Journal of Economic Theory and Econometrics, Vol. 29, No. 1, Mar. 2018, 26–47

Bounds on Effects of Class Size Reduction in Project STAR*

Sang Soo Park†

Abstract

Beginning with the seminal work of Manski (1990), there has been a growing literature on estimation and inference on partially identifiable parameters, including the distribution and/or quantile functions of the heterogeneous treatment effect. This article applies and extends the bounding approaches that Williamson and Downs (1990) and Fan and Park (2010, 2012) to partially identify distribution of treatment effects of class size reduction (CSR). Empirical data I used are from the Project STAR. Conducted by Tennessee State Department of Education in 1985-1988, it was a large-scale, randomized experiment designed to investigate the effect of CSR on student performance.

As an extension of the bounding approach that Fan and Park (2010) used, I proposed bounds for the conditional probability distribution function of treatment effects on pre-treatment outcomes. Although it was hard to find definitive properties of the conditional distribution due to the nature of bounding approach, I find the approach is insightful and has a potential.

Keywords Partial identification, bounds on treatment effects, class size reduction, project STAR

JEL Classification C14, C46, I21.

*This research is supported by a Korea University Grant L1713331.

†Department of Economics, Korea University, e-mail : starpac@korea.ac.kr.

Received February 14, 2018, Revised March 16, 2018, Accepted March 21, 2018

1. INTRODUCTION

The inquiry of the effects of treatment such as policy interventions has expanded toward the identification of the distribution of treatment effects beyond focusing certain functionals of it such as the average treatment effect (ATE), see Heckman et al. (1997), Bitler et al. (2006), Djebbari and Smith (2008) among others. As Abbring and Heckman (2007) summarized it, however, it is not possible to exactly identify the distribution of treatment effects unless assumptions about dependent structure between potential outcomes are imposed or agents' decision rules on the participation are assumed, hence additional information/assumptions have been used: see Abbring and Heckman (2007).

A different approach to the problem of identification of the distribution of treatment effects is to utilize the partial identification nature of the problem. Starting from Manski's seminal work in the late 1980's (Manski(1988) for example) the partial identification nature of the identification of treatment effects was recognized and studied by many researchers, see, for example, Firpo and Ridder (2010), Manski (2003), Fan and Park(2010) and the references therein for a better understanding of the nature. Fan and Park(2010, 2012) dealt with statistical inference on partially identified treatment effects.

This paper is basically an application of bounding approach for identifying distribution of treatment effects that Firpo and Ridder(2010) and Fan and Park(2010) discussed to an experimental project so-called Project STAR. Project STAR, the acronym of The Student/Teacher Achievement Ratio, is a large scale, longitudinal, randomized experiment conducted by the Tennessee State Department of Education during 1985 - 1989 in order to see if the class size reduction improves students' academic achievement.

A large body of existing literature has investigated the effects of CSR and a great portion of them sough to identify the average treatment effects (ATE) of the CSR with a notable exception of Ding and Lehrer (2008). Ding and Lehrer (2008) estimated so-called quantile treatment effects(QTEs) and concluded, "higher ability students gain the most from CSR while many low ability students do not benefit from these reductions in Kindergarten." While their conclusion that the effects of CSR are heterogenous is persuasive, one may bear in mind that the QTEs are not quantiles of treatment effects. In the paper, I will apply and extend the bounding approach that Fan and Park (2010) discussed to think about heterogeneous treatment effects of CSR. The paper is organized as follows: Section 2 briefs about the Project STAR and relevant related researches; Section 3 introduces the bounding approach; empirical work and findings are summarized in Section 4; Section 5 concludes.

2. BRIEF EXPLANATION ON PROJECT STAR

2.1. BRIEF HISTORICAL BACKGROUND

The effect of the CSR has been debated for decades but the conclusion was controversial. Some surveys suggested theoretical channels through which smaller classes helped students attain higher scores. Hallinan and Sorensen (1985) reported that teachers' morale and job satisfaction are higher in small classes and teachers reported that students had better attitudes and motivations. Filby, Cahen, McCutsheon, and Kyle (1980) found teachers were more able to help students when they needed it in smaller classes. In the survey, teachers responded that their work load became lighter, which enabled them to make the classroom climate more positive.

However, empirical researches did not provide conclusive evidences. In their Meta-Analysis with 77 existing studies, Glass, Cahen, and Smith (1978) asserted that they found a trend that the students' achievement decreases as class size increased and they claimed the greatest gains occurred when student-teacher ratio was 1:15 or below. On the contrary, Robinson and Wittebols (1986) found only 35 studies out of the 85 they considered to be relevant reported small classes were better, 18 supported larger classes, and the rest 32 did not support either.

Proir to the launch of Project STAR, Whittington, Bain, and Achilles (1985) investigated the effect of CSR from 1:25 to 1:15 by doing a small scale experimental study with first grade students in the Metro Nashville School District. They reported the students in classes of 15 students performed better than those in classes of 25 in reading and math. On the other hand, Dennis (1986) could not observe any difference between the treated group and control groups in the following year. Bourke (1986) found the class size itself did not affect students' attainment directly. It was, he claimed, teachers' practices that enhanced student achievement. Moreover, teachers do not change their teaching practices when class size is reduced. (Robinson 1990).

There continue to be debates on the effectiveness of CSR. For example, Hanushek (1998) could find "little reason to believe that smaller class sizes systematically yield higher student achievement" , while Krueger (2002) found exactly the opposite and said "when studies are assigned weights in proportion to the 'impact factor'... class size is systematically related to achievement."

Because the CSR was costly, and the results of proceeding researches were not conclusive, the Tennessee State Government decided to conduct a well-designed randomized experiment to investigate whether or not the CSR would be effective before implementing the CSR. In May, 1985, the Tennessee Legisla-

ture passed House Bill (HB) 544, which authorized and funded an experimental study on the effect of CSR, which was Project STAR.

The project was conducted by a consortium of persons from Tennessee State Department of Education, Memphis State University, Tennessee State University, University of Tennessee at Knoxville, Vanderbilt University, representatives from the State Board of Education and the State Superintendents' Association. Only a few months after the pass of the legislation, the consortium was able to set up major parts of the project and to implement it from the fall semester of 1985-1986 schooling year, which continued up until 1988-1989 schooling year.

2.2. DESIGN OF PROJECT STAR

Tennessee had been regulating the student/teacher ratio even before Project STAR.[1] By the time Project STAR started, the ratio could not exceed 1:25. The average of the ratios was 1:22-24. The legislation regulated the ratio in small classes to be between 1:13 and 1:17. So the main question that Project STAR should answer was whether 1:13-17 would be better than 1:22-24 for students' academic achievements.

The consortium decided to divide the class sizes/environment into three categories; small class (teacher:students = 1:13 $\sim$17), regular class (teacher:students = 1:22$\sim$25), and regular class with teacher aide (teacher:students = 1:22$\sim$25). The project schools were chosen out of 180 volunteers from 141 school systems all over the state. Because the consortium designed the project to make the 'within-school' comparison available as well, each school had to have certain number of students so that it had at least one class of each type. After an investigation, the consortium chose 79 schools as the participants for the 1985-1986 schooling year. Initially, a school should plan to remain in the project for the whole years but 1-3 schools left. The initial objective was to have about 100 classes of each type. In the first year, there were 128 small classes, 101 regular classes, and 99 regular classes with teacher aide. Each participating school had to agree to assign students and teachers randomly in three types of classes and not to make any significant changes in their provision of education other than class size. Roughly 6,000 students participated in the project every year.

The kindergarten student academic achievements were measured by Stanford Early School Achievement Test II (SESAT II) in Math, Sounds and Letters, Words and Sentences, and Reading. Higher graders used the Stanford Achievement Tests (SAT), the State of Tennessee's criterian referenced Basic Skills First

[1] This subsection is a summary of the technical report of the STAR project (Word et. al. 1990b), which I will refer as *Technical Report*.

(BSF) tests. In this chapter, I will use the data of kindergarten students on math and reading only.

2.3. EFFECTS OF CSR

After the project was over, Word reported the followings (Word et. al. (1990b), pp.17-19).

1. Small-class advantages are present in all locations and all grades. Students in small classes showed higher performance than those in regular classes or regular classes with teacher aide.

2. Small-class effects diminish after first grade but are significant at the end of third grade.

3. Teacher aides were less effective than CSR in enhancing student performance at each grade level.

4. The effects in Math and reading are similar.

5. Small classes help low socioeconomic students as much as they helped high socioeconomic students. In reading, low socioeconomic students appeared to benefit more whereas in math the high socioeconomic student did.[2]

Many follow-up studies have been conducted on the mid-term effects of CSR found a significantly larger proportion of the small class students than regular or regular with aids class students had passed the Tennessee Competency Examination (TCE) requirement at eighth grade. (Pate-Bain et. al. 1997) Another follow-up study showed similar results. Students in small classes were more likely to take ACT or SAT exams and the difference in proportions of students who took a college entrance exam out of Project STAR participants differed across races. The difference was larger for black students, which indicates CSR benefited black students more in the long run. In addition, the average scores of small class students were significantly higher than that of large class students (Krueger and Whitmore 2001).

Including the official project reports, almost all of existing literature focuses on the average effect of treatment (CSR) gains with the consideration of observable heterogeneities such as sex, age, race, and the socioeconomic status. One notable exception is Ding and Lehrer (2008). They estimated the following quantile regression equation with the kindergarten data:

$$Quantile(Y_{ij}) = \alpha' X_{ij} + \delta' CS_{ij} + v_j + \varepsilon_{ij},$$

[2]Students' socioeconomic status is measured by a dummy variable indicating whether or not student joined a free or reduced price lunch program. If they joined, they were considered to be poor or of low socioeconomic status.

where Y_{ij} is the level of achievement for child i in school j, X_{ij} a vector of student and teacher characteristics, CS_{ij} the actual number of students in the class where child i belonged to, v_j a school fixed effect, and ε_{ij} the random and idiosyncratic unobservable factors. The QTEs are the δ and estimated to be increasing as quantile levels increase for both math and reading. Ding and Lehrer interpreted this as an evidence that higher ability students gain more from the CSR. Their interpretation, though, is only valid when we consider QTE as if they are the quantiles of treatment effects.

3. BOUNDS ON DISTRIBUTION OF TREATMENT EFFECTS

Consider a randomized binary treatment. Let D be an indicator function that takes a value of one if a subject is assigned to the treatment group and zero if he/she is assigned to the control group. Following conventional notations, let $Y_1 (\in \mathscr{Y}_1 \subset \mathscr{R})$ and $Y_0 (\in \mathscr{Y}_0 \subset \mathscr{R})$ be the (potential) continuous outcomes from treatment and control and F_1 and F_0 the marginal distributions of them respectively. Let $\Delta = Y_1 - Y_0$ denote the treatment effect or outcome gain and $F_\Delta(\cdot)$ its distribution function. Due to the randomization assumption, we assume the independence between (Y_1, Y_0) and D i.e. $(Y_1, Y_0) \perp D$.

Since we only observe $Y = DY_1 + (1-D)Y_0$ but not (Y_1, Y_0), we cannot identify Δ nor F_Δ but can identify the (pointwise) lower and upper bounds for F_Δ. Define

$$F^L(\delta) = \max(\sup_y \{F_1(y) - F_0(y-\delta)\}, 0) \text{ and} \tag{1}$$

$$F^U(\delta) = 1 + \min(\inf_y \{F_1(y) - F_0(y-\delta)\}, 0). \tag{2}$$

Then $F^L(\delta) \leq F_\Delta(\delta) \leq F^U(\delta)$ and these bounds are sharp. Similar bounds exist for quantile function of Δ, namely $Q_\Delta(p) = \arg\inf_\delta \{F_\Delta(\delta) \geq p\}$. For any $p \in (0,1)$, $Q_\Delta(\tau) \in \left[Q^L(p), Q^U(p)\right]$, where

$$Q^U(p) = \inf_{u \in (p,1)} [F_1^{-1}(u) - F_0^{-1}(u-p)], \tag{3}$$

$$Q^L(p) = \sup_{u \in (0,p)} [F_1^{-1}(u) - F_0^{-1}(1+u-p)]. \tag{4}$$

See Williamson and Downs (1990) for its proofs. The identification of F^L, F^U, Q^L, and Q^U is straightforward due to the randomized experiment because

$$F_1(y) = F_1(y|D=1) \text{ and } F_0(y) = F_0(y|D=0),$$

i.e. F_1 is identified by the marginal distribution of the treatment group outcomes and F_0 by the marginal distribution of the control group outcomes.

3.1. BOUNDS ON THE DISTRIBUTION OF TREATMENT EFFECTS CONDITIONAL UPON PRE-TREATMENT OUTCOMES

The Y_0 can be viewed as the pre-treatment outcome of the treated. It can be of an important research interest to know how likely the treatment is beneficial to a sub-population defined by pre-treatment outcome. For example, policy makers may want to how much the treatment will be beneficial to the lower first quantile of pre-treatment outcomes. Then the parameter of interest will be $\Pr\left[\Delta \geq 0 | Y_0 \leq F_0^{-1}(0.25)\right]$. The general functions I am considering are

$$\Lambda_y(\delta) = \Pr[\Delta > \delta | Y_0 \leq y] \text{ and} \tag{5}$$

$$\Psi_y(\delta) = \Pr[\Delta > \delta | Y_0 \geq y]. \tag{6}$$

Like the F_Δ or Q_Δ, $\Lambda_y(\delta)$ and $\Psi_y(\delta)$ are only partially identified. Below I provide the bounds for those functions.

Theorem 1. *Let Y_1 and Y_0 be continuous random variables. For y such that $F_0(y) \in (0,1)$, $\Lambda_y^L(\delta) \leq \Lambda_y(\delta) \leq \Lambda_y^U(\delta)$, and $\Psi_y^L(\delta) \leq \Psi_y(\delta) \leq \Psi_y^U(\delta)$, where*

$$\begin{aligned}
\Lambda_y^L(\delta) &= \min\left\{\max\left\{\frac{\sup_{z\leq y}\{F_0(z) - F_1(\delta+z)\}}{F_0(y)}, 0\right\}, 1\right\}, \\
\Lambda_y^U(\delta) &= \max\left\{\min\left\{\frac{\inf_{z\leq y}\{F_0(z) - F_1(\delta+z)\}+1}{F_0(y)}, 1\right\}, 0\right\}, \\
\Psi_y^L(\delta) &= \min\left\{\max\left\{\frac{\sup_{z\geq y}\{F_0(z) - F_1(\delta+z)\} - F_0(y)}{1-F_0(y)}, 0\right\}, 1\right\}, \\
\Psi_y^U(\delta) &= \max\left\{\min\left\{1 + \frac{\inf_{z\geq y}\{F_0(z) - F_1(\delta+z)\}}{1-F_0(y)}, 1\right\}, 0\right\}.
\end{aligned}$$

Although a bit more complicated, the interpretation of these bounds is similar to that of bounds of distribution of treatment effects. See Appendix A for its proof.

3.2. TIGHTENING THE BOUNDS

Although the bounds for the marginal distribution and quantile functions are presented, the idea can be easily extended to the bounds for conditional distribution and quantile functions.

(C1) Let (Y_1, Y_0, D, X) have a joint distribution. For all $x \in \mathscr{X}$ (the support of $X, \subset \mathscr{R}^q \times \mathscr{R}^r$), (Y_1, Y_0) is jointly independent of D conditional on $X = x$.[3]

[3]The randomization assumption implies (Y_1, Y_0, X) is jointly independent of D.

(C2) For all $x \in \mathscr{X}$, $0 < p(x) < 1$, where $p(x) = P(D = 1|x)$.

Define

$$\begin{aligned} F^L(\delta|x) &= \sup_y \max(F_1(y|x) - F_0(y-\delta|x), 0), \\ F^U(\delta|x) &= 1 + \inf_y \min(F_1(y|x) - F_0(y-\delta|x), 0). \end{aligned}$$

Under (C1) and (C2),

$$F^L(\delta|x) \leq F_\Delta(\delta|x) \leq F^U(\delta|x),$$

Here, we use $F_\Delta(\cdot|x)$ to denote the conditional distribution function of Δ given $X = x$. A similar conclusion can be made to the bounds of conditional quantile function. Let $Q_{TE}(p|x) = \arg\inf_\delta \{F_\Delta(\delta|x) \geq p\}$. Then, $Q^L(p|x)$ and $Q^U(p|x)$ defined below consist of the bounds for $Q_{TE}(p|x)$:

$$\begin{aligned} Q^U(p|x) &= \inf_{u\in(p,1)} [F_1^{-1}(u|x) - F_0^{-1}(u-p|x)], \\ Q^L(p|x) &= \sup_{u\in(0,p)} [F_1^{-1}(u|x) - F_0^{-1}(1+u-p|x)]. \end{aligned}$$

Under (C1) and (C2), $F_1(y|x)$ and $F_0(y|x)$ are identified by $F_1(y|x) = F_1(y|x, D = 1)$ and $F_0(y|x) = F_0(y|x, D = 0)$. See Fan and Park (2010, 2012).

Using the bounds for conditional distribution or quantile function of Δ, we can tighten unconditional bounds. Under (C1) and (C2),

$$F^L(\delta|x) \leq F_\Delta(\delta|x) \leq F^U(\delta|x).$$

Since $F_\Delta(\delta) = E_X[F_\Delta(\delta|X)]$, we can construct the following bounds for $F_\Delta(\delta)$:

$$E_X\left[F^L(\delta|X)\right] \leq F_\Delta(\delta) \leq E_X\left[F^U(\delta|X)\right].$$

If X is independent of (Y_1, Y_0), then these new bounds on $F_\Delta(\delta)$ reduce to those in (1) and (2). If X is not independent of (Y_1, Y_0), then the above bounds are in general tighter than them. It can be intuitively understood that $E_X\left[F^L(\delta|X)\right] = E_X\left[F^U(\delta|X)\right]$ if both Y_1 and Y_0 are deterministic functions of X. We can tighten the bounds for $Q_{TE}(p)$, $\Lambda_y(\delta)$, and $\Psi_y(\delta)$ similarly.

3.3. PARTIAL IDENTIFICATION AND HOMOGENEOUS TREATMENT EFFECTS

When the treatment effects are constant over agents i.e. $\Delta = \bar{\Delta}$, then $F_\Delta(\delta) = 0$ if $\delta < \bar{\Delta}$; 1 if $\delta \geq \bar{\Delta}$, which implies $\sup\left\{F^U(\delta) - F^L(\delta)\right\} = 1$. Therefore if

$\sup\left\{F^U(\delta) - F^L(\delta)\right\} < 1$ then it is the evidence against the constant treatment effects.

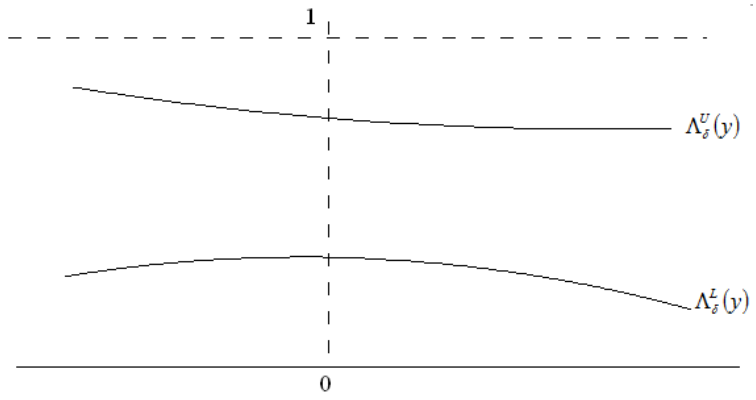
It can also be of interest whether Δ is independent of Y_0 conditional/unconditional on covariates. Suppose Δ (potentially a function of X) is independent of Y_0. Then

$$\begin{aligned}\Lambda_y(\delta,x) &= \Pr[\Delta > \delta | Y_0 \leq y, x] = \Pr[\Delta > \delta | x] \text{ and} \\ \Psi_y(\delta,x) &= \Pr[\Delta > \delta | Y_0 \geq y, x] = \Pr[\Delta > \delta | x]\end{aligned}$$

i.e. they are not functions of Y_0. Therefore $\Lambda_y(\delta) = E_X[\Lambda_y(\delta,X)] = E_X[\Psi_y(\delta,X)] = \Psi_y(\delta)$ are constant in Y_0 for all δ. If $\sup_y \Lambda_y^L(\delta) > \inf_y \Lambda_y^U(\delta)$ or $\sup_y \Psi_y^L(\delta) > \inf_y \Psi_y^U(\delta)$ for some δ then it is the evidence against the independence of Δ to Y_0.[4]

Figure 1 is an example graph of $\Lambda_y^L(\delta)$ and $\Lambda_y^U(\delta)$ that are consistent with assumption of $\Delta \perp Y_0$. The horizontal axis is y in the graph and I changed the notations of $\Lambda_y^L(\delta)$ and $\Lambda_y^U(\delta)$ to $\Lambda_\delta^L(y)$ and $\Lambda_\delta^U(y)$ to visualize that the curves are functions y for a fixed δ. Having a space in the middle of $\Lambda_\delta^L(y)$ and $\Lambda_\delta^U(y)$ throughout the entire domain of y shows $\sup_y \Lambda_y^L(\delta) > \inf_y \Lambda_y^U(\delta)$.

Figure 1: Consistent with $\Delta \perp Y_0$



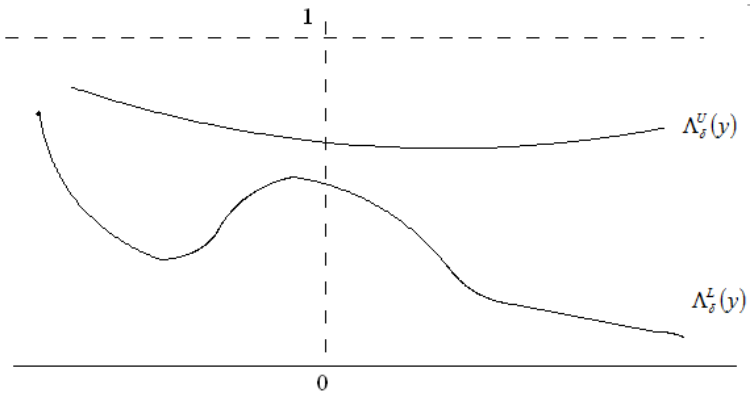
If, on the other hand, graphs of $\Lambda_y^L(\delta)$ and $\Lambda_y^U(\delta)$ against y look like Figure 2 then Δ is clearly dependent of Y_0 because not a single horizontal line

[4]Heckman, Smith, and Clements (1997) showed that, when Δ is independent of Y_0,

$$F_\Delta(\delta) = \frac{1}{2} + \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{it}\left(e^{it\delta}\frac{E\left[e^{itY_1}\right]}{E\left[e^{itY_0}\right]} - e^{-it\delta}\frac{E\left[e^{itY_1}\right]}{E\left[e^{itY_0}\right]}\right)dt.$$

can fit within two curves without crossing one or another, which visualizes $\sup_y \Lambda_y^L(\delta) < \inf_y \Lambda_y^U(\delta)$.

Figure 2: Inconsistent with $\Delta \perp Y_0$



3.4. ESTIMATION

Let the empirical distribution function of Y_j be F_{jn} i.e.

$$F_{nj}(y) = \frac{1}{n_j}\sum_{i=1}^{n_j} 1_{\{Y_{ji}\le y\}} \text{ for } j = 1,0$$

where the sample data consists of $\{Y_{1i}\}_{i=1}^{n_1}$ and $\{Y_{0i}\}_{i=1}^{n_0}$ and $1_{\{A\}}$ is an indicator function that takes a value 1 if A happens and 0 if A does not. Fan and Park (2010, 2012) suggested the following estimators for F^L, F^U, Q^L, and Q^U:

$$\begin{aligned} F_n^L(\delta) &= \max\left\{\sup_{y\in\mathscr{Y}_\delta}\{F_{n1}(y) - F_{n0}(y-\delta)\}, 0\right\}; \\ F_n^U(\delta) &= 1 + \min\left\{\inf_{y\in\mathscr{Y}_\delta}\{F_{n1}(y) - F_{n0}(y-\delta)\}, 0\right\}; \\ Q_n^L(p) &= \sup_{u\in(0,p)}\{F_{n1}^{-1}(u) - F_{0n}^{-1}(1+u-p)\}; \\ Q_n^U(p) &= \inf_{u\in(p,1)}\{F_{n1}^{-1}(u) - F_{0n}^{-1}(u-p)\}. \end{aligned} \tag{7}$$

The inverses of empirical distribution functions are defined as follows:

$$F_{nj}^{-1}(p) = Y_{jn(i)} \text{ for } p \in \left(\frac{i-1}{n_j}, \frac{i}{n_j}\right], i = 1,...,n_j,\ F_{jn}^{-1}(0) = Y_{jn(1)},$$

where $Y_{jn(1)} \leq ... \leq Y_{jn(n)}$ are the order statistics of $\{Y_{ji}\}_{i=1}^{n_j}$. I will summarize in the next subsection the inference on the partially identified $F_\Delta(\delta)$. See Fan and Park (2010, 2012) for the inference on $F_\Delta(\delta)$ or $Q_\Delta(p)$.

Estimatiors for $\Lambda_y^L, \Lambda_y^U, \Psi_y^L, \Psi_y^U$ are:

$$\begin{aligned}
\hat{\Lambda}_y^L(\delta) &= \min\left\{\max\left\{\frac{\sup_{z\leq y}\{F_{n0}(z)-F_{n1}(\delta+z)\}}{F_{n0}(y)},0\right\},1\right\}; \qquad (8)\\
\hat{\Lambda}_y^U(\delta) &= \max\left\{\min\left\{\frac{\inf_{z\leq y}\{F_{n0}(z)-F_{n1}(\delta+z)\}+1}{F_{n0}(y)},1\right\},0\right\};\\
\hat{\Psi}_y^L(\delta) &= \min\left\{\max\left\{\frac{\sup_{z\geq y}\{F_{n0}(z)-F_{n1}(\delta+z)\}-F_0(y)}{1-F_{n0}(y)},0\right\},1\right\};\\
\hat{\Psi}_y^U(\delta) &= \max\left\{\min\left\{1+\frac{\inf_{z\geq y}\{F_{n0}(z)-F_{n1}(\delta+z)\}}{1-F_{n0}(y)},1\right\},0\right\}
\end{aligned}$$

for y such that $F_{n0}(y) \in (0,1)$.

To tighten these bounds, we need to estimate conditional distribution functions nonparametrically. Let $G_j(x)$ be $\Pr[X \leq x|D=j]$ for $j=1,0$ and

$$\operatorname{plim} G_{nj}(x) = \sum_{i=1}^{n_j} K_\gamma(X_i,x) = G_j(x)$$

for a multivariate kernel function $K_\gamma(\cdot,\cdot)$. Then

$$F_{nj}(y|x) = \frac{\frac{1}{n_j}\sum_{i=1}^{n_j} 1_{\{Y_{ji}\leq y\}} G_{nj}(x)}{\frac{1}{n_j}\sum_{i=1}^{n_j} G_{nj}(x)}.$$

We can also define $\hat{\Lambda}_y^L(\delta|x)$, $\hat{\Lambda}_y^U(\delta|x)$, $\hat{\Psi}_y^L(\delta|x)$, and $\hat{\Psi}_y^L(\delta|x)$ by simply replacing $F_{n0}(\cdot)$ by $F_{n0}(\cdot|x)$ and $F_{n1}(\cdot)$ by $F_{n1}(\cdot|x)$ in (8) for a y such that:

$$\begin{aligned}
y &\geq Y_{jn(1)} \text{ for } \hat{\Lambda}_y^L(\delta) \text{ and } \hat{\Lambda}_y^U(\delta);\\
y &< Y_{jn(n_j)} \text{ for } \hat{\Psi}_y^L(\delta) \text{ and } \hat{\Psi}_y^U(\delta).
\end{aligned}$$

The estimators for tightened bounds for $F_\Delta(\delta)$ are:

$$\begin{aligned}
E_X\left[\widehat{F^L(\delta|X)}\right] &= \frac{1}{n_1+n_0}\left(\sum_{i=1}^{n_1} F_n^L(\delta|X_i)+\sum_{i=1}^{n_0} F_n^L(\delta|X_i)\right); \qquad (9)\\
E_X\left[\widehat{F^U(\delta|X)}\right] &= \frac{1}{n_1+n_0}\left(\sum_{i=1}^{n_1} F_n^U(\delta|X_i)+\sum_{i=1}^{n_0} F_n^U(\delta|X_i)\right)
\end{aligned}$$

where $F_n^L(\delta|x)$ and $F_n^U(\delta|x)$ are

$$\begin{aligned}
F_n^L(\delta|x) &= \sup_y \max\{F_{1n}(y|x)-F_{0n}(y-\delta|x),0\},\\
F_n^U(\delta|x) &= 1+\inf_y \min\{F_{1n}(y|x)-F_{0n}(y-\delta|x),0\}.
\end{aligned}$$

The tightened bounds for $\Lambda_y(\delta)$ and $\Psi_y(\delta)$ are defined in the same manner.

Statistical inference on $F_\Delta(\delta)$ and $Q_\Delta(p)$ is dealt in Fan and Park (2010, 2012) but that on $\Lambda_y(\delta)$ and $\Psi_y(\delta)$ is not done yet. I am working on the problem and will only provide estimation results on the bounds of $\Lambda_y(\delta)$ and $\Psi_y(\delta)$ without statistical inference in this paper.

4. ESTIMATION RESULTS

4.1. BOUNDS FOR UNCONDITIONAL DISTRIBUTION OF TREATMENT EFFECTS

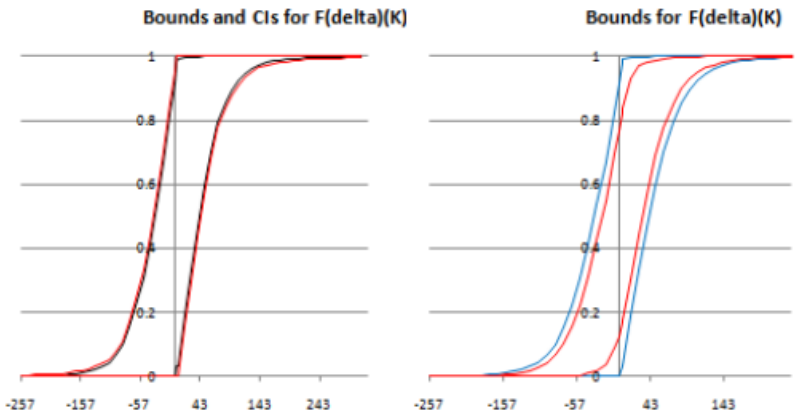
Like Ding and Lehrer (2008), I first used the kindergarten data for the reading test. Of the three types of classes, I considered 'small class' as the treatment($D=1$) and 'regular class' as the control($D=0$). The reading score of a student in a 'small class' is Y_1 and that of a student in a 'regular class' is Y_0. The following table shows some simple descriptive statistics.

	$D=1$	$D=0$
# of students included in the analysis	1731	1977
Avg. reading scores	440.57	434.85
(Std. dev.)	(32.55)	(31.04)

The average treatment effect (ATE) is, then, $440.57-434.85=5.72$ and the t-statistic is $\frac{5.72}{\sqrt{\frac{32.55^2}{1731}+\frac{31.04^2}{1977}}}=5.45$ under $H_0: E[Y_1]=E[Y_0]$, hence the hypotheses is rejected at $\alpha=0.01$ i.e. a statistically significant positive ATE.

Figure 3 shows estimation results for bounds of $F_\Delta(\delta)$. The horizontal axis is δ. The '(K)' in titles of both panels stands for 'Kindergarten'. The left panel of the figure presents $F_n^L(\delta)$ and $F_n^U(\delta)$(black lines) and $CI_{0.95}$ in Fan and Park (2010) (red lines). For $\delta=0$, $F_n^L(0)=0.0$ and $F_n^U(0)=0.9160$. The 95% confidence interval for $F_\Delta(0)$ is $[0.0, 0.9495]$. If the treatment effects was constant at $\Delta=0$ (i.e. no effects) then $F_\Delta(0)$ would be a step function at $\delta=0$ hence $F^L(0)$ had to be 0 and $F^U(0)$ had to be 1. Therefore, the 95% CI for $F_\Delta(0)$ being strictly included in $[0,1]$ implies the constant zero treatment effects (or no treatment effects to all students) would be rejected. $F_n^U(0)=0.9160$ implies about 8.4% of population (kindergarten students) are estimated to be harmed by the CSR. The bounds for $F_\Delta(5.72)$ are estimated $[0.034, 0.994]$ and the 95% CI for it is $[0.002, 1]$. Therefore the hypothesis of $H_0: \Delta=5.72$ for all students is rejected at 5% as well.

Figure 3: Bounds for F_Δ



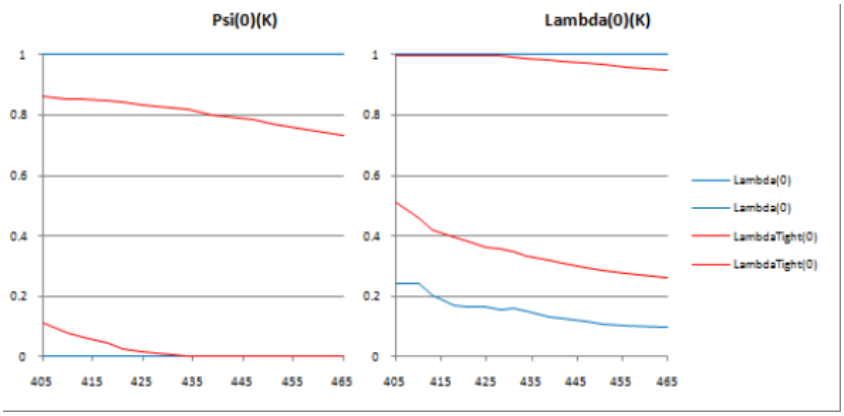
The right panel of 3 compares two bounds: the blue lines are the bounds in (7) and the red lines are the tightened bounds in (9). To estimate the tightened bounds, I used a Gaussian kernel with bandwidths $h_s = 1.06\hat{\sigma}_{X_k^c} n_j^{-1/\left(4+\dim\left(X^C\right)\right)}$ where $\hat{\sigma}_{X_k^c}$ is the sample standard deviation of k-th continuous covariate and $\dim\left(X^C\right)$ the number of continuous variables. I used 0.05 for the tuning parameter for discrete covariates arbitrarily. See Qi and Racine (2007).

The use of covariates tightened the bounds a little bit. The bounds for $F_\Delta(0)$ now became $[0.1238, 0.7688]$, that is to say the CSR was estimated to be harmful at least 12.38% of kindergarten students and at the worst this proportion can go up to 76.88%. The bounds for $F_\Delta(5.72)$ is now $[0.1778, 0.8368]$. At least 17.78% of students benefit less than the estimated ATE.

Figure 4 presents bounds for $\Lambda_y(\delta)$ and $\Psi_y(\delta)$ for $\delta = 0$ (blue lines) and the tightened version of them (red lines) for various values of y from 15%-tile to 85%-tile of Y_0 on the horizontal axis.

The bounds for $\Psi_y(0)$, for $y = F_{n0}^{-1}(0.8) = 456$ or the upper quintile, are estimated $[0.0004, 0.7598]$ by the tightened bounds. This reads that if a student performs within the upper quintile in a regular class room environment, the probability that he/she will do better in a small class environment is at most 0.7598. Or, in other words, at most 75.98% of upper quintile students in an large class environment will benefit by the CSR. On the other hand the bounds for $\Lambda_y(0)$ for $y = F_{n0}^{-1}(0.2) = 410$ are estimated $[0.4556, 0.9994]$ by the the tightened bounds. It reads at least 45.56% of lower quintile students will benefit from the CSR. Had

Figure 4: Estimation of $\Lambda_y(\delta)$ and $\Psi_y(\delta)$



these bounds been tighter so that $\Psi^U_{F^{-1}_{n0}(0.8)}(0) < \Lambda^L_{F^{-1}_{n0}(0.2)}(0)$ then we would have been able to tell definitively that the CSR would benefit the lower quintile students than the upper quintile ones. Since, however, the bounds overlap for all y considered here, nothing can be said definitively.

For $y = 431$, the median of $\{Y_{0i}\}_{i=1}^{n_0}$, the bounds for $\Psi_{431}(0)$ are $[0.0059, 0.8288]$ and $\Lambda_{431}(0)$ are $[0.3467, 0.9921]$: at most 82.88% of 'better-than-median' students benefit meaning that at least 17.12% of 'better-than-median' students will be harmed by the CSR; at least 34.67% of 'worse-than-median' student benefit.

4.2. ANALYSIS OF SUBGROUPS

To see how the heterogeneity in treatment effects differs with students' characteristics, I split the whole sample into eight subgroups according to student's race, sex, and socioeconomic status. The socioeconomic status is measured by whether or not student received free lunch. The subgroups[5] and related descriptive statistics are shown in Table 1.

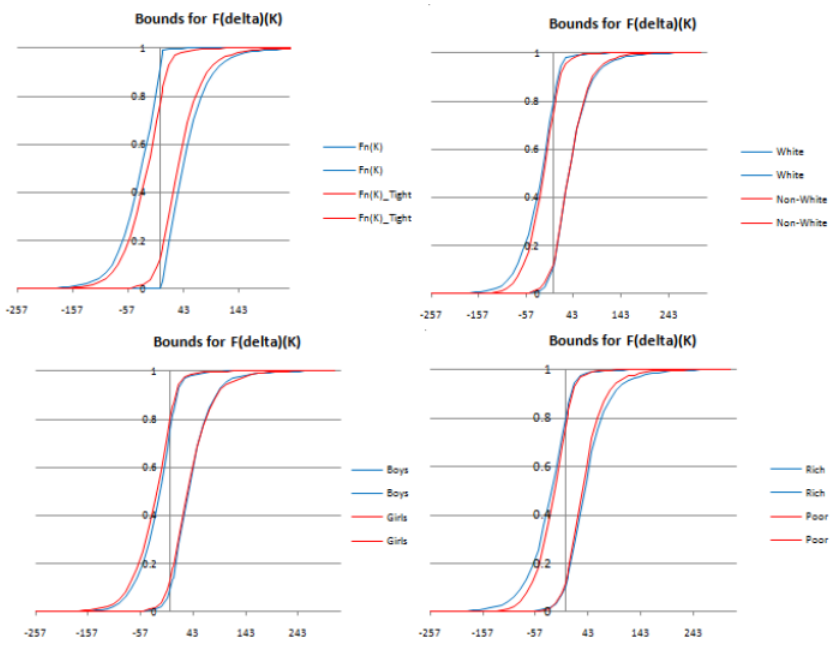
Graphs in Figure 5 show the tightened bounds for $F_\Delta(\delta)$ for indicated subgroups. The horizontal axis is δ.

[5] As is evident, the subgroups are not mutually exclusive. I did not construct mutually exclusive subgroups due to the concern about the number of observations in each subgroup. Another method that can be used is a nonparametric or semiparametric estimation with covariates.

Table 1: Subgroup Categories

Subgroup categorization		Class	Reading	
			# of obs.	Average
Ethnicity	White	S	1183	443.7
		R	644	425.6
	Non-White	S	548	433.8
		R	1333	439.3
Sex	Male	S	889	438.3
		R	962	439.8
	Female	S	842	442.9
		R	1015	430.1
Socio-Economic Status	Rich (No Free Lunch)	S	913	448.3
		R	946	425.5
	Poor (Free Lunch)	S	818	431.9
		R	1031	443.4

Figure 5: Bounds for $F_{\Delta}(0)$ for subgroups



The left panel of Figure 5 presents the bounds and confidence intervals and

the and the right panel the tightened bounds like Figure 3. The graphs in the upper right corner compare White vs. Non-White groups, the graphs in the lower left corner Boys vs. Girls, and finally the lower right ones Rich vs. Poor. The White and Non-White comparison groups show quite noticeable differences on the upper bound. The Rich seems to have strictly larger bounds than the Poor for all δ.

The estimated bounds for $F_\Delta(0)$ are:

	White	Non-White	Boy	Girl	Rich	Poor
$F_n^L(0)$	0.1115	0.1190	0.0932	0.1315	0.1063	0.1190
$F_n^U(0)$	0.7907	0.7485	0.7588	0.8025	0.7970	0.7620

Overall there are not noticeable differences between subgroups. It appears: at least 13.15% of Girls are harmed whereas the minimum fraction of Boys who get harmed is only 9.3%. However Girls have higher upper bounds.; the Poor has slightly higher lower bounds and low upper bound, which may indicate the CSR may work against the Poor.

Next, the bounds for $\Psi_y(0)$ and $\Lambda_y(0)$ are plotted against various y from the15%-tile to 85%-tile of Y_0 for different comparison subgroups on Figures 6 $\sim$ 8.

Figure 6: Bounds for $\Psi_y(0)$ and $\Lambda_y(0)$ for White vs. Nonwhite

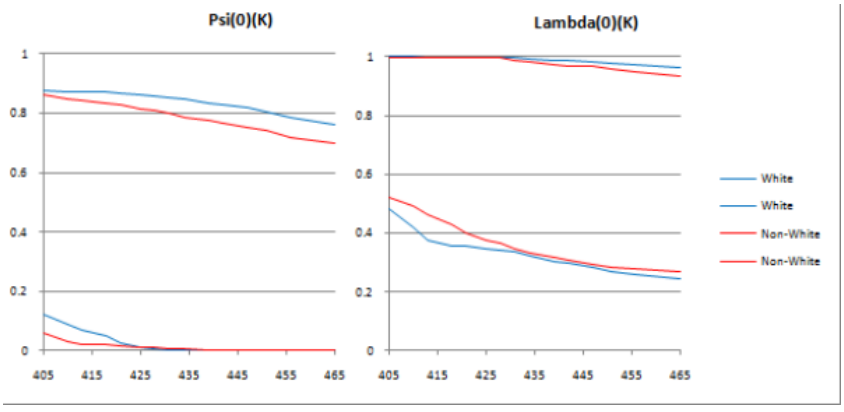


Figure 7: Bounds for $\Psi_y(0)$ and $\Lambda_y(0)$ for Boys and Girls

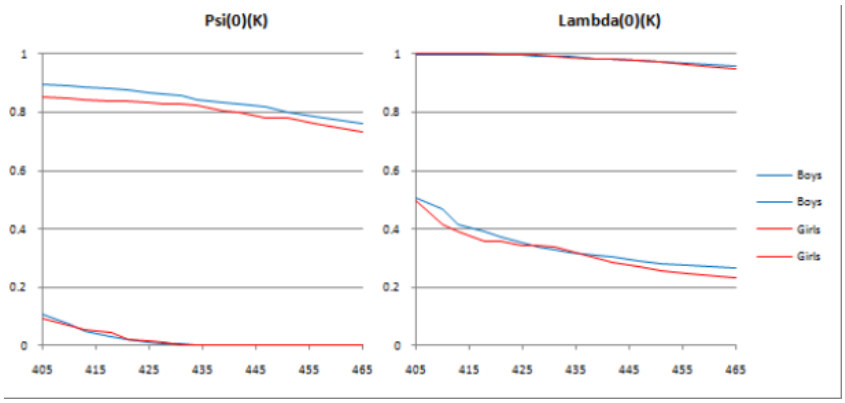
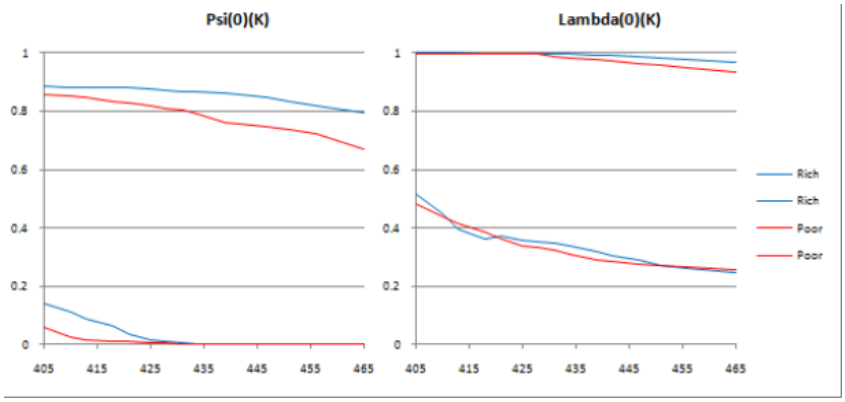


Figure 8: Bounds for $\Psi_y(0)$ and $\Lambda_y(0)$ for the Rich and the Poor



The bounds for $\Psi_y(0)$ reveal some noticeable differences in White-Non White comparison and Rich-Poor comparison. Bounds of White are higher than that of Non-White. Bounds of Rich are higher than that of Poor.

First thing noticeable is that $\Psi_y^U(0)$'s differ across comparison pairs whereas both $\Lambda_y^L(0)$ and $\Lambda_y^U(0)$ are estimated similar in each pair except for Rich-Poor comparison. The Rich-Poor subgroups show somehow different $\Lambda_y^L(0)$'s. The

bounds for $\Psi_y(0)$ and $\Lambda_y(0)$ for some y are:

	White	Non-White	Boy	Girl	Rich	Poor
$\Psi^L_{F_{n0}^{-1}(0.8)}(0)$	0.0001	0.0016	0.0005	0.0003	0.0006	0.0
$\Psi^U_{F_{n0}^{-1}(0.8)}(0)$	0.7910	0.7237	0.7887	0.7656	0.8216	0.7267
$\Lambda^L_{F_{n0}^{-1}(0.2)}(0)$	0.4164	0.4902	0.4662	0.4138	0.4469	0.4352
$\Lambda^U_{F_{n0}^{-1}(0.2)}(0)$	0.9998	0.9991	0.9992	1.0	1.0	0.9992

$\Psi^U_{F_{n0}^{-1}(0.8)}(0) > \Lambda^L_{F_{n0}^{-1}(0.2)}(0)$ in all subgroups hence whether or not the CSR works for or against lower quintile students in each subgroup is not definitive.

5. CONCLUSION

When treatment effects are heterogeneous identification of ATE is not enough as in Bitler et al. (2006). If possible, identification of the entire distribution of treatment effects is desired. The idea and technique introduced and proposed in this paper is using the bounding approach to partially identify the distribution. A new finding is we are able to use bounds for conditional distribution of treatment effects given pre-treatment outcome levels.

Application of the techniques for finding the bounds to Project STAR seemed to have yield a bit wider bounds for the conditional distribution, which may limit applicability or usefulness of the techniques. It is a fundamental limitation that the approach can't help but provide partial, hence imprecise, knowledge about what we want to know however 'how much imprecise' depends on data and probabilistic nature of them so having more tools in our toolbox is good. Even current investigation with the Project STAR that was not very satisfactory in terms of precision of empirical results (i.e. the width of identification region) provided some useful insights about heterogeneity of CSR effects as mentioned in Section 4.

Statistical inference of the conditional distribution of treatment effects on pre-treated outcomes, which is yet to develop, is another unsatisfactory aspect of current paper. It may require another paper.

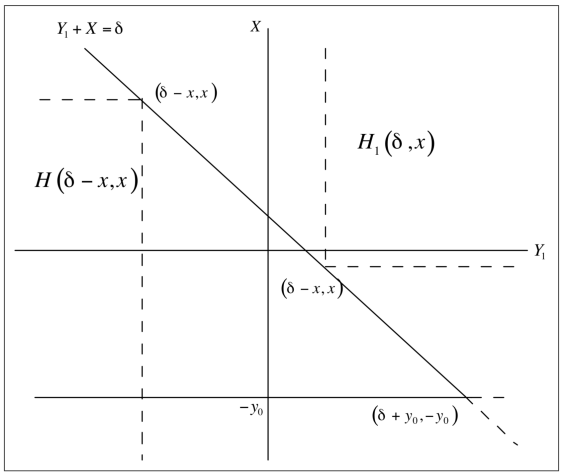
APPENDIX A. PROOF OF THEOREM 1

I will provide the bounds for $\Pr[\Delta \leq \delta, Y_0 \leq y_0]$ then the bounds for $\Lambda_{y_0}(\delta)$ can be computed by

$$\Lambda_{y_0}(\delta) = \Pr[\Delta > \delta | Y_0 \leq y_0] = 1 - \frac{\Pr[\Delta \leq \delta, Y_0 \leq y_0]}{\Pr[Y_0 \leq y_0]}.$$

Let $X = -Y_0$. Then $\Pr[\Delta \leq \delta, Y_0 \leq y_0] = \Pr[\Delta \leq \delta, X \geq -y_0]$. Define $W(u,v) = \max\{u+v-1, 0\}$ and $M(u,v) = \min\{u,v\}$. Also define $H(y_1, x) = \Pr[Y_1 \leq y_1, X \leq x]$.

$\Pr[\Delta \leq \delta, X \geq -y_0]$ is the H-volume of upper left half plane surrounded by the lines $Y_1 + X = \delta$ and $X = -y_0$.



As in Nelson (1999), $\Pr[\Delta \leq \delta, X \geq -y_0]$ is bounded from below by $\sup_{x \geq -y_0} H(\delta - x, x)$. Therefore,

$$\begin{aligned}
& \Pr[\Delta \leq \delta, X \geq -y_0] \\
\geq\ & \max\left\{\sup_{x \geq -y_0} H(\delta - x, x) - H(\delta - x, -y_0), 0\right\} \\
\geq\ & \max\left\{\sup_{x \geq -y_0} [\max\{F_1(\delta - x) + F_X(x) - 1, 0\} - \min\{F_1(\delta - x), F_X(-y_0)\}], 0\right\} \\
=\ & \max\left\{\sup_{y \leq y_0} [\max\{F_1(\delta + y) - F_0(y), 0\} - \min\{F_1(\delta + y), 1 - F_0(y_0)\}], 0\right\} \\
=\ & \max\left\{\sup_{y \leq y_0} \{F_1(\delta + y) - F_0(y) - 1 + F_0(y_0)\}, 0\right\}
\end{aligned}$$

For the upper bound, we know

$$\begin{aligned}
& \Pr[\Delta \leq \delta, X \geq -y_0] \\
\leq\ & \inf_{x \geq -y_0} \{1 - F_X(-y_0) - H_1(\delta, x)\} \\
=\ & \inf_{x \geq -y_0} \{1 - F_X(-y_0) - \{1 - F_1(\delta - x) - F_X(x) + H(\delta - x, x)\}\} \\
\leq\ & \inf_{x \geq -y_0} \{-F_X(-y_0) + F_1(\delta - x) + F_X(x) - \max\{F_1(\delta - x) + F_X(x) - 1, 0\}\} \\
=\ & \inf_{y \leq y_0} \{F_0(y_0) + F_1(\delta + y) - F_0(y) - \max\{F_1(\delta + y) - F_0(y), 0\}\} \\
=\ & F_0(y_0) + \min\{\inf_{y \leq y_0} \{F_1(\delta + y) - F_0(y)\}, 0\}
\end{aligned}$$

The proof for $\Psi_{y_0}(\delta)$ require the bounds for $\Pr[\Delta \leq \delta, Y_0 \geq y_0]$. The proof is analogous except that we have to start with

$$\begin{aligned}
& \sup_{x \leq -y_0} H(\delta - x, x) \\
\leq\ & \Pr[\Delta \leq \delta, X \geq -y_0] \\
\leq\ & \inf_{x \leq -y_0} \{F_X(x) + H(\delta - x, -y_0) - H(\delta - x, -x)\}.
\end{aligned}$$

REFERENCES

[1] Abbring, J. H., and J. J. Heckman (2007). "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation." *Handbook of Econometrics 6b*, 5145-5303.

[2] Bitler, M., J. Gelbach, and H. W. Hoynes (2006). "What Mean Impact Miss: Distributional Effects of Welfare Reform Experiments." *American Economic Review* 96, 988-1012.

[3] Bourke, S. (1986) "How Smaller Is Better: Some Relationships Between Class Size, Teaching Practices and Student Achievement." *American Educational Research Journal*, 23: 558-571.

[4] Ding, W. and S. Lehrer (2004). "Estimating Dynamic Treatment Effects from Project STAR." Mimeo.

[5] Ding, W. and S. Lehrer (2008). "Class Size and Student Achievement: Experimental Estimates of Who Benefits andWho Loses from Reductions." Mimeo.

[6] Djebbari, H. and J. A. Smith (2004). "Heterogeneous Program Impacts in PROGRESA." Mimeo.

[7] Dennis, B. D. (1986), "Effects of Small Class Size (1:15) on the Teaching/Learning Process in Grade Two." Dissertation. Tennessee State University, 177.

[8] Fan, Y. and S. Park (2010), "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference," *Econometric Theory* 26, 1-21.

[9] Fan, Y. and S. Park (2012), "Confidence intervals for the quantile of treatment effects in randomized experiments," *Journal of Econometrics* 167, 330-344.

[10] Firpo, S. P. and Ridder, G (2010). *Bounds on functionals of the distribution treatment effects*. Textos para discussão 201, FGV/EESP - Escola de Economia de São Paulo, Getulio Vargas Foundation (Brazil).

[11] Frank, M. J., R. B. Nelsen, and B. Schweizer (1987). "Best-Possible Bounds on the Distribution of a Sum—a Problem of Kolmogorov." *Probability Theory and Related Fields* 74, 199-211.

[12] Glass, G. V., and M. L. Smith. (1978) *Meta-Analysis of Research on the Relationship of Class Size and Achievement*. San Francisco: Far West Laboratory of Educational Research and Development.

[13] Hallinan, M. T. and A. B. Sorensen. (1985) "Ability Grouping and Student Friendships." *American Educational Research Journal* 22, 485-499.

[14] Hanushek, E. A. (1998), "The Evidence on Class Size." Occasional Paper Number 98-1, W. Allen Wallias Institute of Political Economy, University of Rochester.

[15] Heckman, J., J. Smith, and N. Clements (1997). "Making The Most Out of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts." *Review of Economic Studies* 64, 487-535.

[16] Li, Q. and J. Racine (2007), *Nonparametric Econometrics, Theory and Practice*. Princeton University Press.

[17] Nelsen, R. B. (1999). *An Introduction to Copulas*. Springer, New York.

[18] Williamson, R. C. and T. Downs (1990). "Probabilistic Arithmetic I: Numerical Methods for Calculating Convolutions and Dependency Bounds." *International Journal of Approximate Reasoning* 4, 89-158.