P. H., Doctoroff, G. L. & Dobbs, J. (2002). Accelerating Math Development in Head Start Classrooms. *Journal of Educational Psychology*, 94(4), 762-770. doi: 10.1037/0022-0663.94.4.762.

Bitler, M. P., Gelbach, J. B., & Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *The American Economic Review*, 96(4), 988-1012. doi:10.1257/aer.96.4.988.

Glazerman, S., Mayer, D., & Decker, P. (2006). Alternative routes to teaching: The impacts of Teach for America on student achievement and other outcomes. *Journal of Public Policy Analysis and Management*, 25(1), 75-96. doi: 10.1002/pam.20157.

Howell, W. G., Wolf, P. J. , Campbell, D. E., & Peterson, P. E. (2002). School Vouchers and Academic Performance: Results from Three Randomized Field Trials. *Journal of Policy Analysis and Management*, vol. 21(2), 191–217. doi: 10.1002/pam.10023

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about effectiveness? Evidence from New York City. *Economics of Education Review*, 27(6), 615-631. doi: 10.1016/j.econedurev.2007.05.005

Noell, G. H., & Gansle, K. A. (2009). Teach for America Teachers' Contribution to Student Achievement in Louisiana in Grades 4-9: 2004-2005 to 2006-2007. Technical Report. *Baton Rouge, Louisiana State University*. Retrieved from: http://www.givebetter.net/files/unitedstates/TFA/Noell%20and%20Gansle.%202009.%20Teach%20For%20America%20teachers%E2%80%99%20contribution.pdf

Rathburn, A., West, J., & Germino-Hausken, E. G. (2004). *From Kindergarten through Third Grade: Children's Beginning School Experiences.* National Center of Education Statistics, Washington, DC. Retrieved from: http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED483078&ERICExtSearch_SearchType_0=no&accno=ED483078.

Riverside Publishing. (2012). Iowa Test of Basic Skills (ITBS), Forms A, B, and C. Retrieved from: http://www.riversidepublishing.com/products/itbs/details.html.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics* 125(1),175–214. doi: 10.1162/qjec.2010.125.1.175.

Xu, Z., Hannaway, J., & Taylor, C. (2011). Making a difference? The effects of Teach for America in high school. *Journal of Policy Analysis and Management*, 30(3), 447-469. doi: 10.1002/pam.20585

Table 1. Demographic Characteristics of Students in Study Sample

	Percent of Sample	N[1]
Female	49%	881
Black	68%	1220
Hispanic	29%	465
Overage for grade	21%	321
Free/Reduced Lunch Eligible	98%	1365
Did not move classes during school year (Stayer)	93%	1664
Grade		
First	19%	332
Second	10%	175
Third	35%	619
Fourth	27%	490
Fifth	10%	173
Total	100%	1789

[1] Sample includes only those with a post-test

Table 2. Occurrence of Mis-Coded Non-Response Values in Normal Curve Equivalent Pretest Scores, Overall and by Demographic Characteristics

	Mathematics					Reading				
	N with 00s	N without 00s	Total N	coefficient for group differences	p-value for group differences in 00s	N with 00s	N without 00s	Total N	coefficient for group differences	p-value for group differences in 00s
Full Sample[1]	55	1,734	1,789		-	322	1,467	1,764		-
Treatment	33	767	800	-0.04	0.02	128	672	800	0.02	0.05
Control	22	967	989			194	795	989		
First Grade	1	331	332	0.01	0.00	258	74	332	-0.20	0.00
Second Grade	1	174	175			34	141	175		
Third Grade	15	604	619			24	595	619		
Fourth Grade	33	457	490			5	485	490		
Fifth Grade	5	168	173			1	172	173		
Female	29	852	881	0.00	0.60	175	706	881	0.04	0.04
Male	26	882	901			147	761	908		
Black	241	979	1220	0.03	0.00	51	1169	1220	0.06	0.01
Non-Black	81	488	569			4	565	569		
Hispanic	3	462	465	-0.04	0.00	68	397	465	-0.02	0.31
Non-Hispanic	49	1077	1126			188	938	1126		
Overage	10	311	321	0.00	0.83	27	294	321	-0.12	0.00
Not overage	41	1178	1219			250	969	1219		
Free/Red. Lunch	22	1343	1365	-0.02	0.40	269	1096	1365	0.20	0.01
Non-Free/Red. Lunch	1	26	27			0	27	27		
Stayer	52	1612	1664	0.01	0.65	301	1363	1664	0.01	0.72
Mover	3	122	125			21	104	125		

[1] Full sample includes only those with a post-test

Table 3. Regression Results, Full Sample and Student Subgroups

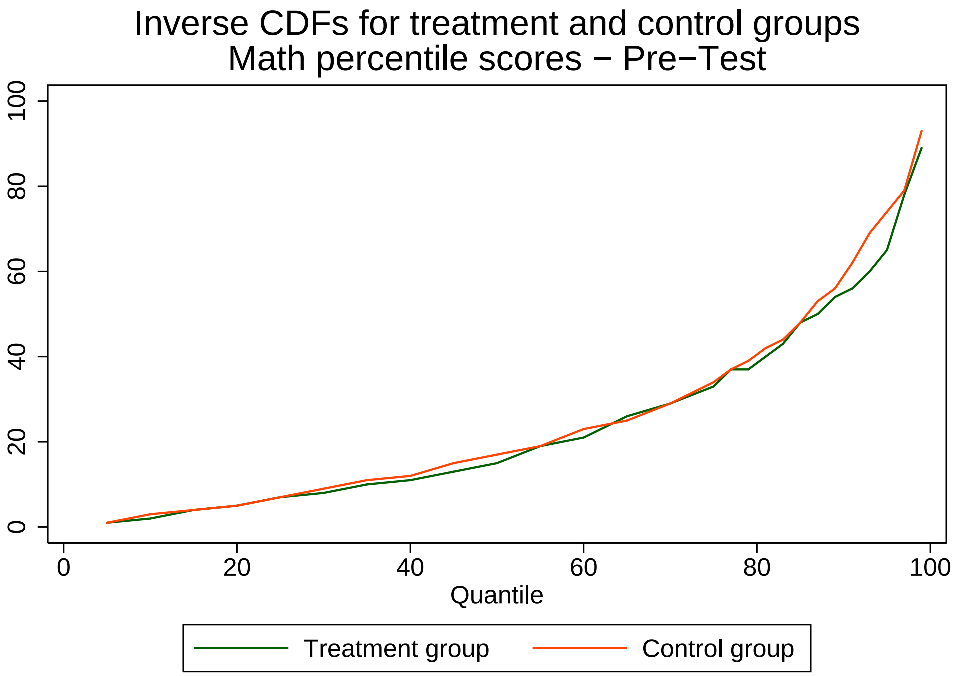
	Mathematics						Reading					
	Control Mean[1]	TFA Mean	Impact	P-value	Students[2]	Classes	Control Mean	TFA Mean	Impact	P-value	Students	Classes
Total	**31.59**	**34.22**	**2.63**	**0.01**	**1710**	**100**	32.70	33.79	1.09	0.27	1495	82
Subgroups												
Females	**31.27**	**34.61**	**3.34**	**0.00**	**837**	**100**	34.18	35.07	0.89	0.44	721	82
Males	31.87	33.87	2.01	0.11	873	100	31.32	32.61	1.29	0.30	774	82
Race/Ethnicity												
African American	**27.31**	**29.80**	**2.49**	**0.04**	**1150**	**88**	28.93	28.38	-0.55	0.60	1007	69
Hispanic	37.20	39.24	2.05	0.14	650	48	39.10	40.97	1.87	0.31	528	38
Overage for Grade												
Overage	29.70	31.75	2.05	0.18	292	78	26.79	27.87	1.08	0.41	295	66
Not Overage	**33.06**	**36.06**	**3.01**	**0.01**	**1173**	**87**	36.16	37.56	1.40	0.27	997	71
Missing Age	26.41	28.91	2.50	0.52	245	28	24.31	22.86	-1.45	0.46	203	21
Mobility Status												
Stayers	**31.71**	**34.21**	**2.50**	**0.01**	**1614**	**100**	33.15	33.75	0.60	0.56	1409	82
Movers	31.56	32.06	0.50	0.94	96	47	27.82	31.43	3.62	0.50	86	40
Initial Achievement												
Low	**19.11**	**21.94**	**2.83**	**0.01**	**552**	**91**	18.79	17.51	-1.28	0.95	450	71
Middle	**27.22**	**30.79**	**3.57**	**0.01**	**534**	**94**	29.64	29.58	-0.06	0.96	461	74
High	**47.68**	**50.99**	**3.31**	**0.00**	**556**	**94**	47.25	49.67	2.42	0.08	430	74
Grade Level												
Grade 1	30.82	31.56	0.73	0.74	324	23	**43.87**	**101.16**	**57.29**	**0.02**	**63**	**5**
Grade 2	22.09	28.21	6.12	0.08	172	10	**31.69**	**34.91**	**3.22**	**0.01**	**170**	**10**
Grade 3	34.02	37.60	3.59	0.10	593	34	33.14	33.62	0.48	0.75	607	34
Grade 4	32.22	34.49	2.26	0.15	452	24	30.30	31.32	1.02	0.45	484	25
Grade 5	30.41	35.64	5.23	0.34	169	8	**28.88**	**30.32**	**1.44**	**0.02**	**171**	**8**

Source: Scores are from the Iowa Test of Basic Skills. Scores are reported as Normal Curve Equivalent Scores, whose national average score is 50 with a standard deviation of 21.06. Rows appearing in bold represent statistically significant estimates at the .05 level or better, two tailed test.

[1] Means and impacts are regression adjusted. Regression models include controls for baseline test scores, gender, race/ethnicity, eligibility for free/reduced price lunch, age (whether over age for grade), and percentage of students who were not in the research sample. Models also include randomization block fixed effects. Standard errors are clustered at the block level to account for non-independence of same-school observations.

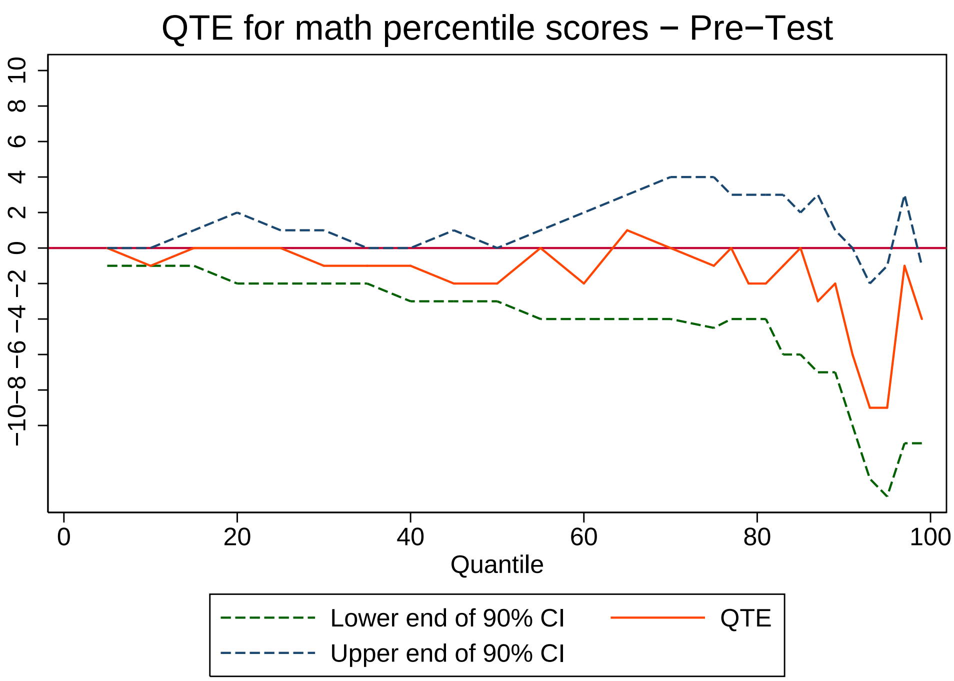
[2] Sample differs from previous page due to missing values on math post-test.

Figure 1: Inverse CDF for math percentile pre-test scores



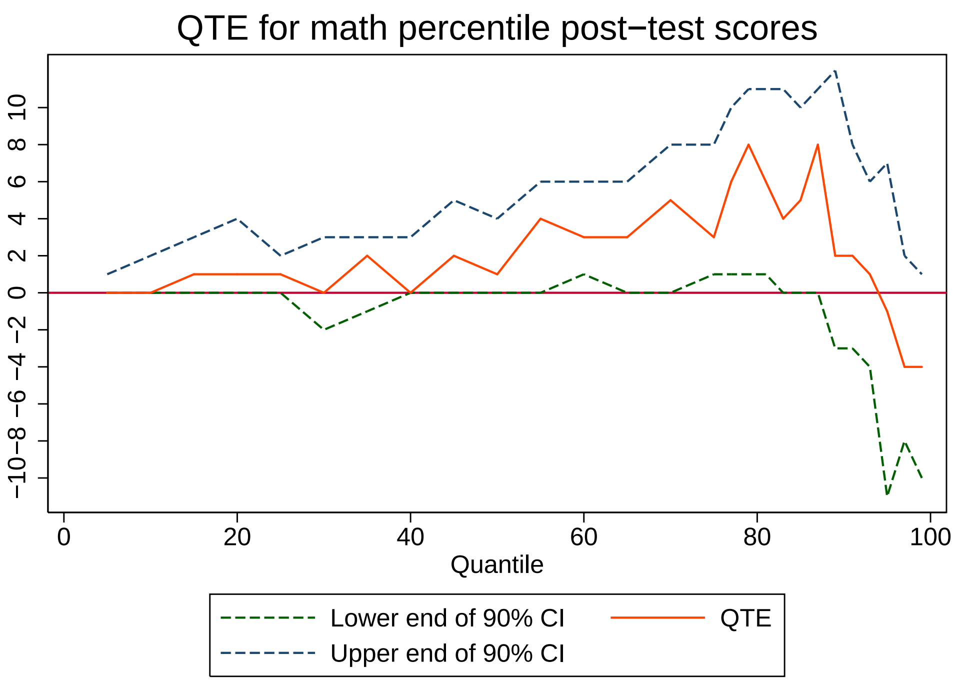
Notes: Figure shows inverse CDF for pre-treatment differences on math percentile scores - Assessment: Iowa Test of Basic Skills.

Figure 2: Quantile treatment effects for math percentile pre-test scores



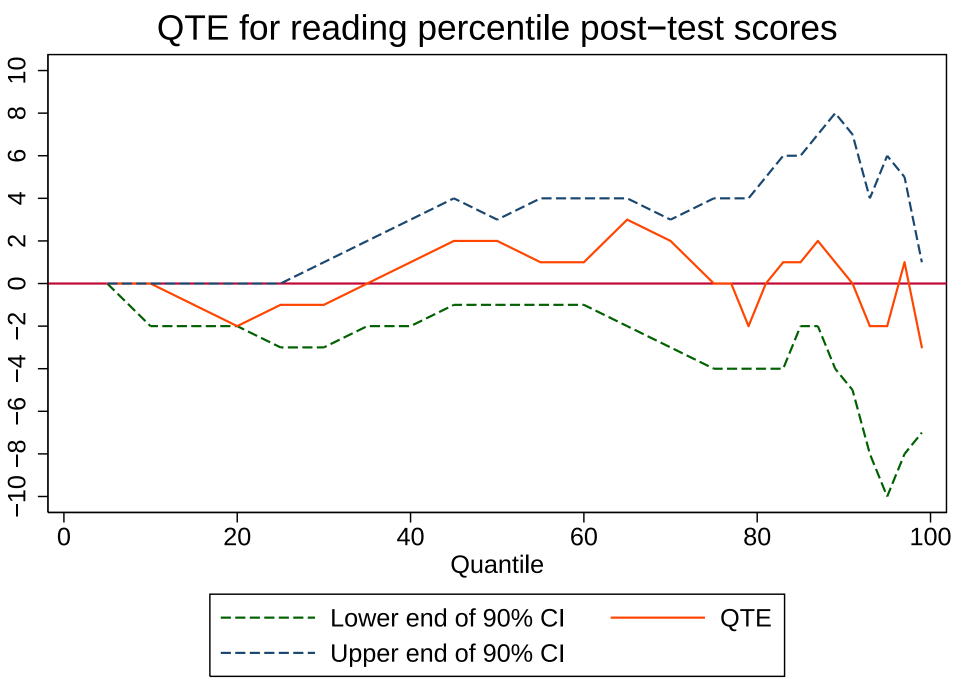
Notes: Figure shows QTE for pre-treatment differences on math percentile score - Assessment: Iowa Test of Basic Skills. CIs are obtained by bootstrapping by blockid and treatment condition.

Figure 3: Quantile treatment effects for math percentile post-test scores



Notes: Figure shows QTE for the effect of assigment to a Teach for America classroom on math percentile scores - Assessment: Iowa Test of Basic Skills. CIs are obtained by bootstrapping by blockid and treatment condition.

Figure 4: Quantile treatment effects for reading percentile post-test scores



Notes: Figure shows QTE for the effect of assigment to a Teach for America classroom for reading percentile scores - Assessment: Iowa Test of Basic Skills. CIs are obtained by bootstrapping by blockid and treatment condition.