Working Paper No. 102

Heterogeneous effects of pupil-to-teacher ratio policies – A look at class size reduction and teacher aide

Simone Balestra and Uschi Backes-Gellner

Universität Zürich
IBW – Institut für Betriebswirtschaftslehre

UNIVERSITÄT BERN

Working Paper No. 102

Heterogeneous effects of pupil-to-teacher ratio policies – A look at class size reduction and teacher aide

Simone Balestra and Uschi Backes-Gellner

April 2017 (first version: May 2014)

This paper was previously circulated under the title "Revisiting Class-Size Effects: Where They Come From and how Long They Last" (2014).

Please cite as:

"Heterogeneous Effects of Pupil-to-Teacher Ratio Policies - A Look at Class Size Reduction and Teacher Aide." Swiss Leading House "Economics of Education" Working Paper No. 102, 2017 (first version 2014). By Simone Balestra and Uschi Backes-Gellner.

Heterogeneous effects of pupil-to-teacher ratio policies – A look at class size reduction and teacher aide*

Simone Balestra† and Uschi Backes-Gellner‡

April 2017

Abstract

This paper investigates the effect of two pupil-to-teacher ratio policies on test scores for children with different achievement levels. Using data from a large randomized experiment in early childhood, we estimate unconditional quantile treatment effects of small class and teacher aide, as compared to regular classes. For the small class intervention, results show that pupils in the middle of the achievement distribution profit the most from being assigned to a smaller class, whereas pupils at the bottom or at the top of the achievement distribution experience almost no gain in test scores. For the teacher aide intervention, the analysis reveals positive and significant effects for students at the bottom of the achievement distribution, an effect stronger for boys and disadvantaged pupils. The findings suggest that the average effects reported in traditional empirical studies on pupil-to-teacher ratio interventions provide an incomplete characterization of the impact on the achievement distribution, thus constituting a weak guide for policymakers.

Keywords: class size; teacher aide; unconditional quantile regression; kindergarten.

JEL Classification: C21; I21; J13.

* We thank Eric Bettinger, Tor Eriksson, Simon Janssen, Ofer Malamud, Steffen Müller, Carmit Segal, Ludger Wössmann, participants in the APPAM fall research conference, and participants in the Canadian Economics Association annual conference for their constructive comments. This study is partly funded by the Swiss Federal Office for Professional Education and Technology through its Leading House on the Economics of Education, Firm Behavior and Training Policies. The authors declare that they have no conflict of interest.

† University of St. Gallen, Switzerland. Telephone: +41 (0)71 224 23 18; e-mail: simone.balestra@unisg.ch. Address: Rosenbergstrasse 51, CH-9000 St. Gallen, Switzerland. Simone Balestra is the corresponding author.

‡ University of Zurich, Switzerland. Telephone: +41 (0)44 634 42 81; e-mail: backes-gellner@business.uzh.ch. Address: Plattenstrasse 14, CH-8032 Zurich, Switzerland.

1 Introduction

In the last two decades, the subject that has received the most academic attention in education policy is pupil-to-teacher ratio. Reducing pupil-to-teacher ratio to improve student achievement is a policy measure that has gained major consideration in the U.S., Europe, and Australia. Despite the remarkable amount of literature on pupil-to-teacher ratio, almost all studies focus on the average effects that pupil-to-teacher ratio has on a wide spectrum of very heterogeneous students (West and Woessmann, 2006). This focus on average effects is surprising, given the policy relevance of class size effects for children at different levels of achievement. In addition, the exclusive attention on class size reduction (CSR) is also surprising, because CSR is only one way of diminishing pupil-to-teacher ratio. Another possibility is the use of teacher aide (TA), which is often viewed as a less expensive alternative to smaller classes (Hough, 1993). Still, there is little empirical research investigating the effect of TA on student achievement, and none in a distributional framework.

This study uses a recently developed estimator, unconditional quantile regression (Firpo et al., 2009), to examine heterogeneous effects of CSR and TA, as compared to a regular class. We use data from the kindergarten cohort of the Tennessee Student/Teacher Achievement Ratio experiment (Project STAR), the largest education experiment in U.S. history. Although Project STAR data are three decades old, no other intervention has randomly assigned students to smaller and larger classes in a substantial number of schools (Chingos, 2013). Moreover, the introduction of new technologies in schools such as computers appear to have not impacted student achievement (Bulman and Fairlie, 2016). For these reasons, the STAR data are still being used in the present days (Chetty et al., 2011; Dynarski et al., 2013; Jackson and Page, 2013).

Our results show that pupils in the middle of the achievement distribution profit the most from a CRS, whereas pupils at the bottom or at the top of the achievement distribution experience almost no gain in test scores. For the TA intervention, the analysis reveals positive and significant effects for students at the bottom of the achievement distribution. For pupils at the bottom of the achievement distribution, the effect size of the TA intervention is as large as that

of a CSR. According to our analysis, both CSR and TA interventions are particularly beneficial for boys, ethnic minorities, and children with lower socio-economic background.

We make three contributions to the literature in applied economics. First, from a methodological perspective, unconditional quantile regression estimates a direct measure of how a marginal change in pupil-to-teacher ratio affects the distribution of achievement in the population, keeping the distribution of other characteristics equal. This differs from the commonly used conditional quantile regression (Koenker and Bassett, 1978), because conditional quantile regression estimates treatment effects conditional on the mean value of included covariates, and the interpretation of such treatment effects change when different sets of covariates are entered into the regression equation. In such cases, the estimated effects do not translate to relevant policy questions that are linked to the covariates of interest (Borah and Basu, 2013). For this reason, we argue that unconditional quantile regression is more appropriate than conditional quantile regression in policy analyses such as that of the present study.

Second, we provide a comprehensive analysis of two different pupil-to-teacher ratio policies, revealing heterogeneous effects that are useful to policymakers and school managers. Our analysis shows that only CSR increases average test scores, an increase driven by students in the middle of the achievement distribution. On the contrary, TA does not increase average test scores, but it has positive and significant effects for students at the bottom of the achievement distribution. We thus conclude that the two policies are not interchangeable, because they affect different populations of students.

Third, in terms of equity, the effect of CRS on the gap between low and high achieving students is not clear, because the impact of CSR is concentrated in the middle of the distribution. By contrast, TA would be an effective policy for reducing the achievement gap, because the TA intervention raises the test scores at the bottom of the distribution. TA policies may thus be particularly helpful for classes with large proportions of boys, minorities, or low-income students.

The remainder of this study proceeds as follows. Section 2 gives an concise overview of the literature on pupil-to-teacher ratio policies. Section 3 presents the STAR experiment, the data set, and some descriptive statistics. Section 4 introduces our empirical strategy for identifying the effect of CSR and TA on student test scores in a distributional framework. Section 5 presents

and discusses the results and section 6 performs the attrition analysis and a series of robustness checks. Finally, section 7 concludes.

2 Background

Researchers have invested a lot of effort in studying pupil-to-teacher ratio, mainly for two reasons. First, pupil-to-teacher ratio is readily measurable, and both teachers and parents usually perceive class size to be negatively correlated with student achievement (Angrist and Lavy, 1999). Second, traditional economic theory suggests that smaller classes have a positive impact on achievement (Lazear, 2001; Todd and Wolpin, 2003). In every education production function, class size is always listed along with other relevant school resources such as teacher qualification and school funding (Hanushek, 2002; Woessmann and West, 2006).

Although theoretical research predicts that a lower pupil-to-teacher ratio would have positive effects on student achievement, the empirical evidence is mixed and exclusively focused on class size. On the one hand, studies relying on experimental data consistently find positive causal effects of CSR on student test scores (Fletcher, 2009; Krueger, 1999). On the other hand, non-experimental studies present a less optimistic view of class size effects. Indeed, most non-experimental studies report either relatively small CSR effects (Angrist and Lavy, 1999; Woessmann and West, 2006) or no effects at all (Hoxby, 2000; Woessmann, 2005). Lazear (2001) explains this lack of empirical support in the non-experimental literature by arguing that teachers (and schools) adjust their behavior to smaller classes, therefore bringing no significant effects on test scores. This hypothesis is supported by Fredriksson et al. (2015), who show that public and private investments in pupils are substitutes in CSR interventions.

While many authors have been studying CSRs, we know very little about the impact of TA. The few papers analyzing TA only looked at its mean impact, finding that TA has, on average, no significant effect on pupils' test scores (Finn and Achilles, 1990; Gerber et al., 2001; Krueger, 1999). These results have led many academics to not consider the presence of TA when performing inference on education data (e.g., Chetty et al., 2011; Jackson and Page, 2013).

Similarly, the effects of pupil-to-teacher ratio policies in a distributional framework are also

overlooked by most of the literature. While we might expect that CSRs could have different impacts on children with different cognitive abilities, the empirical literature left this question unanswered. Only few recent studies (Jackson and Page, 2013; Konstantopoulos, 2008; Müller, 2013) have attempted to tackle this heterogeneity issue, but the evidence remains mixed. No study analyzes the distributional impact of TA, which is potentially beneficial for low-achieving pupils (Gerber et al., 2001).

To better understand the effects of CSR on the achievement gap, Konstantopoulos (2008) examines the variance of test scores within smaller and larger classes. He finds that CSR increases not only level of achievement but also variance in achievement. In addition, he finds no significant evidence that a CSR would reduce the achievement gap between low and high achievers. Jackson and Page (2013) employ a different econometric approach and find that the largest test score gains are at the top of the achievement distribution. They estimate the class-size effect as a difference between treatment and control group at each percentile of the observed achievement distribution. However, doing so provides only a measure of within-group variation, while researchers and policymakers are really interested in the total effect (Borah and Basu, 2013; Maclean et al., 2014). In a related study, Müller (2013) analyses teacher experience as a moderating factor for the effect of CSR on student achievement using Project STAR data. His results reveal that class size effects exist only for senior teachers, and that although such effect is present at all deciles of the achievement distribution, it is weaker at lower deciles.

Overall, previous research suggests that there are differences in the effects of CSR on kids with different achievement levels, but the heterogeneous effects have not systematically studied before. Moreover, there is no paper focusing on the other policy to reduce pupil-to-teacher ratio, i.e., TA policies.

3 Data

3.1 Project STAR

We use data from the Project STAR experiment, a large-scale, randomized class-size experiment that took place between 1985 and 1989 and involved 11,600 students from kindergarten

through third grade. Project STAR was commissioned by the Tennessee state legislature and implemented by a consortium of Tennessee universities and the Tennessee State Department of Education. The total cost of the experiment, including the cost of hiring new teachers and classroom aides, was approximately USD 12 million.

The 42 participating districts had to agree to enter the experiment for four years and to allow site visitations for verification of class sizes, interviewing, and data collection. Similarly, all the 79 participating schools had to allow the random assignment of pupils and teachers to class types from kindergarten through third grade. The experiment randomly assigned kindergarten pupils to small classes (target enrollment between 13 and 17 students), regular classes (target enrollment between 22 and 26 students), or regular classes with a full-time teacher's aide. These teachers' aides did not have to possess any specific educational background and they did not receive any particular training in their duties, their job was helping teachers prepare materials for class and tutoring individual students with learning difficulties.

Although Project STAR covered only one state and one cohort, the experiment included a heterogeneous set of schools from across Tennessee, including large and small, urban and rural, and wealthy and poor districts. Consequently, the schools included in the STAR data represent most of the educational conditions that exist in the United States (Finn and Achilles, 1990; Krueger and Whitmore, 2001). Additionally, the data includes detailed information on pupils, teachers, and schools, for which we can control for. For the present study, we focus on the kindergarten cohort. The reason for this choice is dictated by the experimental design (Krueger, 1999). After the kindergarten year of Project STAR, an initial assessment of the TA intervention revealed no significant average effect on test scores. The experimental committee thus decided to re-randomize pupils who were originally assigned to one of the treatments in a regular-size class (TA and control group). This re-randomization makes it impossible for researchers to know which treatment the pupils in regular classes experienced after the kindergarten year.

The measure of achievement is the Stanford Achievement Test (SAT–9), which all Project STAR children took at the end of each grade. The SAT–9 is a standardized, multiple choice test that includes math, reading, and word identification as subject areas. Because there are no standard units for the test results, we follow Krueger (1999) and scale the test scores into

percentile ranks. Specifically, at each grade level and for each treatment condition, we assign percentile scores based on students' raw test scores, ranging from 0 (lowest score) to 100 (highest score). For each subject test (math, reading, word), we generate a separate percentile distribution and, to summarize overall achievement, we calculate the average of the three SAT–9 percentile rankings for each student. As a robustness check, we also perform the analysis using standardized scores (z-scores) instead of percentile ranks as in Jackson and Page (2013), with virtually identical results.

3.2 Descriptive Statistics

The main advantage of a randomized experiment is that it provides a solution to the problem of causal inference. In principle, if randomization is done appropriately, the mean outcomes of the treatment and control groups can stand as counter-factual for one another, making inference about the effects of the treatment relative to the control transparent. If the treatment conditions were truly randomly assigned to pupils in each school, then individuals in the treatment and control groups should have—in expectation—the same pre-intervention characteristics. One way to test this condition is to check whether assignment to a treatment condition is predictive of pupil and teacher characteristics. If pupils and teachers were properly randomized within schools, we should observe no statistically significant relationship between each treatment condition and all pre-intervention covariates.

The randomization check for Project STAR has been performed several times in the literature (e.g., Krueger, 2003; Krueger and Whitmore, 2001). For this reason, we briefly repeat the analysis for kindergarten, focusing on both treatment conditions (CSR and TA) instead of comparing the CSR with the rest—as the literature usually does. To check for random assignment, which was performed at entry into Project STAR, we regress student and teacher characteristics on the treatment indicators. We also have to include school fixed effects, because random assignment was only valid within schools. We finally perform an F-test of the hypothesis that the class-type dummies had no joint effect.

Table 1 presents descriptive statistics and covariate balance. The table shows that children assigned to the CSR treatment were actually put in smaller classes. The typical small class has

seven pupils less than a regular class. In terms of outcome, pupils assigned to smaller classes score significantly higher than pupils in the control group. The difference in test scores is less pronounced for pupils assigned to the TA treatment, which is in line with the literature on average impacts of TAs (Krueger, 1999).

[Table 1 here]

For pupils and teachers characteristics, all p-values of the treatment dummies exceed statistical significance, suggesting that there is no significant difference in pupils and teachers characteristics across treatments. We thus find no evidence that initial assignment to class types was highly correlated with either pupil or teacher characteristics. Therefore, we can be reasonably confident that, within schools, both pupils and teachers were randomly assigned to the treatment conditions in kindergarten. When we estimate the treatment effects, we nevertheless condition not only on school fixed effects but also on all observable student and teacher characteristics to increase the precision of the point estimates.

4 Methods

In this section, we present our empirical strategy for identifying the effect of CSR and TA on student test scores in a distributional framework. Our core model is a standard equation of student achievement of the following form:

$$Y_{ics} = \beta_0 + \beta_1 \cdot CSR_{cs} + \beta_2 \cdot TA_{cs} + X'_{ics}\beta_3 + Z'_{cs}\beta_4 + \alpha_s + \varepsilon_{ics} \tag{1}$$

where Y_{ics} is the average percentile score on the SAT–9 test of student i in class c at school s, CSR_{cs} is a dummy variable indicating whether the student was assigned to a small class, TA_{cs} is a dummy variable indicating whether the student was assigned to a regular-size class with an aide, X_{ics} is a vector of observed student covariates (gender, age, ethnicity, and free-lunch eligibility), and Z_{cs} is a vector of observed teacher covariates (gender, ethnicity, years of experience, and qualifications). Given that the randomization was done at entry within schools, we also include school fixed effects (α_s).

Because student unobserved ability also contributes to the achievement level (Ashenfelter and Rouse, 1998), in any non-experimental application researchers generally lack data on some relevant characteristics, whether observed or not. These omitted variables will then appear in the error term ε_{ics}, and if the omitted variables are correlated with the included variables, then the estimated parameters will be biased. However, if pupil-to-teacher ratio is determined by random assignment, it will be independent of the omitted variables. For this reason, the coefficients β_1 and β_2 in equation 1 represent the causal effect of being assigned to a small-size or regular-size with aide class on the percentile score of the SAT test. In the literature, such effects are referred as "reduced-form" or "intention-to-treat" effect. Given that the dependent variable is expressed in percentile ranks, we interpret the effect size of β_1 and β_2 as a percentage point change in the distribution of achievement.

To estimate the effect of CSR and TA over the entire distribution of test scores, we use the unconditional quantile regression (UQR) approach of Firpo et al. (2009). Firpo, Fortin, and Lemieux's estimator allows for a direct measure of how a marginal change in the level of one variable (in our case, the treatment dummies) affects the distribution of achievement in the population, keeping the distribution of other characteristics equal. UQR differs from the commonly used conditional quantile regression (CQR) by Koenker and Bassett (1978). CQR estimates treatment effects conditional on the mean value of included covariates, and the interpretation of such treatment effects change when different sets of covariates are entered into the regression equation. Consequently, the interpretation of effects is limited when the effects for different conditional quantiles vary. In such cases, the estimated effects do not translate to relevant policy questions that are linked to the covariates of interest (Borah and Basu, 2013; Maclean et al., 2014).

A simple example adapted from Frölich and Melly (2010) illustrates this advantage. As in our Project STAR setting, assume that a treatment has been completely randomized and is thus independent of both potential outcomes and other covariates. In such a situation, a comparison of the distribution of the outcome in the treated and non-treated populations has a causal interpretation. Also suppose that, either for efficiency or because of block randomization (as in our case), we may wish to include covariates or fixed effects in the estimation. If we are

interested in mean effects, it is well known that including covariates that are independent of the treatment in a linear regression leaves the estimated treatment effect unchanged. This property is lost for quantile treatment effects, because including covariates that are independent of the treatment changes the limit of the estimated *conditional* quantile treatment effect. However, including those covariates does not change the *unconditional* treatment effect, as long as the exogeneity assumptions of the model are satisfied (which are indeed in our randomized setting).

Additionally, because a conditional quantile is the relative position of an individual among a (virtual) population of individuals that share precisely the same observed characteristics, CQR yields only the within-group effect, whereas UQR estimates the total effect, i.e., the sum of the between-group and within-group effects (Fournier and Koske, 2013). This means that UQR allows us to compare estimated effects at different quantiles with each other, whereas CQR does not allow for such comparison.

In practice, UQR consists of running a regression of a relatively simple transformation—the re-centered influence function—of the outcome variable on the explanatory variables. Because of its policy relevant interpretation and its computational attractiveness, UQR has been used in several studies on quantile effects (Firpo et al., 2009; Maclean et al., 2014; Stueber and Beissinger, 2012) and decomposition analyses (Heywood and Parent, 2012; Sakellariou, 2012; Tang and Long, 2013).

An influence function is an analytic tool assessing the effect of removing or adding an observation on the value of a certain statistic $v(F)$, without having to recalculate that statistic. The influence function is defined as follows:

$$IF[y, v(F)] = \lim_{h \to 0} \frac{v[(1-h) \cdot F + h \cdot \delta_y] - v(F)}{h}, 0 \leq h \leq 1 \tag{2}$$

where F represent the cumulative distribution function for Y, and δ_y is a distribution that puts mass at the value y. We obtain the re-centered influence function (RIF) by adding the statistics of interest to its influence function:

$$RIF(y, v) = v(F) + IF(y, v) \tag{3}$$

If the statistic of interest is a specific quantile τ of the distribution of the outcome of interest, we have:

$$IF[y, v(F)] = (\tau - I[Y \leq q_\tau])/f_Y(q_\tau) \tag{4}$$

where q_τ is the τ^{th} quantile of the unconditional distribution of Y, $f_Y(q_\tau)$ is the probability density function of Y evaluated at q_τ, and $I[Y \leq q_\tau]$ indicates whether an outcome value is less than the specified quantile q_τ. In the case of quantiles, the re-centered influence function is then:

$$RIF(y, q_\tau) = q_\tau + IF(y, q_\tau) \tag{5}$$

Firpo et al. (2009) show that if the conditional expectation of $RIF(y, q_\tau)$ is modeled as a function of explanatory variables, a RIF regression can be viewed as an UQR estimator.[1]

The implementation of the UQR method is a two-step procedure. For a specific quantile τ, we first have to estimate the RIF of the τ^{th} quantile of Y following (4) and (5). We calculate q_τ using the sample estimate of the unconditional τ^{th} quantile of Y. Similarly, we estimate the density $f_Y(q_\tau)$ at q_τ using kernel methods. The second step is to run an ordinary least squares (OLS) regression of the $RIF(y, q_\tau)$ on the treatment variables and other observed covariates. In this two-step procedure, the unconditional quantile partial effects are simply the estimated coefficients.[2]

A limitation of UQR is that no valid method to cluster standard errors currently exists. While we are aware of the importance of clustering standard errors at the treatment level (the class in our study), Chetty et al. (2011) provide evidence that, in the case of Project STAR, avoiding clustering does not overstate precision. We thus rely on robust standard errors for UQR and standard errors clustered at the class level in the OLS estimations.

[1] This is because, as $\mathrm{E}_X\mathrm{E}[RIF(Y,\tau)|x] = q_\tau$ by the definition of RIF, Firpo et al. (2009) demonstrate that $\mathrm{E}_X[dm_\tau(x)/dX]$ is the marginal effect of a covariate on the τ^{th} unconditional quantile of Y, ceteris paribus.

[2] To compute the unconditional quantile treatment effects, we use the Stata routine `rifreg`, available at http://faculty.arts.ubc.ca/nfortin/datahead.html

5 Results

This section presents our results in two parts. The first part presents the quantile treatment effect of CSR and TA on student test scores using the UQR approach. The second part shows sub-sample analysis, to investigate heterogeneous effects across observable characteristics.

5.1 Distributional Effects of CSR and TA

Table 2 presents the unconditional treatment effects on kindergarten test scores for both the CSR and TA intervention. We gradually include pupil and teacher covariates in our models, and while doing so does not largely alter the effect sizes, pupil and teacher characteristics are jointly significant. Thus our preferred estimates are those shown in in column 3.

The coefficients of CSR reveal that being assigned to a small class has positive effects on test scores throughout the entire achievement distribution ($p < 0.00$). However, there is a high degree of heterogeneity in the effects. At the bottom decile of the distribution, students assigned to smaller classes score 2.5 percentage points higher than those assigned to the control group. The CSR effect is larger at the median, reaching 7.7 percentage points. Then the CSR effect declines in the upper part of the achievement distribution, down to about 5 percentage points.

[Table 2 here]

In Figure 1, we compute the unconditional quantile treatment effect of CSR for each percentile of the achievement distribution, along with the respective 90-percent confidence intervals. The figure further illustrates the heterogeneous effect of being assigned to a small class. It shows that the mean effect—the dashed line—is a poor representation of the CSR effect. Mid-achievers (fourth to eighth decile) profit the most from a CSR intervention. Pupils at the bottom and at the top of the achievement distribution experience only a little from being in a smaller class.

[Figure 1 here]

Table 2 also presents unconditional quantile treatment effects for the TA intervention. Consistent with previous studies, we find almost no effect on test scores at the median. However, low achievers actually benefit from TAs. For pupils at the bottom of the achievement distribution,

we estimate a positive and highly significant effect of roughly 2.5 percentage points. This effect is as large as that of being in a small class. To emphasize the relevance of the TA effect, Figure 2 presents unconditional quantile treatment effect of TA at each percentile of the achievement distribution. Being assigned to the TA condition clearly has a positive effect on test scores for low-achieving students. The effect is highly significant for the first two deciles of the achievement distribution and ranges between 2-3 percentage points. For the rest of the distribution, the TA condition has no impact on test scores.

[Figure 2 here]

From a policy perspective, the TA result is important for at least two reasons. First, TA is a less expensive alternative to CSR, not only because a TA has both lower training costs and wage but also because no additional classrooms and materials are needed to implement a TA policy. Second, in terms of general equilibrium effects, finding teacher aides for a TA policy is less demanding than finding additional teachers to sustain a CSR policy. This responds to the common critique to CSR interventions on where to find the needed teachers. In the Project STAR case, for example, TAs did not have to possess any specific background and they did not receive any particular training to perform their duties. In spite of this, we still observe a significant positive effect on test scores for low-achieving children.

Given the heterogeneity we observe in the data for both the CSR and the TA intervention, we need to explain the potential mechanism behind our results. Understanding why high achievers benefit less from CSRs compared to the mid-achievers is somehow intuitive. Being a high achiever usually correlates with both higher motivation and higher socio-economic status (Heckman and Masterov, 2007). Children with such characteristics would probably perform well regardless of the treatment they receive, or at least they would benefit less compared to those students who lack of either resources or motivation.

Understanding why low achievers benefit less from a CSR intervention, instead, might be less intuitive. Theory would suggest that, in smaller classes, teachers are more able to identify low achievers and thus more likely to provide instruction designed to benefit these students (Konstantopoulos, 2008). However, such practice is difficult to implement when there is only

one teacher in the classroom, because he or she would need to focus only on a group of children leaving the majority of the class without supervision. By contrast, when the adults in the classroom are two as in the TA treatment, teachers are not only likely to identify low achievers but also able to provide targeted instruction to benefit such students through the TA. This might explain the positive impact of being in a class with a TA for those pupils who need more help or support.

5.2 Sub-Sample Analysis

To further understand the heterogeneous effects of CSR and TA, we perform our main analysis for groups of pupils with specific observable characteristics. We conduct a separate analysis according to gender, ethnicity, and socio-economic status (SES). As it is common in the literature, we proxy lower SES with free-lunch eligibility. We perform this analysis because both economists and psychologists suggest that these sub-groups tend to be more disruptive (Bertrand and Pan, 2013; Kristoffersen et al., 2015) and more likely to lose focus during instructional time (Feingold, 1994).

Figure 3 provides graphical representation of the effect of CSR and TA for boys, black children, and free-lunch eligible children.[3] The respective regression outputs are presented in Appendix Table A.1. The CSR intervention has larger effects for boys, black children, and free-lunch eligible children. This is consistent with previous studies that focus on average effects (Dee, 2007), but our analysis also underlines a high level of heterogeneity in the effect of CSR. The inverted u-shaped pattern over the percentiles of the achievement distribution persists, but now the effect drops only at the top decile (instead of the top quintile as for the full sample).

[Figure 3 here]

Similarly, the TA intervention is very beneficial for boys, black children, and free-lunch eligible children. The treatment TA has a strong positive impact on low-achievers' test scores and this effect is significant almost for half of their achievement distribution. Comparing the

[3]We also performed the analysis for white children, girls, and higher SES children. We do not report the results here, because they are very similar to the estimates using the full sample (slightly smaller effect sizes). However, results are available upon request.

effect of TA for the sub-samples to the one of the entire sample, we observe two phenomena. First, the effect is larger in terms of magnitude; and, second, the effect is significant up to the fourth decile (whereas for the full sample it is significant only for the bottom two deciles).

6 Attrition Analysis and Robustness Checks

Although over the entire Project STAR attrition was high (Hanushek, 1999), in kindergarten attrition was less severe. In our data, we have missing kindergarten test scores for 488 children, corresponding to 7.7 percent of the kindergarten cohort. If outcome data are missing for some pupils, we might be concerned that the potential outcomes for those who are observed in the treatment groups differ from the potential outcomes for those observed in the control group. Even if attrition is not different across treatment groups, departures could yield analytic samples that vary significantly from the original sample, limiting external validity of estimated causal effects.

While there is no outcome data on students who left Project STAR before the test scores were collected, we can look for evidence of non-random attrition by examining differences in attrition rates across treatments and in observable characteristics across treatments. To do so, we regress an indicator of whether a student left Project STAR during kindergarten on the treatment dummies and an interaction between the treatment dummies and the pre-intervention student and teacher characteristics. We define attrition as a binary variable that equals one if a pupil left the study during kindergarten and zero otherwise.

Table 3 presents our attrition analysis. The table is divided in two parts as follows: panel A shows attrition rates for each treatment condition and tests whether attrition rates are significantly different across treatment groups; panel B shows whether any student or teacher characteristic moderates the attrition rates. In general, we find that attrition rates tend to be lower for the treatment groups than the control group, especially for the CSR treatment. In terms of observable characteristics, black children, older children, and pupils with a black teacher are more likely to leave the sample. However, most of these differences are not statistically significant. The only characteristic that is marginally significant ($p = 0.08$) is whether

pupils had a black teacher in kindergarten. It appears that attrition rates in the TA treatment were disproportionately high for classes with a black teacher.

[Table 3 here]

Overall, the attrition patterns just described are relatively low and not highly significant. Nonetheless, missing 7.7 percent of the test scores may cause some problems to the interpretation of our findings. In our first robustness check we thus adjust for attrition by imputing test scores for children who left the sample. We adopt a worst-case scenario and predict the scores of pupils who left the control group as if they received the treatment CSR. Conversely, we predict for pupils who left one of the treatment groups as if they received no treatment. This imputation technique should lead to an increase in average achievement for the control group and, at the same time, a decrease in average achievement for the treatment groups.

Column 1 (no student and teacher controls) and column 2 (with student and teacher controls) of Appendix Table B.1 present the quantile treatment effects of CSR and TA on the imputed scores. Overall, using the imputed scores reduces the estimated treatment effects but does not entirely wipe out either the CSR effect or the TA effect. This makes us confident that attrition, although present, is not threatening the validity of our main analyses.

Our second robustness check tests the sensitivity of the results to alternative specifications of the dependent variable. We might suspect that the distributional results we obtain depend on how we specify the outcome variable, i.e., in percentile ranks. This should not be the case for unconditional quantile treatment effects, but we present the results also to relate the magnitude of our effects to those reported in the literature. To do so, we standardize the dependent variable in z-scores. One advantage of z-scores is that we can interpret the estimated coefficients as standard deviations units.

Column 3 (no student and teacher controls) and column 4 (with student and teacher controls) of Appendix Table B.1 present the results. As we expected, specifying the dependent variable in z-scores does not affect the findings of our main analysis. Although the median effect of CSR we estimate is similar to the mean effect reported by the literature (Krueger and Whitmore, 2001), our distributional effects for CSR are slightly larger than those reported in

previous studies (Jackson and Page, 2013). This is because we estimate, for each quantile, the total effect of the treatments (between and within) instead of the effect within groups with exactly the same observable characteristics. The total effect is what policymakers are interested in when designing educational reforms, because it is not based exclusively upon a homogeneous group in the population (Maclean et al., 2014).

7 Conclusions

The vast majority of the literature on pupil-to-teacher ratio focuses on CSR effects for the average student. This is undoubtedly relevant for both researchers and policymakers, but it might also present an incomplete picture of the heterogeneous effect of reforms that aim to reduce pupil-to-teacher ratio. In this paper, we provide a distributional analysis of both CSR and TA policies, showing that these policies not only have highly heterogeneous effects but also that they affect different parts of the achievement distribution.

Our results contribute to the literature on pupil-to-teacher ratio policies in at least three ways. First, we show that, given the large amount of heterogeneity in the treatment effects, mean regression provides only a poor description of the underlying relationship between CSR and achievement. Similarly, not even standard sub-sample analysis is a sufficient tool for studying heterogeneity and heterogeneity patterns over the achievement distribution, because—as our results show—no standard approach could reveal the distributional effect we present here.

Second, we find that mid-achieving students gain the most from being assigned to smaller classes, whereas students at the bottom and top of the achievement distribution experience only minimal gains. This result differs from what those few studies that investigate the distributional effects of class size suggest due to a new, improved econometric approach (i.e., unconditional quantile regression), and our findings are robust across alternative specifications and estimation techniques.

Third, we report positive and significant effects of TA for the low-achieving pupils. Not only is the effect significant for the first two deciles of the achievement distribution, but it is even larger for boys and disadvantaged children. Interestingly, the effect size of TA is as large

as that of CSR for the bottom third of the achievement distribution. In terms of equity, while the net effect of CSR on the achievement gap is not clear, our estimates show that adding a TA would be an effective policy for reducing the achievement gap, especially for classes with large percentages of boys, black students, or low-income students.

This paper shows that typical estimates of the average gain from CSR and TA provide an incomplete characterization of their real impact on the achievement distribution, thus constituting a weak guide for public educational policy. While CSRs have the largest impact on mid-achievers, having an in-class TA constitutes an effective measure for raising the test scores of low achievers. Similarly, while a TA appears to have no impact on test scores at the mean, having a TA has a significant impact on low-achieving pupils. We conclude that policymakers, when designing educational reforms, need to think more carefully about their distributional goals and how these goals are affected by an intervention, rather than its impact on the average pupil.

References

Angrist, J. D. and Lavy, V. (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114(2):533–575.

Ashenfelter, O. and Rouse, C. (1998). Income, schooling, and ability: Evidence from a new sample of identical twins. *The Quarterly Journal of Economics*, 113(1):253–284.

Bertrand, M. and Pan, J. (2013). The trouble with boys: social influences and the gender gap in disruptive behavior. *American Economic Journal: Applied Economics*, 5(1):32–64.

Borah, B. J. and Basu, A. (2013). Highlighting differences between conditional and unconditional quantile regression approaches through an application to assess medication adherence. *Health Economics*, 22(9):1052–1070.

Bulman, G. and Fairlie, R. W. (2016). Technology and education: Computers, software, and the internet. NBER Working Papers 22237, National Bureau of Economic Research, Inc.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly Journal of Economics*, 126(4):1593–1660.

Chingos, M. M. (2013). Class size and student outcomes: Research and policy implications. *Journal of Policy Analysis and Management*, 32(2):411–438.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources*, 42(3):528–554.

Dynarski, S., Hyman, J., and Schanzenbach, D. W. (2013). Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion. *Journal of Policy Analysis and Management*, 32(4):692–717.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, 116(3):429–456.

Finn, J. D. and Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal*, 27(3):557–577.

Firpo, S., Fortin, N. M., and Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica*, 77(3):953–973.

Fletcher, J. M. (2009). Is identification with school the key component in the black box of education outcomes? Evidence from a randomized experiment. *Economics of Education Review*, 28(6):662–671.

Fournier, J.-M. and Koske, I. (2013). Public employment and earnings inequality: An analysis based on conditional and unconditional quantile regressions. *Economics Letters*, 121(2):263–266.

Fredriksson, P., Oosterbeek, H., and ckert, B. (2015). Parental responses to public investments in children: evidence from a maximum class size rule. Working Paper Series 2015:27, IFAU - Institute for Evaluation of Labour Market and Education Policy.

Frölich, M. and Melly, B. (2010). Estimation of quantile treatment effects with stata. *Stata Journal*, 10(3):423–457.

Gerber, S. B., Finn, J. D., Achilles, C. M., and Boyd-Zaharias, J. (2001). Teacher aides and students academic achievement. *Educational Evaluation and Policy Analysis*, 23(2):123–143.

Hanushek, E. A. (1999). Some findings from an independent investigation of the tennessee star experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis*, 21(2):143–163.

Hanushek, E. A. (2002). Publicly provided education. In Auerbach, A. J. and Feldstein, M., editors, *Handbook of Public Economics*, volume 4 of *Handbook of Public Economics*, chapter 30, pages 2045–2141. Elsevier.

Heckman, J. J. and Masterov, D. V. (2007). The productivity argument for investing in young children. *Review of Agricultural Economics*, 29(3):446–493.

Heywood, J. S. and Parent, D. (2012). Performance pay and the white-black wage gap. *Journal of Labor Economics*, 30(2):249–290.

Hough, J. (1993). Educational finance issues in North America. *Education Economics*, 1(1):35–42.

Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics*, 115(4):1239–1285.

Jackson, E. and Page, M. E. (2013). Estimating the distributional effects of education reforms: A look at project star. *Economics of Education Review*, 32(C):92–103.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Konstantopoulos, S. (2008). Do small classes reduce the achievement gap between low and high achievers? evidence from project star. *The Elementary School Journal*, 108(4):275–291.

Kristoffersen, J. H., Obel, C., and Smith, N. (2015). Gender differences in behavioral problems and school outcomes. *Journal of Economic Behavior & Organization*, 115:75–93.

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, 114(2):497–532.

Krueger, A. B. (2003). Economic considerations and class size. *The Economic Journal*, 113(485):F34–F63.

Krueger, A. B. and Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project star. *The Economic Journal*, 111(468):1–28.

Lazear, E. P. (2001). Educational production. *The Quarterly Journal of Economics*, 116(3):777–803.

Maclean, J. C., Webber, D. A., and Marti, J. (2014). An application of unconditional quantile regression to cigarette taxes. *Journal of Policy Analysis and Management*, 33(1):188–210.

Müller, S. (2013). Teacher experience and the class size effect—Experimental evidence. *Journal of Public Economics*, 98(C):44–52.

Sakellariou, C. (2012). Unconditional quantile regressions, wage growth and inequality in the philippines, 2001–2006: The contribution of covariates. *Applied Economics*, 44(29):3815–3830.

Stueber, H. and Beissinger, T. (2012). Does downward nominal wage rigidity dampen wage increases? *European Economic Review*, 56(4):870–887.

Tang, Y. and Long, W. (2013). Gender earnings disparity and discrimination in urban china: Unconditional quantile regression. *African Journal of Science, Technology, Innovation and Development*, 5(3):202–212.

Todd, P. E. and Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113(485):F3–F33.

West, M. R. and Woessmann, L. (2006). Which school systems sort weaker students into smaller classes? international evidence. *European Journal of Political Economy*, 22(4):944–968.

Woessmann, L. (2005). Educational production in Europe. *Economic Policy*, 20(43):445–504.

Woessmann, L. and West, M. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review*, 50(3):695–736.

Tables and Figures

Table 1: Descriptive statistics and covariate balance

	CSR (1)	TA (2)	Control (3)	N (4)	Joint p-value (5)
Test score percentile	52.15	47.41	47.49	5,837	0.00***
Class size	15.12	22.78	22.38	6,253	0.00***
Girl	0.49	0.48	0.49	6,253	0.85
Black	0.31	0.34	0.33	6,253	0.35
Age (in 1985)	4.65	4.65	4.64	6,253	0.30
Free-lunch eligible	0.47	0.50	0.48	6,253	0.41
Black teacher	0.14	0.15	0.20	6,253	0.34
Teacher with master	0.31	0.36	0.36	6,253	0.57
Teacher experience	9.90	10.70	10.03	6,253	0.30

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Ordinary least squares models with robust standard errors clustered at the class level. Project STAR data for the kindergarten cohort.

Table 2: CSR and TA effects in kindergarten

	Percentile rank score		
	(1)	(2)	(3)
Quantile .10			
CSR treatment	2.423***	2.349***	2.475***
	(0.846)	(0.838)	(0.841)
TA treatment	2.265***	2.346***	2.461***
	(0.814)	(0.805)	(0.811)
Quantile .25			
CSR treatment	4.080***	3.896***	3.990***
	(1.143)	(1.118)	(1.122)
TA treatment	1.325	1.485	1.352
	(1.099)	(1.075)	(1.082)
Quantile .50			
CSR treatment	7.824***	7.590***	7.749***
	(1.349)	(1.306)	(1.312)
TA treatment	–1.024	–0.799	–0.813
	(1.297)	(1.255)	(1.264)
Quantile .75			
CSR treatment	7.767***	7.528***	7.516***
	(1.201)	(1.173)	(1.178)
TA treatment	–0.061	0.092	–0.213
	(1.155)	(1.128)	(1.136)
Quantile .90			
CSR treatment	5.149***	5.015***	5.098***
	(1.069)	(1.055)	(1.061)
TA treatment	–1.035	–0.935	–1.015
	(1.028)	(1.014)	(1.022)
School fixed effects	✓	✓	✓
Student covariates		✓	✓
Teacher covariates			✓
N	5,837	5,837	5,837

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Robust standard errors are in parentheses below the coefficients. Project STAR data for the kindergarten cohort.

Figure 1: Distributional effect of CSR on test scores

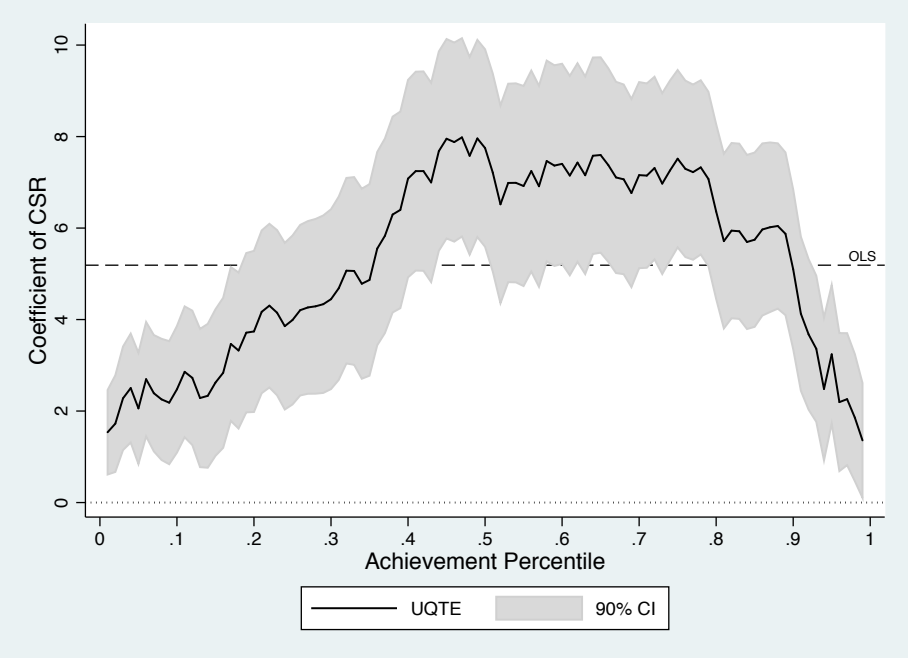

Figure 2: Distributional effect of TA on test scores

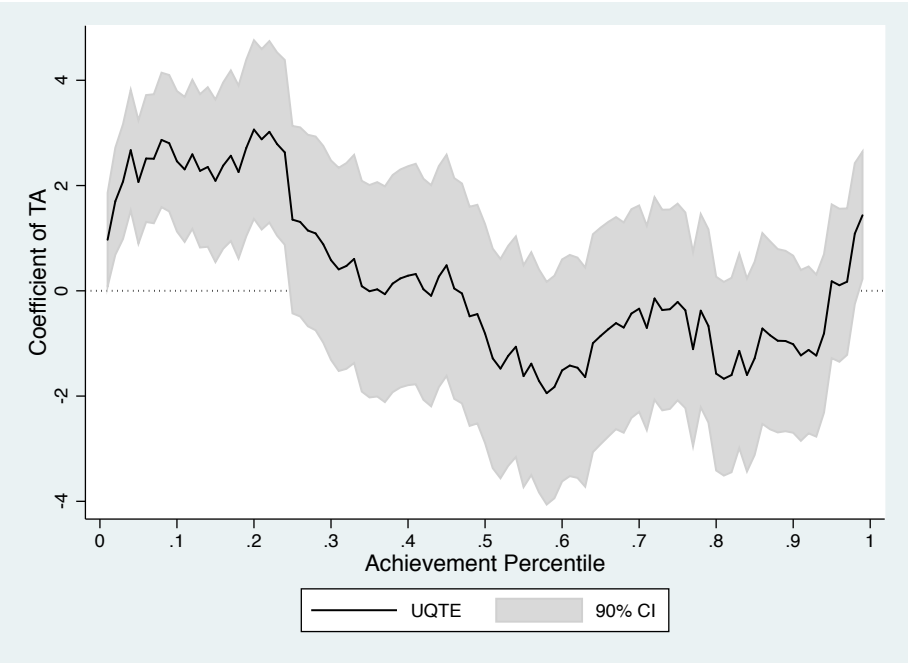

Figure 3: Distributional effect of CSR and TA, by sub-group

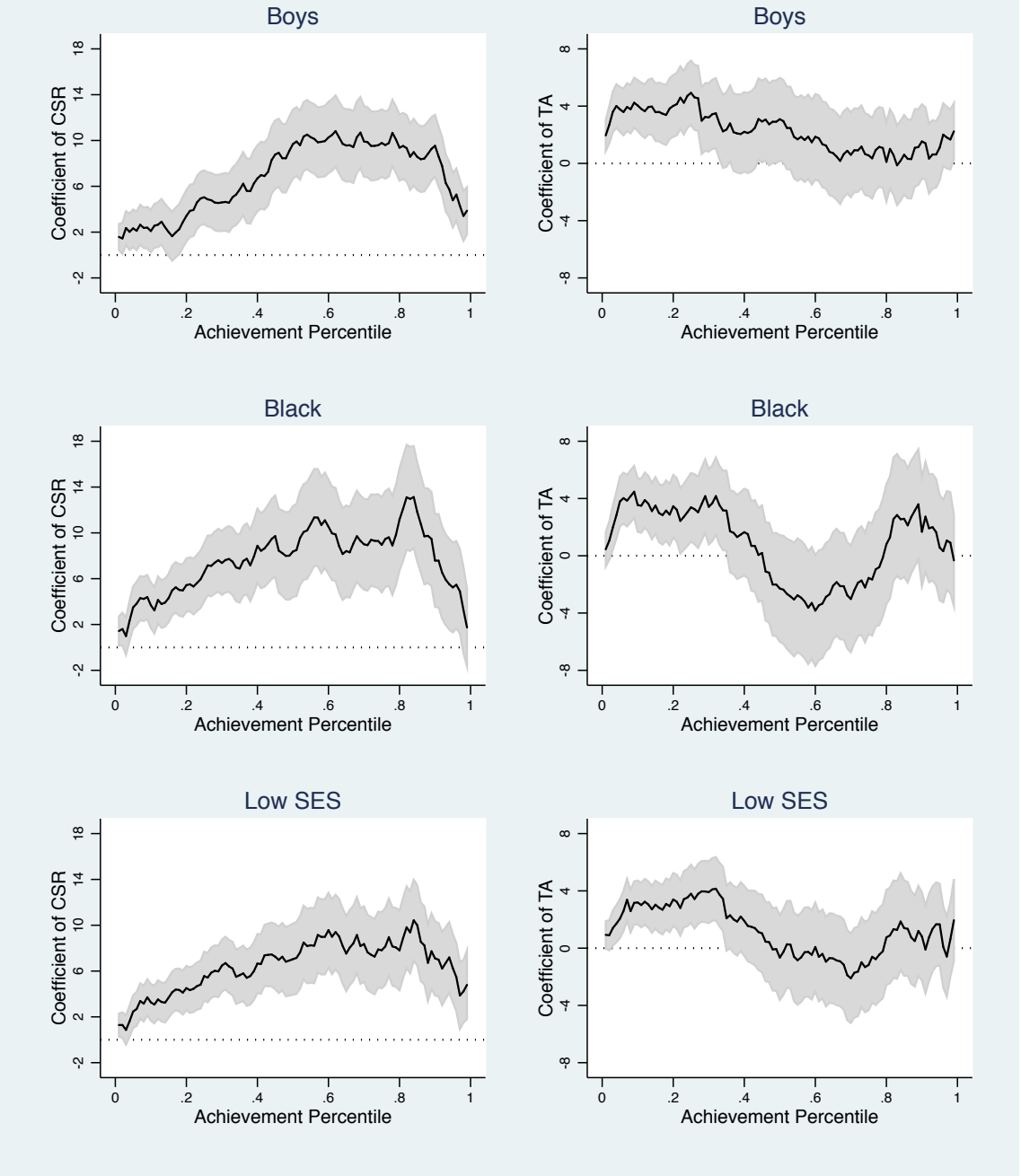

Table 3: Attrition analysis

	CSR (1)	TA (2)	Control (3)	N (4)	*p*-value (3) – (1)	*p*-value (3) – (2)
A. *Attrition*	0.067	0.064	0.070	6,325	0.88	0.33
B. *Non-random attrition*						
Girl	0.065	0.058	0.069	6,325	0.95	0.69
Black	0.067	0.072	0.082	6,322	0.42	0.85
Age > 5	0.081	0.060	0.084	6,317	0.96	0.43
Free-lunch eligible	0.064	0.072	0.072	6,300	0.64	0.58
Black teacher	0.076	0.115	0.095	6,282	0.98	0.08*
Teacher with master	0.067	0.057	0.051	6,304	0.28	0.41
Teacher experience > 10	0.075	0.057	0.068	6,304	0.17	0.25

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Ordinary least squares models with robust robust standard errors clustered at the class level. All models include school fixed effects. Project STAR data for the kindergarten cohort.

APPENDIX

A Sub-Sample Analysis

Table A.1: CSR and TA effects on kindergarten test scores, sub-samples

	Percentile rank score		
	Boy	Black	Free-lunch
	(1)	(2)	(3)
Quantile .10			
CSR treatment	2.093*	3.679***	3.267***
	(1.128)	(1.212)	(0.978)
TA treatment	4.045***	3.542***	3.191***
	(1.083)	(1.137)	(0.931)
Quantile .25			
CSR treatment	5.041***	6.551***	5.562***
	(1.425)	(1.553)	(1.297)
TA treatment	4.928***	3.386**	3.797***
	(1.368)	(1.456)	(1.235)
Quantile .50			
CSR treatment	9.692***	8.369***	7.037***
	(1.829)	(2.413)	(1.890)
TA treatment	3.087*	–2.288	–0.656
	(1.755)	(2.263)	(1.800)
Quantile .75			
CSR treatment	9.782***	8.955***	7.817***
	(1.792)	(2.527)	(2.020)
TA treatment	0.560	–1.551	–1.178
	(1.720)	(2.370)	(1.924)
Quantile .90			
CSR treatment	9.539***	7.578***	7.107***
	(1.643)	(2.459)	(1.969)
TA treatment	1.533	–1.678	0.818
	(1.577)	(2.307)	(1.875)
School fixed effects	✓	✓	✓
Student covariates	✓	✓	✓
Teacher covariates	✓	✓	✓
N	2,991	1,897	2,823

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Robust standard errors are in parentheses below the coefficients. Project STAR data for the kindergarten cohort.

B Robustness Checks

Table B.1: CSR and TA effect on imputed scores and Z-scores

	Imputed test scores		Z-scores	
	(1)	(2)	(3)	(4)
Quantile .10				
CSR treatment	2.606***	2.604***	0.108***	0.110***
	(0.869)	(0.861)	(0.037)	(0.037)
TA treatment	2.179***	2.337***	0.119***	0.125***
	(0.835)	(0.831)	(0.035)	(0.035)
Quantile .25				
CSR treatment	3.611***	3.439***	0.157***	0.156***
	(1.104)	(1.088)	(0.038)	(0.037)
TA treatment	0.737	0.969	0.063*	0.063*
	(1.062)	(1.050)	(0.036)	(0.036)
Quantile .50				
CSR treatment	5.912***	5.853***	0.254***	0.251***
	(1.180)	(1.151)	(0.042)	(0.041)
TA treatment	–1.589	–1.437	–0.018	–0.014
	(1.134)	(1.112)	(0.040)	(0.039)
Quantile .75				
CSR treatment	7.124***	6.917***	0.320***	0.311***
	(1.186)	(1.161)	(0.046)	(0.046)
TA treatment	–0.062	–0.048	–0.019	–0.025
	(1.140)	(1.122)	(0.045)	(0.044)
Quantile .90				
CSR treatment	5.539***	5.581***	0.241***	0.236***
	(1.084)	(1.078)	(0.061)	(0.060)
TA treatment	–0.790	–0.735	–0.087	–0.086
	(1.042)	(1.041)	(0.058)	(0.058)
School fixed effects	✓	✓	✓	✓
Student covariates		✓		✓
Teacher covariates		✓		✓
N	6,325	6,253	5,837	5,837

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Robust standard errors are in parentheses below the coefficients. Project STAR data for the kindergarten cohort.