

# Class Size and Educational Policy: Who Benefits from Smaller Classes?

Esfandiar Maasoumi

Southern Methodist University

Daniel L. Millimet\*

Southern Methodist University

Vasudha Rangaprasad

Southern Methodist University

September 2005

## Abstract

The impact of class size on student achievement remains an open question despite hundreds of empirical studies and the perception amongst parents, teachers, and policymakers that larger classes are a significant detriment to student development. This study sheds new light on this ambiguity by utilizing nonparametric tests for stochastic dominance to analyze unconditional and conditional test score distributions across students facing different class sizes. Analyzing the conditional distributions of test scores (purged of observables using class-size specific returns), we find that there is little causal effect of marginal reductions in class size on test scores within the range of 20 or more students. However, reductions in class size from above 20 students to below 20 students, as well as marginal reductions in classes with fewer than 20 students, increase test scores for students *below the median*, but decrease test scores *above the median*. This non-uniform impact of class size suggests that compensatory school policies, whereby lower-performing students are placed in smaller classes and higher-performing students are placed in larger classes, improves the academic achievement of not just the lower-performing students, but also the higher-performing students.

**JEL:** C14, C33, I21, I28

**Keywords:** Class Size, Student Achievement, School Quality, Quantile Treatment Effects, Stochastic Dominance, Program Evaluation

---

\*The authors are extremely grateful to David Drukker for assistance, Julian Betts, Eric Hanushek, and Christopher Jepsen for comments, seminar participants at the 8th Annual Texas Camp Econometrics, Georgetown University, University of Texas - Arlington, and the SOLE Labor Economics Internet Seminar. Corresponding address: Daniel Millimet, Department of Economics, Box 0496, Southern Methodist University, Dallas, TX 75275-0496. Tel: (214) 768-3269. Fax: (214) 768-1821. E-mail: millimet@mail.smu.edu.

# 1 Introduction

The effect of school quality, notably class size, on student achievement is one of the most debated educational policy issues (e.g., Hanushek 1986, 1996, 2003; Krueger 2003). While a number of researchers have analyzed the issue, no consistent relationship between class size and student outcomes has been identified. In a recent review, Akerhielm (1995) noted that of 112 studies in the literature, 23 documented a statistically significant relationship between class size and student achievement: 14 finding a negative relationship, nine showing a positive relationship.<sup>1</sup> Consequently, Krueger (2003, p. F61) labels the effect “subtle and easily obscured,” Fertig (2003b, p. 2) calls the mixed results “disconcerting,” and Card and Krueger (1996, p. 47) note that “decisions about educational resources and reform have to be made in an environment of much uncertainty.” In spite of these ambiguous findings, the common perception among parents, teachers, and policymakers is that larger classes are a significant detriment to student development. Several states, including Texas, have mandated maximum class size levels. As a result, Dustmann (2003, p. F1) labels class size reduction policies as “easily the most popular policy for school improvements in the US,” despite the fact that the efficacy of such policies is “an issue of ongoing debate.” Hanushek (2002, p. 61) concludes: “Despite the political popularity of overall class size reduction, the scientific support of such policies is weak to nonexistent.”

Given the importance of student development, future labor market success, racial achievement and wage gaps, economic growth, and the limited budgets that school districts confront, improved understanding of any benefits of reductions in class size, and who benefits, are crucial to informed policymaking. Unfortunately, all existing empirical studies (to our knowledge) are lacking in at least one of three important respects, contributing to the inconsistent results noted above. First, in the past, most studies utilized the average pupil-teacher ratio at an individual’s school to proxy for the actual class size faced by the individual, given the absence of data on actual class size. However, the relationship between student performance and average pupil-teacher ratios is likely to be weaker (Ehrenberg et al. 2001). Second, nearly all studies focus on only one aspect of the distribution of student achievement: the (conditional) mean.<sup>2</sup> This is helpful only to a decision maker who places equal weights on all student groups. Finally, class size is treated as an intercept effect/shift, potentially missing important interactions between class size and

---

<sup>1</sup>More recently, Todd and Wolpin (2003, p. F4) state that despite “hundreds of empirical studies of the school quality-achievement relationship,” such studies “do not appear to be converging toward a consensus.” See also Hanushek (2002).

<sup>2</sup>Noteable exceptions are Eide and Showalter (1998), Levin (2001), and Fertig (2003a), who use quantile regression methods to analyze the effect of school attributes on student achievement at select points in the distribution. However, Eide and Showalter (1998) only have data on the pupil-teacher ratio at the school level, not actual class size, and Levin (2001) and Fertig (2003a) analyze Dutch and German data, respectively.

other educational inputs. For example, it may be that effect of smaller classes arise because they increase the productivity of other inputs, such as teacher quality or parental involvement.

To simultaneously account for these three shortcomings, we use a nationally representative sample of public high school students in the US. The data, taken from the National Educational Longitudinal Study of 1988 (NELS), contain several test score measures, the class size corresponding to the subject being tested, as well as an exhaustive set of conditioning variables. We examine the entire distribution of tenth and twelfth grade test scores, both unconditional and conditional, across several class size groupings, controlling for a host of student, family, and school attributes and incorporating the logic of the Oaxaca-Blinder decomposition to allow for interaction effects between class size and the returns to other educational inputs.

Our comparisons of the distributions of test scores rely on recently developed nonparametric tests for stochastic dominance (SD). Such tests of first, second, and higher order stochastic dominance (hereafter FSD and SSD, respectively) were examined in McFadden (1989) and Kaur et al. (1994). Our implementation of these tests treats the distribution of student test scores as an unknown to be estimated nonparametrically, and draws upon bootstrap techniques developed in Maasoumi and Heshmati (2000) and Linton et al. (2005) to assess the level of statistical confidence regarding various relations.

Testing for SD is a very useful and insightful companion to standard regression analysis for several reasons. First, it allows one to examine effects of the ‘treatment’ in question at different parts of the distribution, as viewed by very different evaluation functions. Because many policy interventions have different effects on high-, middle-, and low-achieving students, any summary (index) measure of this distributed outcome must assign implicit weights to these different groups. For instance, averages assign equal weights, while focusing on the lowest quintile, for instance, places zero weight on others, etc. Second, SD analysis endows a context to averages and estimators of Quantile Treatment Effects (QTEs) which exposes the welfare functions, or policy evaluation functions, underlying the assessment of the differences in the distributions of the outcomes of different groups exposed to different treatments (see, e.g., Bitler et al. 2005). For example, SD analysis may make clear that the only way two situations can be (could have been) ranked is by a summary measure that prefers higher scores and cares about ‘inequality’ or certain trade-offs between groups. Third, finding a SD ranking (of a particular order) indicates that all utility/welfare functions belonging to a particular class would agree that one policy is preferable to another. Thus, summarized comparisons based on specific indices (e.g., mean comparisons in the case of first order SD) are only needed for quantification of the ‘treatment’ effect. In addition, the inability to infer a dominance relation is equally informative, indicating that any (implicit) welfare ordering based on a particular index (such as the conditional mean) is highly subjective; different indices – even within the

same general class of utility/welfare functions – will yield different substantive conclusions. As a result, summary measures may be used for *complete*, strong rankings of student outcomes across classes of varying size, consistent with explicitly revealed welfare preferences. More sophisticated techniques are needed for weak, yet *uniform* ranking of student outcomes over large classes of welfare functions. Such uniform rankings (i.e., rankings robust to the specific choice of preference function *and* distribution of outcomes) are needed for ‘majority’ rankings of policy outcomes. Empirical examination of such uniform rankings, based on the notion of stochastic dominance (SD), is the approach utilized in the current paper.

The power of such SD relations, combined with the recently developed theory necessary to conduct rigorous statistical tests for the presence of such relations, has led to their growing application. For example, Maasoumi and Millimet (2005) examine changes in US pollution distributions over time and across regions at a point in time. Maasoumi and Heshmati (2000) analyze changes in the Swedish income distribution over time as well as across different population subgroups. Fisher et al. (1998) compare the distribution of returns to different length US Treasury Bills. Particularly relevant to the analysis at hand are previous applications of SD to the analysis of various treatment effects on the distribution of the outcome of interest. For instance, Amin et al. (2003) analyze the effect of a micro-credit program in Bangladesh on the distribution of consumption of participants versus non-participants. Abadie (2002) analyzes the impact of veteran status on the distribution of civilian earnings. Bishop et al. (2000) compare the distribution of nutrition levels across populations exposed to two different types of food stamp programs. Anderson (1996) compares pre- and post-tax income distributions in Canada over several years.

The results are quite striking. In particular, we reach four main conclusions. First, we find that the *unconditional* distributions of test scores favor students in *larger* classes below the 70<sup>th</sup> percentile or so, and students in *smaller* classes in the upper tail. Specifically, while marginal changes in class size in classes with at least 20 students are not associated with changes in the unconditional test score distributions, the unconditional test scores of low-achieving (high-achieving) students are *superior* in classes with at least (less than) 20 students. Second, analyzing the distribution of test scores purged of a multitude of individual, family, class, teacher, and school attributes, including lagged test scores and grade point average, using *class-size specific returns*, we find that there is little causal effect of marginal reductions in class size on test scores within the range of 20 or more students. However, reductions in class size from above 20 students to below 20 students, as well as marginal reductions in classes with fewer than 20 students, *increase* test scores for students *below the median*, but *decrease* test scores for students *above the median*. Third, our conclusions hinge on the fact that we allow the returns to observables to vary by class size. In fact, unknown heretofore, we conclude that the majority of the beneficial impact of class size reduction arises because of the productivity-enhancing effect it has on other educational inputs. However, the improvement in the

returns to observables is non-monotonic; the overall returns are ‘maximized’ (in some sense) in classes with 16-19 students. Finally, the fact that class size reductions rarely have a uniform impact across the test score distributions (i.e., the QTEs vary across the distributions in terms of sign and magnitude) implies that the current focus on mean ‘treatment’ effects is misleading. Only by expanding one’s analysis to also incorporate the ‘dispersion’ of test scores can policymakers hope to arrive at uniform rankings of class sizes.

In terms of informing educational policy, our results provide two key insights. First, since marginal reductions in class size in classes with more than 20 students have little impact on test scores, absent other considerations (e.g., student discipline), class size reductions in schools with current class sizes well above 20 should only be undertaken if schools are willing to significantly reduce class size close to or below 20 students per class. Second, the non-uniform impact of class size suggests that compensatory school policies, whereby lower-performing students are placed in smaller classes and higher-performing students are placed in larger classes, improves the academic achievement of not just the lower-performing students, but also the higher-performing students.

The remainder of the paper is organized as follows. Section 2 presents an overview of the literature. Section 3 defines the various dominance relations and describes the tests used to identify such relations in the data. Section 4 discusses the data. Section 5 presents the initial results, while Section 6 presents some further detailed analysis. Section 7 offers some concluding remarks.

## 2 Literature Review

Given that student achievement has been linked to variation in economic growth across countries (Hanushek and Kimko 2000; Barro 2001), and that racial disparities in student achievement have been well-documented (Jenks and Phillips 1998; Cook and Evans 2000; Hanushek 2001), a considerable literature has developed attempting to understand the factors contributing to student development. The vast majority of previous studies focus on the pupil-teacher ratio at the school level to proxy for class size (e.g., Coleman et al. 1966; Chubb and Moe 1990; Eide and Showalter 1998), with the resulting link between student achievement and pupil-teacher ratios found to be tenuous at best (Hanushek 1986, 1989). More recent work has utilized actual measures of class size (e.g., Akerhielm 1995; Boozer and Rouse 2001).

In early work, Summers and Wolfe (1977) analyzed data on 627 sixth grade elementary school students across 103 randomly selected elementary schools from the Philadelphia School District in 1970-1971, as well as 553 eighth grade students and 716 twelfth grade students. The authors concluded that disadvantaged students and students from low socioeconomic backgrounds benefited from smaller class sizes. Several

more recent studies have reaffirmed this conclusion. Angrist and Lavy (1999) find a beneficial impact of smaller classes on the achievement of fourth and fifth grade students in Israel using an instrumental variable (IV) approach. Rivkin et al. (2002) use a complex fixed effects model and data on class size within Texas schools across multiple cohorts, documenting some significant positive effects of smaller classes on fourth and fifth grade students, although the effects disappear by sixth grade and in any event are of much smaller magnitude than the impact of teacher quality. Akerhielm (1995) and Boozer and Rouse (2001) use data from the NELS, relying on IV methods to control for potential endogeneity of class size. Both find statistically significant, negative effects of larger classes on test scores. Utilizing experimental evidence on kindergarten through third grade students from Tennessee’s Project STAR (Student/Teacher Achievement Ratio), Krueger (1999) finds a modest, beneficial impact of smaller classes during kindergarten and first grade, but no impact on subsequent achievement.<sup>3</sup> Krueger and Whitmore (2002), also examining the Project STAR data, document a positive impact of class size reductions, particularly for minorities.

Conversely, several studies have offered compelling empirical support for the notion that (i) smaller classes are not as beneficial as the above studies suggest, or (ii) larger classes are actually beneficial for students. Jepsen and Rivkin (2002) examine the effectiveness of California’s Class Size Reduction Program (CSR) on student achievement using longitudinal data from the 1990s. Employing a difference-in-differences approach, the authors find that smaller class sizes raised third grade mathematics and reading test scores, particularly for low income students. However, the reduction in class size lowered the quality of teachers on the margin, and the deterioration in teacher quality at least partially offsets the gains from smaller classes (see also Rivkin et al. (2002) for a similar argument). Hoxby (2000a) employs a similar approach to many of the above studies, relying on exogenous variation in class size in Connecticut arising from idiosyncratic variation in the population, and finds no significant relationship between class size and student performance on tests in fourth and sixth grade. Goldhaber and Brewer (1997) control for school and teacher fixed and random effects, finding a positive impact of class size on mathematics achievement of tenth grade students using the NELS. Wößmann (2003) uses international data from the Trends in International Mathematics and Science Study (TIMSS) on seventh and eighth grade students, spanning 39 countries, and presents statistically significant evidence (using IV and other estimation techniques) that smaller class size is related to *lower* student achievement in mathematics and science. Fertig (2003a), using German data on the reading achievement of 15-16 year old students, also finds a statistically significant positive impact of *larger* classes using both OLS and quantile regression methods and controlling for a host

---

<sup>3</sup>Interestingly, Hanushek (2002) implicitly suggests the benefit of examining achievement distributions, rather than average outcomes, noting that while overall average kindergarten achievement is higher in smaller classes under the STAR experiment, this ranking only holds in 40 of 79 schools participating in the experiment.

of other attributes, including the homogeneity (in terms of ability) of the school population. The author argues that heterogeneity, not class size, is a more important detriment to student performance (see also Fertig (2003b) and Lazear (2001)). Finally, Dobbelsteen et al. (2002, p. 36) utilize Dutch data on fourth, sixth, and eighth grade students and exogenous variation in class size arising from rules linking total school enrollment to the number of teachers. The authors find that “after correcting for endogeneity, pupils in large classes do no worse – and sometimes even better – than identical pupils in small classes.” Consonant with Fertig (2003a,b), the authors test and find support for their hypothesis that students learn from other students of similar ability; thus, larger classes increase the probability of being surrounded by others from whom one may learn (see also Levin 2001).

Finally, Krueger (2003) performs a meta-analysis of the literature, finding in general a small, not particularly robust, positive impact of smaller class sizes. Hanushek (1996, 2002, 2003) also performs several meta-analyses, concluding that there is no consistent relationship between resources and student achievement. This lack of consensus suggests that examination of the QTEs, along with the application of SD testing, may be particularly fruitful. To contrast, it is worth re-emphasizing that practically all the above studies utilizing regression-based inferences face two limitations: (i) the assumed structure of the various educational production functions estimated may be too simplistic as they restrict class size to only an intercept shift, and (ii) the focus is on the *conditional mean* of the distribution of scores/performance and the impact of different conditioning variables thereon.

### 3 Empirical Methodology

#### 3.1 Test Statistics

Our distributional comparisons are based on the notion of SD. Several tests for SD have been proposed in the literature; the approach herein is based on a generalized Kolmogorov-Smirnov test.<sup>4</sup> To begin, let  $X$  and  $Y$  denote two outcome (test score) variables being compared (e.g.,  $X$  ( $Y$ ) might refer to test scores of students exposed to a short (long) school year).  $\{x_i\}_{i=1}^N$  is a vector of  $N$  possibly dependent observations of  $X$ ;  $\{y_i\}_{i=1}^M$  is an analogous vector of realizations of  $Y$ . In the spirit of the historical development of such two-sample tests,  $\{x_i\}_{i=1}^N$  and  $\{y_i\}_{i=1}^M$  each constitute one sample. Thus, we refer to dependence between  $x_i$  and  $x_j$ ,  $i \neq j$ , as *within-sample dependence* (similarly for observations of  $Y$ ), and dependence between  $X$  and  $Y$  as *between-sample dependence*.

Assuming general von Neumann-Morgenstern conditions, let  $\mathcal{U}_1$  denote the class of (increasing) utility functions  $u$  such that utility is increasing in test scores (i.e.  $u' \geq 0$ ), and  $\mathcal{U}_2$  the class of social welfare

---

<sup>4</sup>Maasoumi and Heshmati (2000) provide a brief review of the development of alternative tests.

functions in  $\mathcal{U}_1$  such that  $u'' \leq 0$  (i.e. concavity). Concavity represents risk aversion, or an aversion to inequality in the achievement of students; a high concentration of both high- and low-achieving students is undesirable. Let  $F(x)$  and  $G(y)$  represent the cumulative density functions (CDF) of  $X$  and  $Y$ , respectively, which are assumed to be continuous and differentiable.

Under this notation,  $X$  First Order Stochastically Dominates  $Y$  (denoted  $X$  FSD  $Y$ ) iff  $E[u(X)] \geq E[u(Y)]$  for all  $u \in \mathcal{U}_1$ , with strict inequality for some  $u$ .<sup>5</sup> Equivalently,

$$F(z) \leq G(z) \quad \forall z \in \mathcal{Z}, \text{ with strict inequality for some } z. \quad (1)$$

where  $\mathcal{Z}$  denotes the union of the supports of  $X$  and  $Y$ . If  $X$  FSD  $Y$ , then the expected welfare from  $X$  is at least as great as that from  $Y$  for all increasing welfare functions, with strict inequality holding for some utility function(s) in the class. Equivalent conditions may be given in terms of quantiles. Loosely speaking, all quantiles of  $F(\cdot)$  will be at least as large as those of, and  $G(\cdot)$ . The distribution of  $X$  Second Order Stochastically Dominates  $Y$  (denoted as  $X$  SSD  $Y$ ) iff  $E[u(X)] \geq E[u(Y)]$  for all  $u \in \mathcal{U}_2$ , with strict inequality for some  $u$ . Equivalently,

$$\int_{-\infty}^z F(v)dv \leq \int_{-\infty}^z G(v)dv \quad \forall z \in \mathcal{Z}, \text{ with strict inequality for some } z. \quad (2)$$

If  $X$  SSD  $Y$ , then the expected social welfare from  $X$  is at least as great as that from  $Y$  for all increasing and concave utility functions in the class  $\mathcal{U}_2$ , with strict inequality holding for some utility function(s) in the class. FSD implies SSD and higher orders, and SSD is equivalent to Generalized Lorenz Dominance.

Define the following generalizations of the Kolmogorov-Smirnov test criteria:

$$d = \sqrt{\frac{NM}{N+M}} \min \sup_{z \in \mathcal{Z}} [F(z) - G(z)] \quad (3)$$

$$s = \sqrt{\frac{NM}{N+M}} \min \sup_{z \in \mathcal{Z}} \int_{-\infty}^z [F(u) - G(u)] du \quad (4)$$

where min is taken over  $F - G$  and  $G - F$ , in effect performing two tests in order to leave no ambiguity between the ‘equal’ and ‘unrankable’ cases. Our nonparametric tests for FSD and SSD are based on the empirical counterparts of  $d$  and  $s$  using the empirical CDFs, where the empirical CDF for  $X$  is given by

$$\hat{F}_N(x) = \frac{1}{N} \sum_{i=1}^N \mathbf{I}(X \leq x) \quad (5)$$

---

<sup>5</sup>Note that SD relations offer insights into which distribution provides greater welfare *considering only the outcome of interest*, without regard for other considerations such as cost. If the ‘treatment’ is costly, a separate cost-benefit analysis is required to decide if the welfare gains exceed the costs.

and  $I(\cdot)$  is an indicator function;  $\widehat{G}_M(y)$  is defined similarly for  $Y$ . If  $\widehat{d} \leq 0$  ( $\widehat{s} \leq 0$ ) to a degree of statistical confidence, then the null hypothesis of FSD (SSD) is not rejected (see Appendix A for details).

To this point  $X$  and  $Y$  have represented two *unconditional* test score variables. However, dependence between class size and other determinants of student achievement may confound the effect of class size with the impact of other characteristics (Hanushek 1979; Hanushek et al. 1996; Dustmann 2003). In particular, student achievement may be related to family background variables (such as family structure and income), teacher attributes (such as salary and experience), and school characteristics (such as school size and the ability of peers).<sup>6</sup> Concern over the ability to control for all such reasonable determinants of student achievement, however, has led researchers to focus on the various IV methods cited in the previous section. In particular, two potential sources of endogeneity of class size are frequently cited in the literature (Hoxby 2000b; Boozer and Rouse 2001; Ehrenberg et al. 2001; Dobbelsteen et al. 2003; Todd and Wolpin 2003). First, family attributes may be correlated with both student achievement and class size through endogenous residential decisions (so-called Tiebout (1956) sorting). Second, schools may have compensatory policies specifically designed to assign less able students to smaller classes and/or to superior teachers.

To control for the myriad of determinants of student performance, as well as circumvent the potential endogeneity issue, we follow in the spirit of Dearden et al. (2002) and perform dominance tests on conditional test score distributions derived via two approaches. Under the first approach, we control for a host of observable attributes that may generate a spurious correlation between class size and test scores and conduct dominance tests on *the distributions of test scores purged of the ‘average’ effects of these attributes*. Under the second approach, we incorporate differential ‘average’ effects by class size into the conditional distributions. In both cases, the conditioning covariates (discussed below) represent individual (such as race, gender, and lagged test scores), family (such as parental education, socioeconomic status, and family composition), class (such as subject and average student ability), teacher (such as race, gender, experience, and education), and school (such as enrollment, number of teachers, and average teacher salaries) attributes. Because our conditioning set is quite exhaustive, the implicit *selection on observables* assumption required to identify the ‘treatment’ effect on the distribution of test scores appears reasonable.

To proceed with the first approach, we estimate separate educational production function models for students in each size category, obtain the intercept–adjusted residuals, and perform the dominance tests on these residuals.<sup>7</sup> Specifically, in the first-stage, we estimate

---

<sup>6</sup>For a theoretical account of the cognitive development of students, see Todd and Wolpin (2003).

<sup>7</sup>The intercepts are included as part of the residuals, otherwise the conditional distributions will all be mean zero, precluding the possibility of first order dominance.

$$t_{ijk} = \alpha_k + h_{ij}\beta_k + \tilde{\epsilon}_{ijk}, \quad k = 1, \dots, K \quad (6)$$

where  $t_{ijk}$  is the test score for individual  $i$  in school  $j$  in class size group  $k$ ,  $h$  is a lengthy vector of individual, family, class, teacher, and school attributes,  $\tilde{\epsilon}$  is the error term, and there are  $K$  class size categories ( $K = 3$  in the application). In the second-stage, we analyze the distributions of  $\hat{\epsilon}_{ijk} \equiv \hat{\alpha}_k + \hat{\tilde{\epsilon}}_{ijk}$ , which correspond to test scores *net of all observable characteristics* (evaluated at the *size-specific* returns,  $\beta_k$ ).<sup>8</sup>

Under the above approach, the intercept-adjusted residuals,  $\hat{\epsilon}_{ijk}$ , reflect test scores net of all observable characteristics evaluated at the size-specific returns,  $\beta_k$ . Since this method nets out test score differences due to observables as well as the size-specific returns to such observables, we refer to these tests as being based on ‘Partial Residuals’ (PR). As an alternative, we implement a second approach based on the ‘Full Residuals’ (FR), where we denote the full residuals as inclusive of differences in the return to observables. Specifically, we re-write the first-stage regression (6) for class size  $k$  as

$$\begin{aligned} t_{ijk} &= \alpha_k + h_{ij}\beta_k + \tilde{\epsilon}_{ijk} \\ &= \alpha_k + h_{ij}\beta_k + \tilde{\epsilon}_{ijk} + (h_{ij}\beta_{k'} - h_{ij}\beta_{k'}) \\ &= \alpha_k + h_{ij}\beta_{k'} + h_{ij}(\beta_k - \beta_{k'}) + \tilde{\epsilon}_{ijk} \end{aligned} \quad (7)$$

where class size  $k'$  is implicitly treated as the ‘dominant’ category (Neuman and Oaxaca 2003). Consequently, we amend the residual tests to compare the previous intercept-adjusted residual distribution of  $\hat{\epsilon}_{ijk'}$  with  $\hat{\epsilon}_{ijk}^{FR} \equiv (\hat{\alpha}_k + h_{ij}(\hat{\beta}_k - \hat{\beta}_{k'}) + \hat{\tilde{\epsilon}}_{ijk})$ ,  $k \neq k'$ .

To aid in the comparison of the various residual dominance results, we also present the results from standard Oaxaca-Blinder parametric decompositions. Specifically, the mean test score gap between any two class sizes,  $k$  and  $k'$ , may be expressed as

$$\bar{t}_k - \bar{t}_{k'} = \underbrace{(\alpha_k - \alpha_{k'})}_U + \underbrace{(\bar{h}_k - \bar{h}_{k'})\beta_{k'}}_E + \underbrace{\bar{h}_k(\beta_k - \beta_{k'})}_C \quad (8)$$

where class  $k'$  is treated as the ‘dominant’ category. If the difference in returns (term  $C$ ) in (8) is large in absolute value, then the two residual tests may be expected to yield disparate results. Moreover, if the three sets of results (two sets of residual tests and the one set of unconditional tests) offer different inferences, one may infer a ‘significant’ association between class size, the distribution of test scores, the set of conditioning variables, and the returns to the conditioning variables.

---

<sup>8</sup>Note, controls for class size are omitted from (6), thereby allowing the error term to capture the residual effect of class size not captured by the included regressors.

### 3.2 Inference

The asymptotic null distribution of the test statistics,  $d$  and  $s$ , depend on the unknown distributions,  $F$  and  $G$ . In the analysis below, we first approximate the empirical distribution of the test statistics using *simple bootstrap* methods as in Maasoumi and Heshmati (2000) and Maasoumi and Millimet (2005), and report the estimated significance level. To evaluate the null  $H_o : d \leq 0$ , we first report in our tables whether the observed empirical distributions are *seemingly* rankable by FSD or SSD; we present the sample values of  $\max\{d_1\}$ ,  $\max\{d_2\}$ ,  $\hat{d}$ ,  $\max\{s_1\}$ ,  $\max\{s_2\}$ , and  $\hat{s}$  (see Appendix A). We then obtain bootstrap estimates of the probability that  $d$  lies in the non-positive interval (i.e.  $\Pr\{d \leq 0\}$ ) using the relative frequency of  $\{\hat{d}^* \leq 0\}$ , where  $\hat{d}^*$  is the bootstrap estimate of  $d$  (500 repetitions are used).<sup>9</sup> If this interval has a large probability, say 0.90 or higher, and  $\hat{d} \leq 0$ , we may infer dominance to a desirable degree of confidence. If this interval has a low probability, say 0.10 or smaller, and  $\hat{d} > 0$ , we may infer the presence of significant crossings of the empirical CDFs, implying an inability to rank the outcomes. Finally, if the probability lies in the intermediate range, say between 0.10 and 0.90, there is insufficient evidence to distinguish between equal and unrankable distributions. This is a classic confidence interval test; specifically, we are assessing the likelihood that the event  $d \leq 0$  has occurred. Similarly, we estimate  $\Pr\{s \leq 0\}$  to evaluate the second order dominance proposition given by  $H_o : s \leq 0$ .<sup>10</sup>

As an alternative, we also evaluate the less decisive dominance proposition  $H_o : d = 0$  via the Linton et al. (2005) *recentered bootstrap* procedure, which the authors demonstrate provides a consistent test. It is known that  $\hat{d}$  converges to  $d$  under general conditions (likewise for the SSD statistic). However, under the null  $H_o : d = 0$ , centering of computations around their corresponding sample values introduces second order errors that are negligible for first order (asymptotic) approximations, but is desirable for removing some uncertainties due to estimation of unknown parameters and distributions. This is the source of improvement in bootstrap power gained from recentering. The other source of improvement arising from recentering pertains to the technique’s robustness to within-sample dependence.

Utilizing the algorithm detailed in Appendix A, we obtain recentered bootstrap p-values in the classical

---

<sup>9</sup>Note, we also report simple bootstrap estimates of the  $\Pr\{d^* \geq \hat{d}\}$ . These are provided to facilitate visualization of the simple bootstrap distribution.

<sup>10</sup>Note, we do not impose and test the Least Favorable Case (LFC) of equality of the distributions. This could be done by combining the data on  $X$  and  $Y$  and bootstrapping from the combined sample (e.g., Abadie 2002). Our bootstrap samples still contain  $N$  ( $M$ ) observations from  $X$  ( $Y$ ). As argued in Linton et al. (2005), working under LFC has some undesirable power consequences as it can produce biased tests that are not similar on the boundary of the null. This happens when the boundary of the null itself is composite. The bootstrap methods employed herein, combined with a fixed critical value at zero (the boundary of the null hypothesis), renders our tests ‘asymptotically similar’ and unbiased on the boundary (Maasoumi and Heshmati 2005).

sense as the relative frequency of  $\{\widehat{d}^{**} > \widehat{d}\}$ , where  $\widehat{d}^{**}$  is the recentered bootstrap estimate of  $d$ . If the  $\Pr\{\widehat{d}^{**} > \widehat{d}\}$  is low, say 0.10 or smaller, we reject the null  $H_o : d = 0$ ; if this p-value is greater than 0.10, we fail to reject the null.<sup>11</sup> It is important to emphasize, however, that while rejection of the null provides valuable insight in the recentered bootstrap case, failure to reject the null provides less information. If we reject the null and  $\widehat{d} < 0$ , we may infer dominance to a desirable degree of confidence. Conversely, if we reject the null and  $\widehat{d} > 0$ , we may infer unequal, but unrankable, distributions. These are both strong findings, as the former (latter) indicates that all (not all) increasing social welfare functions will concur on the relative rankings of the distributions in question. On the other hand, failure to reject the null merely implies that we cannot eliminate the possibility that  $F = G$ ; strict dominance also cannot be ruled out to some degree of confidence. Seen in this light, the recentered bootstrap is a conservative test. In the discussion of the results, we focus more heavily on the more decisive simple bootstrap for inference. Similarly, we report the relative frequency of  $\{\widehat{s}^{**} > \widehat{s}\}$  and  $\{\widehat{s}^{**} > 0\}$  to evaluate the null  $H_o : s = 0$ .

A final, necessary comment pertains to inference in the FR tests (i.e., those incorporating the Oaxaca-Blinder decomposition). Due to the usage of a common set of coefficient estimates in obtaining both residual distributions being compared, there *necessarily* exists between-sample dependence. For example, the FR test using data on small ( $k = 1$ ) and medium ( $k = 2$ ) classes compares the distributions of  $\widehat{\epsilon}_{ij1}$  and  $\widehat{\epsilon}_{ij2}^{FR}$ . The former depends on  $\{t_{ij1}, h_{ij1}, \beta_1(t_1, h_1)\}$ , where  $t_1$  and  $h_1$  represent the full data vector for  $t$  and  $h$  for the sample of students in small classes; the latter,  $\widehat{\epsilon}_{ij2}^{FR} = t_{ij2} - h_{ij2}\beta_1$ , depends on  $\{t_{ij2}, h_{ij2}, \beta_1(t_1, h_1)\}$ . This source of dependence is atypical. Between-sample dependence usually arises when the same individuals appear in the two samples being compared (e.g., distributions of pre- and post-tax incomes for a sample of individuals). To handle this more common type of between-sample dependence, pairwise (or ‘clustered’) bootstrap samples are drawn in order to maintain the dependence in the resampled data (Linton et al. 2005). In the current situation, the between-sample dependence is maintained by re-estimating the first-stage equations (??) and (7) on each bootstrap resample. Specifically, by resampling  $N$  observations  $\{t_{ij1}^*, h_{ij1}^*\}$  and  $M$  observations  $\{t_{ij2}^*, h_{ij2}^*\}$  nonparametrically and re-estimating (??), we obtain the resampled distributions of  $\widehat{\epsilon}_{ij1}^*$  and  $\widehat{\epsilon}_{ij2}^{FR*}$ , where the former depends on  $\{t_{ij1}^*, h_{ij1}^*, \beta_1^*(t_1^*, h_1^*)\}$  and the latter depends on  $\{t_{ij2}^*, h_{ij2}^*, \beta_1^*(t_1^*, h_1^*)\}$ . Thus, as in the usual pairwise bootstrap case, the source of between-sample dependence is maintained in the resampling procedure.

---

<sup>11</sup>We also report the  $\Pr\{\widehat{d}^{**} \leq 0\}$  in the tables. This allows the reader to see the significance level (size) of the test associated with the special critical value ‘zero.’ In our tables, these are obtained simply as  $\Pr\{\widehat{d}^{**} > 0\} = 1 - \Pr\{\widehat{d}^{**} \leq 0\}$ .

## 4 Data

The data are obtained from the National Education Longitudinal Study (NELS) of 1988, a large-scale longitudinal study of high school students conducted by the National Center for Education Statistics (NCES). The NELS contains a nationally representative sample of eighth graders first surveyed in the spring of 1988. Follow-up surveys were administered to the respondents in 1990 and 1992. The original sample was chosen by initially sampling some 1,000 public and private schools from a universe of approximately 40,000 schools containing eighth grade students, and then drawing random samples of approximately 24-26 eighth grade students per school. The original sample, therefore, contains roughly 25,000 eighth grade students.

The students were administered cognitive tests in reading, social studies, mathematics and science during the base year (1988), first follow-up (1990), and second follow-up (1992). The grade-specific tests contained material appropriate for each grade, but included sufficient overlap with the exams for the other grades to permit measurement of academic growth.<sup>12</sup> For each sampled student, the teachers from two of the four subjects were surveyed, thereby yielding information on class size. Thus, we have two observations for each student in each wave of the survey after we match the class size with the test score for that particular subject. The NELS also supplies general descriptive information about the school obtained from the chief administrator of each school in the sample.

Following Boozer and Rouse (2001), we construct two samples, each including only students attending public schools. The first sample uses test score results from the first follow-up (tenth grade) and information on eighth grade class size, and contains 12,412 students from 762 schools (23,549 total observations).<sup>13</sup> The second sample is analogously defined using information from tenth and twelfth grades, and contains 8,685 students from 756 schools (14,796 total observations), considerably smaller than the previous sample due to attrition (either dropping out of the survey or relocating to a out-of-sample school). All results are weighted using the appropriate sample weights.

Because there is a continuum of class sizes to which students belong, we classify each student-test observation in each sample into one of three groups to make the number of SD tests manageable. The groupings are: *small* (19 or fewer students), *medium* (20-30 students), and *large* (more than 30 students).

To obtain the residual test scores, we utilize an extensive set of individual, family, class, teacher, and school characteristics available from the NELS. Specifically, the vector  $h$  in (6) includes controls for the

---

<sup>12</sup>We follow Goldhaber and Brewer (1997) and Boozer and Rouse (2001) and utilize the raw item response theory (IRT) scores for each test.

<sup>13</sup>The exams may be administered as early as January, mid-way through the school year. As a result, we follow the lead of Boozer and Rouse (2001) and initially examine the effect of previous class size on current test scores. See also (Hoxby 2000a).

following (in addition to a constant term):

**Individual:** race, gender, an indicator for limited English proficiency (LEP), and lagged test scores;

**Family:** father’s education, mother’s education, family composition, number of siblings, an indicator for the presence of a home computer, family socioeconomic status, and dummy variables indicating if family composition and number of siblings are missing;

**Class:** subject and the overall relative ability of the class;

**Teacher:** race, gender, experience, education, indicators if the student and teacher are of the same race and of the same gender;

**School:** urban/rural status, region, total school enrollment, grade-level enrollment, number of full-time teachers, number of teachers by race, level of student disruptions in class, percentage of minority students in the school, teacher salaries, length of school year, percentage of students in remedial reading, remedial math, and bilingual education, and dummy variables indicating if teacher salary information, school year length, the number of full-time teachers by race, and the percentage of students in remedial reading, remedial math, and bilingual education are missing.

Summary statistics for the samples are available upon request.

Before continuing, a few comments are warranted. First, the inclusion of lagged test scores proxies for innate ability, following the strategy of Eide and Showalter (1998), Dearden et al. (2002), and others. Lagged test scores also control for all previous inputs into the educational production process, giving the results a ‘value-added’ interpretation (Goldhaber and Brewer 1997; Rivkin et al. 2002; Todd and Wolpin 2003). In addition, we also condition on the teacher’s subjective assessment of the overall ability of the class from which the test score is taken.<sup>14</sup> Such ability controls are vital to circumventing the potential endogeneity arising from endogenous residential choice and nonrandomness in the assignment of students to classes by schools (Betts and Shkolnik 2000). The ability level of the class also captures peer effects which have been shown to be important (Levin 2001; Dobbelsteen et al. 2003). Second, to further minimize any potential spurious correlation between class size and other determinants of student achievement, we condition on a fairly substantial vector of school attributes. Since *actual* class size is only observed *ex*

---

<sup>14</sup>The response options are (i) the class is of “above average” ability (relative to the school), (ii) the class is of “average” ability, (iii) the class is of “below average” ability, or (iv) the class is comprised of students of “widely varying” ability. Figlio and Paige (2002) also make use of this variable.

*post*, while school attributes are observable *ex ante*, controlling for school characteristics not only removes their direct effect on student performance, but also proxies for family background traits and parental involvement.<sup>15</sup> Third, as argued in Hanushek (1979), controlling for family attributes such as socioeconomic status and parental education levels also severely mitigates any bias resulting from endogenous residential choice.

## 5 Results

### 5.1 Unconditional SD Tests

The initial SD tests involve comparing the unconditional test score distributions across students differentiated by class size. Results are provided in Table 1; Panel A examines tenth grade test scores as a function of eighth grade class size, and Panel B examines twelfth grade test scores as a function of tenth grade class size. The corresponding CDFs, integrated CDFs, and differences in the CDFs are plotted in Figures 1 and 2, respectively.

Comparing the empirical distributions of tenth grade test scores across small, medium, and large classes (Panel A), we find no instance where we are able to rank the distributions in either the first- or second-degree sense, despite the fact that a clear ordering exists for mean test scores.<sup>16</sup> Moreover, the simple bootstrap indicates that the crossings of the CDFs in each case are statistically meaningful at the  $p < 0.01$  confidence level. The recentered bootstrap confirms this finding, indicating rejection of the null  $H_o : d = 0$  when comparing small and medium classes (p-value = 0.000), rejecting strict dominance and equality of the distributions; the null is nearly rejected when comparing small and large classes as well (p-value = 0.126). The recentered bootstrap also nearly rejects the null  $H_o : s = 0$  when comparing medium and large classes (p-value = 0.114), rejecting strict second order dominance and equality of the distributions.

The lack of first- or second-degree SD is an extremely powerful result. For example, if the test score level for Small FSD Medium (or Large), then any policymaker with a social welfare function increasing in test scores would prefer smaller class sizes. Similarly, a finding of SSD would imply that any policymaker with a social welfare function that is increasing and averse to dispersion in test scores would prefer smaller class sizes. However, we find no such dominance relations; individuals with different preference functions in the class  $\mathcal{U}_1$  or  $\mathcal{U}_2$  can reasonably disagree about the efficacy of smaller classes.

---

<sup>15</sup>For instance, when looking at homes for sale at <http://www.realtor.com>, links are provided for (virtually) every house in the US in order to view the average pupil-teacher ratio and total enrollment for that location (along with average SAT scores and percentage of students continuing to college). See also Todd and Wolpin (2003).

<sup>16</sup>The average test score is highest in small classes (31.24), followed by large classes (31.21), and then medium classes (31.17).

Examining the actual plots (Figure 1), we see that the three CDFs and integrated CDFs are extremely similar, never differing by more than four test points in any part of the distribution.<sup>17</sup> Nonetheless, the plot of the differences in the CDFs – corresponding to estimates of the QTEs – is still revealing, indicating that medium and large classes are most similar, while small classes are quite distinct. Specifically, small classes outperform medium and large classes at the upper end of the distribution (above the 70<sup>th</sup> percentile), with the converse holding below the 70<sup>th</sup> percentile.

Turning to the examination of twelfth grade test scores (Panel B) reveals several instances where the empirical distributions of test scores are rankable, and in all such cases it is the distribution from the larger class size that dominates the empirical distribution from the smaller class size.<sup>18</sup> However, not all of these rankings are statistically meaningful at conventional levels, highlighting the need for formal statistical testing. Specifically, we observe Medium SSD Small, Large SSD Small, and Large FSD Medium. The simple bootstrap yields a marginally significant  $\Pr(s \leq 0) = 0.882$  and  $0.870$  in the first two cases, and a  $\Pr(d \leq 0) = 0.134$  ( $\Pr(s \leq 0) = 0.408$ ) in the final case. Moreover, the recentered bootstrap fails to reject the null  $H_o : s = 0$  in any of the three cases; it does reject the null  $H_o : d = 0$  in the test of small versus medium classes (p-value = 0.000), thereby rejecting strict dominance and equality of the distributions. Thus, there is at best modest evidence that of uniform rankings favoring larger classes when (i) unconditional test scores and (ii) the dispersion of test scores are considered.

Examining the actual plots (Figure 2), we see that – as in Figure 1 – the three CDFs and integrated CDFs are extremely similar, never differing by more than four points in any part of the distribution.<sup>19</sup> However, as above, the plot of the differences in the CDFs (i.e., the QTEs) is very informative, indicating that large classes are preferable to medium classes over the entire distribution (consonant with the observed FSD ranking), and that the gains from large classes (relative to medium classes) is fairly uniform across the distribution. Furthermore, as was the case with eighth grade class size, small classes outperform medium and large classes at the upper end of the distribution (roughly above the 70<sup>th</sup> percentile), with the reverse holding at the lower end.

In sum, the unconditional tests assessing the impact of eighth grade class size refute the existence of a uniform ranking – a ranking robust to the choice of specific preference function – of test score distributions across class size categories. This finding highlights the false sense of decisiveness that one gets from summary comparisons, such as those based on mean ‘treatment’ effects. The unconditional tests based on

---

<sup>17</sup>The standard deviation of tenth grade test scores is 12.08.

<sup>18</sup>Focusing on the mean test scores by class size grouping also favors larger classes. The average twelfth grade test score is 36.6 in large classes, 35.5 in medium classes, and 35.0 in small classes.

<sup>19</sup>The standard deviation of twelfth grade test scores is 13.60.

tenth grade class size, however, provide modest evidence of a uniform ranking of test score distributions across class size categories. However, such rankings – to the extent that they exist – are only obtained when one incorporates dispersion of test scores into the welfare criteria. Moreover, the distributional approach reveals exactly who gains and who loses (in the unconditional sense) from smaller classes: students in the upper (lower) tail of the distribution gain from smaller (larger) classes. These findings are merely suggestive, however, as they fail to control for observables correlated with both class size and student achievement. For instance, if low-achieving students are allocated to smaller classes, this may explain the lower test scores found in smaller classes below the median. To determine if the unconditional results hold once we purge test scores of such potential confounders, we turn to the conditional SD tests.

## 5.2 Conditional SD Tests

### 5.2.1 Tenth Grade Test Scores

The PR and FR dominance test results for the pairwise comparisons involving tenth grade test scores are displayed in Table 2.<sup>20</sup> The corresponding plots are displayed in Figures 3-5. Examining the PR results (Panel A), we find that in every pairwise comparison, the empirical distribution for larger classes first order dominates the corresponding distributions for smaller and medium classes. Moreover, both rankings are statistically significant at conventional levels according to the simple bootstrap (Large FSD Small:  $\Pr(d \leq 0) = 0.986$ ; Large FSD Medium:  $\Pr(d \leq 0) = 0.924$ ). The recentered bootstrap fails to reject the null  $H_0 : d = 0$  or  $s = 0$  (Large FSD Small: p-value = 0.922; Large FSD Medium: p-value = 0.874). In addition, we observe that Medium FSD Small, although this ranking is only marginally significant according to the simple bootstrap (simple:  $\Pr(d \leq 0) = 0.882$ ); the recentered bootstrap fails to reject the null of equality (p-value = 0.604). However, the second order relation is statistically significant at conventional levels according to the simple bootstrap ( $\Pr(s \leq 0) = 0.938$ ). Examination of Figure 3 reveals that disparities in PR test scores (the QTEs) are fairly uniform across the entire distribution: at any quantile of the distribution, students in large (medium) classes score roughly ten (two) points higher than students in small classes.

These striking results indicate that not only do smaller classes not map into improvements in the *subsequent* distribution of test scores, net of a host of individual, family, class, teacher, and school attributes, but they actually result in inferior student performance according to the PR tests. Furthermore, this finding is robust across *any* social welfare function that is increasing in test scores. Finally, the fact that the corresponding unconditional distributions are unrankable (Table 1, Panel A) indicates that observable

---

<sup>20</sup>First-stage regression results are not presented, but are available from the authors upon request.

attributes that are positively associated with test scores are negatively correlated with class size, consonant with positive selection into smaller classes through Tiebout sorting.

As noted previously, however, a potential shortcoming of the PR tests is that differences in the class size-specific returns to observables are not included in the PR distributions. As a result, any differences in the returns to observable characteristics are not attributable to the effects of class size, which may yield misleading inferences. The results of FR dominance tests, which account for such impacts of class size, are displayed in Panel B.<sup>21</sup> The results indicate that the empirical distributions are rankable for two of the three comparisons: Small SSD Medium and Medium SSD Large.<sup>22</sup> However, neither ranking is statistically significant according to the simple bootstrap. In the former (latter) case, the simple bootstrap yields  $\Pr(s \leq 0) = 0.526(0.586)$ . The recentered bootstrap gives p-values above 0.70 in both cases, failing to reject the null  $H_0 : s = 0$ . In terms of the comparison of small and large classes, the FR distributions are unrankable.<sup>23</sup>

Examination of Figures 4 and 5 yield three additional findings. First, while small classes second order dominate medium classes, but not large classes, according to the empirical distributions, there is little difference in actuality between the observed distributions for medium and large classes (when small classes are the ‘dominant’ group). Second, small classes fare better than medium and large classes at the lower tail of the distribution (below the median), while the converse holds above the median. This fact highlights the concavity assumption explicitly required of the social welfare function under the SSD ranking. Finally, medium classes outperform large classes (when medium classes are the ‘dominant’ group) across the majority of the distribution; if not for a few crossings in the extreme tails, one would have observed a FSD ranking. However, the difference is of low magnitude as test scores differ by less than half a point (in absolute value) over the majority of the distribution.

These results – which suggest a modest, but statistically insignificant, monotonic ranking in favor of smaller classes – reverse the findings from the PR tests and indicate the importance of allowing the returns to observable attributes to vary by class size, as well as incorporating these differential returns into the

---

<sup>21</sup>In the tables of FR results, the distribution in the ‘X’ column is treated as the ‘dominant’ category; see equation (7).

<sup>22</sup>While the CDFs for small and medium classes appear to cross in the extreme lower tail in the top row of Figure 3 (favoring medium classes and thereby precluding an SSD ranking in favor of small classes), Table 2 reports a finding of Small SSD Medium because of a ‘trimming’ procedure utilized. Specifically, in the empirical implementation, the support points used to obtain the test statistics (see Appendix A) are chosen to be equally-spaced, beginning at the first percentile and ending at the 99<sup>th</sup> percentile of the empirical support,  $\mathcal{Z}$ . This process focuses attention away from extreme outliers.

<sup>23</sup>Note that SSD relations, especially using the FR distributions, do not obey a transitivity property; Small SSD Medium and Medium SSD Large is not sufficient to guarantee Small SSD Large. While this is true in general for SSD (not FSD), it is especially true in the FR tests since the Small versus Medium and Small versus Large comparisons utilize Small as the ‘dominant’ category, while the Medium versus Large comparison uses Medium as the ‘dominant’ group.

residual distributions being compared. To verify that this indeed the case, Table 3 presents the results from standard, parametric Oaxaca-Blinder decompositions. Consonant with the change in results from the PR to the FR tests, Panel A indicates that the coefficients differ considerably across the three class size groups, with a significant advantage belonging to small classes (followed by medium classes). In particular, these results are driven primarily by greater returns to teacher race and education, school size, and lagged test scores. Thus, an important benefit of smaller classes (in terms of subsequent test scores) appears to be a greater return to other observable attributes such as teacher education and students' innate ability and/or previous educational inputs.

In the end, we believe the FR tests to best isolate the effects of class size on the distribution of test scores, and the simple bootstrap to be the more informative method of inference. Accordingly, we conclude that the unconditional distributions of test scores favor students in classes with 20 or more students below the 70<sup>th</sup> percentile, and students in classes with fewer than 20 students in the upper tail. However, after purging test scores of the average impact of a host of observable attributes using class size-specific returns to observables and invoking the selection on observables assumption, we find that test scores exhibit a monotonic second order SD relationship, albeit statistically insignificant at conventional levels, favoring smaller classes. Moreover, the advantage for small classes is confined solely to below the median and is mainly attributable to the advantageous returns to observables enjoyed by students in smaller classes. As a result, uniform rankings are only possible when the debate over class size is broadened to incorporate the dispersion of test scores into the evaluation of educational policy. Furthermore, these results suggest the possibility of a Pareto-improving reallocation of students, whereby lower-performing students are placed in smaller classes and higher-performing students are placed in larger classes.

### 5.2.2 Twelfth Grade Test Scores

The PR and FR results for twelfth grade test scores are given in Panels C-D of Table 2, respectively. The corresponding plots are displayed in Figures 6-8. The PR results suggest a non-monotonic effect of class size. Specifically, we observe Small FSD Medium, Large FSD Small, and Large FSD Medium. However, all three observed rankings are statistically insignificant at conventional levels according to the simple bootstrap. A second order ranking of Large over Small is statistically significant ( $\Pr(s \leq 0) = 0.934$ ), while the other two comparisons are only marginally significant (Small SSD Medium:  $\Pr(s \leq 0) = 0.846$ ; Large SSD Medium:  $\Pr(s \leq 0) = 0.890$ ). Examination of Figure 6 confirms the previous results based on eighth grade class size (Figure 3): disparities in PR test scores (i.e., the QTEs) are fairly uniform across the entire distribution. Specifically, at any quantile of the distribution, students in large (small) classes score roughly two (one-quarter) points higher on twelfth grade tests than students in medium classes.

As documented in the eighth grade class size results, however, the PR tests ignores a potentially important difference across class sizes: the returns to observable determinants of student achievement. The results of the FR dominance tests, displayed in Panel D, reveal that the three distributions are unrankable in the first- and second-degree sense. Examination of Figures 7 and 8 yield two additional findings. First, small classes fare better than medium and large classes only in the extreme lower tail of the distribution (roughly below the 15<sup>th</sup> percentile) when small classes are treated as the ‘dominant’ group. Second, large classes outperform medium classes across the majority of the distribution, although the difference is of low magnitude as test scores differ by less than one-half point (in absolute value) over the majority of the distribution. This result is invariant to the choice of small or medium classes as the ‘dominant’ group.

These results – which suggest little impact of class size on subsequent test scores– refute the findings from the PR tests and confirm our previous finding that the returns to observable attributes vary in an important manner by class size. Specifically, Panel B of Table 3 verifies that, according to the standard parametric Oaxaca-Blinder decomposition, the return to observables is most favorable to medium classes, followed by small and then large classes. In particular, these results are driven by differences in the returns to teacher education and salary, teacher and student race, number of full-time teachers, and lagged test scores.

As stated previously, we believe the FR tests to best isolate the effects of class size on the distribution of test scores, and the simple bootstrap to be the more informative method of inference. Accordingly, the analysis using tenth grade test scores largely reaffirms the conclusions we drew from the analysis of tenth grade test scores. Specifically, we conclude that the unconditional distribution of test scores favors smaller classes roughly only above the 70<sup>th</sup> percentile. However, after purging test scores of a host of observable attributes using class size-specific returns, we conclude that the distribution of subsequent test scores are largely unaffected by class size, although students in the extreme lower tail are, if anything, aided by larger classes.

### 5.3 Disaggregation of ‘Small’ Classes

As noted in Ehrenberg et al. (2001), target class size varies across states, with some states targeting classes with 15 students and others targeting classes of 20 students. In addition, the few experiments that have been conducted in the US typically involve changes in class size at the lower end of the class size distribution. For example, Tennessee’s Project STAR compared classes of 13-17 students with classes of 22-26 students. Wisconsin’s SAGE (Student Achievement Guarantee in Education) program reduced many classes from between 21-25 students to only 12-15 students. California’s CSRFP reduced classes from roughly 30 to 20 students. To see if marginal changes in this range matter, we define two new class size

groupings: Small I (10-15 students) and Small II (16-19 students).

### 5.3.1 Tenth Grade Test Scores

The results pertaining to the tenth grade test score distributions are displayed in Panels A-C of Table 4. The CDFs, integrated CDFs, and differences in the CDFs are given in Figures 9-11. In terms of the unconditional empirical distribution (Panel A), we observe Small II SSD Small I. The observed dominance relation favoring the relatively larger class size is consonant with the previous unconditional tests examining tenth grade class size (Table 1). However, the second order ranking is not statistically significant at conventional levels according to the simple bootstrap ( $\Pr(s \leq 0) = 0.696$ ). The recentered bootstrap fails to reject the null  $H_o : s = 0$  (p-value = 0.926).

Examining the actual plots (Figure 9), we see that the Small II distribution lies to the right of the Small I distributions over the majority of the support. In fact, if not for a few crossings, we would have observed a first order ranking. Moreover, it is interesting to note that the unconditional test score gap favoring Small II classes gets wider in the upper tail (i.e., the QTEs are increasing across the quantiles). This contrasts with previous unconditional eighth grade class size results in Figure 1, where (the previously defined) small classes fared better than medium and large classes only in the upper tail.

The PR test result is displayed in Panel B. Now, we observe Small I FSD Small II. Given the reversal in ranking from the unconditional comparison, this implies that attributes associated with higher test scores are positively correlated with class size over the range being analyzed, consistent with compensatory policies in this range by schools. Moreover, the first order ranking is statistically significant at conventional levels according to the simple bootstrap ( $\Pr(d \leq 0) = 0.916$ ). The recentered bootstrap fails to reject the null  $H_o : d = 0$  or  $s = 0$  (p-values above 0.46). Examination of Figure 10 reveals sizeable disparities in PR test scores that are fairly uniform across the entire distribution: at any quantile of the distribution, students with 10-15 students score roughly four points higher than students in classes with 16-19 students.

The results of the FR dominance tests are displayed in Panel C. Examination of the results reveals that the first order ranking found in the PR test disappears. The elimination of the previous FSD ranking implies that the returns to observables favors classes with 16-20 students. Indeed, this is confirmed in Panel A of Table 5, where the returns on the whole favor Small II classes. In particular, the returns to teacher education and salary, as well as the number of white teachers, yield the primary discrepancies in favor of classes with 16-19 students. Thus, the improvement in returns as class size diminishes documented in the previous section (Panel A, Table 3) is not monotonic.

Viewing the actual plots in Figure 11 shows that the disparity in test scores favors classes with 10-15 students in the lower tail of the distribution (below the 30<sup>th</sup> percentile). However, since the area between

the CDFs (i.e., the cumulative sum (in absolute terms) of the QTEs) is greater above the 30<sup>th</sup> percentile, this precludes a finding of SSD in favor of classes with 10-15 students. In the end, then, given our stated preference for the FR tests, disaggregation of eighth grade small classes indicates that reduction in class size below 16 students only aids the lowest-scoring of students and does not enjoy unambiguous support by all preference functions in the class  $\mathcal{U}_2$ .

### 5.3.2 Twelfth Grade Test Scores

The results analyzing the twelfth grade test score distributions are given in Panels D-F of Table 4. The CDFs, integrated CDFs, and differences in the CDFs are given in Figures 12-14. In terms of the unconditional distributions, we are unable to rank the distributions in either the first- or second-degree sense. Noteworthy, however, is the fact that the recentered bootstrap rejects the null  $H_o : d = 0$  (p-value = 0.056), rejecting both strict dominance and equality of the distributions. Examining Figure 12, we see that – consonant with the earlier results comparing small, medium, and large classes – the relatively large Small II classes outperform Small I classes up to approximately the 70<sup>th</sup> percentile. Thus, as is the case for eighth grade class size, there is modest evidence of an unconditional advantage for classes with 16-19 students over those with 10-15 students, except for the highest-performing students.

The PR dominance test results are displayed in Panel E. As with tenth grade test scores, we find that Small I FSD Small II. This continues to imply a positive correlation between attributes improving test scores and class size for students in classes with fewer than 20 students. However, unlike the previous results for eighth grade class size, this ranking is only marginally statistically significant according to the simple bootstrap ( $\Pr(d \leq 0) = 0.882$ ); the second order ranking is statistically significant ( $\Pr(s \leq 0) = 0.940$ ). The recentered bootstrap fails to reject the null  $H_o : d = 0$  or  $s = 0$ .

Examination of Figure 13 reveals a sizeable, fairly uniform disparity in PR test scores: at any quantile of the distribution, students in tenth grade classes with 10-15 students score roughly three to four points higher than students in classes with 16-19 students. However, the distributions come close to crossing in the extreme lower tail, and apparently do cross sufficiently frequently in the bootstrap resamples to preclude a stronger statistically significant FSD ranking.

Lastly, the FR test results are displayed in Panel F. As with in the previous section examining tenth grade test scores, we are unable to rank the empirical distributions. Moreover, the recentered bootstrap rejects the null  $H_o : s = 0$  (p-value = 0.034), rejecting both strict second order dominance and equality of the distributions. The elimination of the FSD ranking obtained in Panel E implies that the returns to observables favors classes with 16-19 students. Panel B in Table 5 confirms this.<sup>24</sup> Figure 14 shows

---

<sup>24</sup>Some of the returns that differed the most across class size groupings are for size of the tenth grade student body, student

further consistency with the previous tenth grade test score results: the disparity in test scores favors classes with 10-15 students in the lower tail of the distribution (below the 40<sup>th</sup> percentile). However, since the area between the CDFs (i.e., the cumulative sum (in absolute terms) of the QTEs) is greater above the 40<sup>th</sup> percentile, this precludes a finding of SSD in favor of classes with 10-15 students. Consequently, as stated above, preferences for classes of 10-15 or 16-19 students may differ amongst individuals with different welfare functions in the class  $\mathcal{U}_2$ .

## 6 Conclusion

Identifying the factors most relevant to student achievement is important to parents, as well as policymakers concerned with maximizing the use of funds available for public schools. However, the impact of improved student achievement is potentially even more far-reaching, extending beyond school walls through its effect on completed education, future earnings, racial disparities, and economic competitiveness. While the factor most consistently discussed is class size, given the ease at which it may be manipulated by policy, previous empirical examinations of the impact of class size reductions have been mixed, with the results suggesting at best a modest impact. Despite this lack of convincing support, the US federal government allocated \$12 billion (over a seven-year period) to reduce class sizes (Hoxby 2000a), the state of California has spent over \$3.6 billion on class size reduction since 1996, 20 US states are currently undertaking or discussing policies to reduce class sizes, and the Dutch government decided to allocate approximately \$500 million (in US dollars) to reduce class sizes (Levin 2001).

To investigate these previous mixed results, we analyze the relationship between class size and student achievement using recently developed tests for stochastic dominance. The tests are nonparametric and utilize information on the entire distribution of test scores. Moreover, through the use of bootstrap techniques, we are able to report the results of the dominance tests to a degree of statistical certainty. Thus, a finding of a first or second degree dominance is extremely powerful, implying that any social welfare function that is increasing in test score levels (FSD) or increasing and concave (SSD) will prefer one distribution over another. This type of analysis is very useful for policy decisions as it lends itself to broad-based, consensus ranking of outcomes. Furthermore, the absence of such rankings are equally informative, implying that any judgement of one distribution over another is subjective.

Using data from the National Education Longitudinal Study of 1988, we estimate the effects of eighth and tenth grade class size on the unconditional and conditional distributions of test scores in reading, social studies, mathematics, and science. The results shed some light on the previous ambiguous findings,

---

race, the number of white teachers, and teacher experience.

and should prove quite informative for in future educational policy discussions. First, we find that the *unconditional* distributions of test scores favor students in *larger* classes below the 70<sup>th</sup> percentile or so, and students in *smaller* classes in the upper tail. In particular, while marginal changes in class size in classes with at least 20 students are not associated with changes in the unconditional test score distributions, the unconditional test scores of low-achieving (high-achieving) students are *superior* in classes with at least (less than) 20 students. Second, analyzing the conditional distribution of test scores, using *class-size specific returns*, we find little causal effect of marginal reductions in class size on test scores within the range of 20 or more students. However, reductions in class size from above 20 students to below 20 students, as well as marginal reductions in classes with fewer than 20 students, *increase* test scores for students *below the median*, but *decrease* test scores for students *above the median*.

Our methodology and findings point to several possible culprits when trying to make sense of the current empirical evidence on class size. First, reducing class size by one student does not have a constant effect: reductions in classes above 20 students have essentially no impact, while reductions in classes with 20 or fewer students *raise* the test scores of some students (those below the median) and *lower* the test scores of other students (those above the median). Second, whereas the majority of the beneficial impact of class size reduction arises because of the productivity-enhancing effect it has on other educational inputs, the majority of previous studies allow, at most, the impact of class size to vary by race or gender.

## References

- [1] Abadie, A. (2002), "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97, 284-292.
- [2] Akerhielm, K. (1995), "Does Class Size Matter?," *Economics of Education Review*, 14, 229-241.
- [3] Amin, S., A.S. Rai, and G. Topa (2003), "Does Microcredit Reach the Poor and Vulnerable? Evidence From Northern Bangladesh," *Journal of Development Economics*, 70, 59-82.
- [4] Anderson, G. (1996), "Nonparametric Tests of Stochastic Dominance in Income Distributions," *Econometrica*, 64, 1183-1193.
- [5] Angrist, J. and V. Lavy (1999), "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114, 533-575.
- [6] Barrett, G. and S. Donald (2003), "Consistent Tests for Stochastic Dominance," *Econometrica*, 71, 71-104.
- [7] Barro, R.J. (2001), "Human Capital and Growth," *American Economic Review*, 91, 12-17.
- [8] Betts, J.R. and J.L. Shkolnik (2000), "The Effects of Ability Grouping on Student Math Achievement and Resource Allocation in Secondary Schools," *Economics of Education Review*, 19, 1-15.
- [9] Bishop, J.A., J.P. Formby, and L.A. Zeager (2000), "The Effect of Food Stamp Cashout on Undernutrition," *Economics Letters*, 67, 75-85.
- [10] Bitler, M.P., J.B. Gelbach, and H.H. Hoynes (2005), "Distributional Impacts of the Self-Sufficiency Project," NBER Working Paper No. 11626.
- [11] Boozer, M.A. and C. Rouse (2001), "Intraschool Variation in Class Size: Patterns and Implications," *Journal of Urban Economics*, 50, 163-189.
- [12] Card, D. and A.B. Krueger (1996), "School Resources and Student Outcomes: An Overview of the Literature and New Evidence from North and South Carolina," *Journal of Economics Perspectives*, 10, 31-50.
- [13] Chubb, J.E. and T.M. Moe (1990), *Politics, Markets and America's Schools*, Washington, DC: Brookings Institution Press.

- [14] Coleman, J.S., and others (1966), *Equality of Educational Opportunity*,. Washington, DC: Department of Health Education, and Welfare.
- [15] Cook, M.D. and W.E. Evans (2000), “Families or Schools? Explaining the Convergence in White and Black Academic Performance,” *Journal of Labor Economics*, 18, 729-754.
- [16] Dearden, L., J. Ferri, and C. Meghir (2002), “The Effect of School Quality on Educational Attainment and Wages,” *Review of Economics and Statistics*, 84, 1-20.
- [17] Dobbelsteen, S., J. Levin, and H. Oosterbeek (2002), “The Causal Effect of Class Size on Scholastic Achievement: Distinguishing the Pure Class Size Effect from the Effect of Changes in Class Composition,” *Oxford Bulletin of Economics and Statistics*, 64, 17-38.
- [18] Dustmann, C. (2003), “The Class Size Debate and Educational Mechanisms: Editorial,” *Economic Journal*, 113, F1-F2.
- [19] Ehrenberg, R.G., D.J. Brewer, A. Gamoran, and J.D. Willms (2001), “Class Size and Student Achievement,” *Psychological Science in the Public Interest*, 2, 1-30.
- [20] Eide, E. and M.H. Showalter (1998), “The Effect of School Quality on Student Performance: A Quantile Regression Approach,” *Economics Letters*, 58, 345-350.
- [21] Fertig, M. (2003a), “Who’s to Blame? The Determinants of German Students’ Achievement in the PISA 2000 Study,” IZA Discussion Paper 739.
- [22] Fertig, M. (2003b), “Educational Production, Endogenous Peer Group Formation and Class Composition - Evidence from the PISA 2000 Study,” IZA Discussion Paper 714.
- [23] Figlio, D.N. and M.E. Paige (2002), “School Choice and the Distributional Effects of Ability Tracking: Does Separation Increase Inequality?” *Journal of Urban Economics*, 51, 497-514.
- [24] Fisher, G., D. Wilson, and K. Xu (1998), “An Empirical Analysis of Term Premiums Using Significance Tests for Stochastic Dominance,” *Economics Letters*, 60, 195-203.
- [25] Goldhaber, D. and D. Brewer (1997), “Why Don’t Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity,” *Journal of Human Resources*, 32m 505-523.
- [26] Hanushek, E.A. (1979), “Conceptual and Empirical Issues in the Estimation of Educational Production Functions,” *Journal of Human Resources*, 14, 351-388.

- [27] Hanushek, E.A. (1986), "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24, 1141-1177.
- [28] Hanushek, E.A. (1989), "The Impact of Differential Expenditures on School Performance," *Educational Researcher*, 18, 45-51.
- [29] Hanushek, E.A. (1996), "School Resources and Student Performance," in G. Burtless (ed.) *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*, Washington, D.C.: Brookings Institution.
- [30] Hanushek, E.A. (2001), "Black-White Achievement Differences and Government Interventions," *American Economic Review*, 91, 24-28.
- [31] Hanushek, E.A. (2002), "Evidence, Politics, and the Class Size Debate," in L. Mishel and R. Rothstein (ed.) *The Class Size Debate*, Washington, DC: Economic Policy Institute, 37-65.
- [32] Hanushek, E.A. (2003), "The Failure of Input-Based Schooling Policies," *Economic Journal*, 113, F64-F98.
- [33] Hanushek, E.A. and D.D. Kimko (2000), "Schooling, Labor-Force Quality, and the Growth of Nations," *American Economic Review*, 90, 1184-1208.
- [34] Hanushek, E.A., S.G. Rivkin, and L.L. Taylor (1996), "Aggregation and Estimated Effects of School Resources," *Review of Economics and Statistics*, 78, 611-627.
- [35] Hoxby, C.M. (2000a), "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics*, 90, 1239-1285.
- [36] Hoxby, C.M. (2000b), "Does Competition Among Public Schools Benefit Students and Taxpayers?," *American Economic Review*, 90, 1209-1238.
- [37] Jepsen, C. and S. Rivkin (2002), "What is the Tradeoff Between Smaller Classes and Teacher Quality?," NBER Working Paper 9205.
- [38] Kaur, A., B.L.S. Prakasa Rao, and H. Singh (1994), "Testing for Second-Order Stochastic Dominance of Two Distributions," *Econometric Theory*, 10, 849-866.
- [39] Klecan, L., R. McFadden, and D. McFadden (1991), "A Robust Test for Stochastic Dominance," unpublished manuscript, Department of Economics, MIT.

- [40] Krueger, A. B. (1999), "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, 114, 497-532.
- [41] Krueger, A.B. (2003), "Economic Considerations and Class Size," *Economic Journal*, 113, F34-F63.
- [42] Krueger, A.B. and D.M. Whitmore (2002), "Would Smaller Classes Help Close the Black-White Achievement Gap?," in J. Chubb and T. Loveless (eds.) *Bridging the Achievement Gap*, Washington, D.C.:Brookings Institute Press.
- [43] Lazear, E.P. (2001), "Educational Production Function," *Quarterly Journal of Economics*, 116, 777-801.
- [44] Levin, J. (2001), "For Whom the Reductions Count: A Quantile Regression Analysis of Class Size and Peer Effects on Scholastic Achievement," *Empirical Economics*, 26, 221-246.
- [45] Linton, O., E. Maasoumi, and Y.J. Whang (2005), "Consistent Testing for Stochastic Dominance: A Subsampling Approach," *Review of Economic Studies*, 72, 735-765.
- [46] Maasoumi, E. and A. Heshmati (2000), "Stochastic Dominance Amongst Swedish Income Distributions," *Econometric Reviews*, 19, 287-320.
- [47] Maasoumi, E. and A. Heshmati (2005), "Evaluating Dominance Ranking of PSID Incomes by Various Household Attributes," IZA Discussion Paper No. 1727.
- [48] Maasoumi, E. and D.L. Millimet (2005), "Robust Inference Concerning Recent Trends in U.S. Environmental Quality," *Journal of Applied Econometrics*, 20, 55-77.
- [49] McFadden, D. (1989), "Testing for Stochastic Dominance," in Part II of T. Fomby and T.K. Seo (eds.) *Studies in the Economics of Uncertainty* (in honor of J. Hadar), Springer-Verlag.
- [50] Rivkin, S., E.A. Hanushek, and J. Kain (2002), "Teachers, Schools, and Academic Achievement," unpublished manuscript, Hoover Institution, Stanford University.
- [51] Summers, A. and B. Wolfe (1977), "Do Schools Make a Difference?" *American Economic Review*, 56, 639-652.
- [52] Tiebout, C.M. (1956), "A Pure Theory of Local Expenditures," *Journal of Political Economy*, 64, 416-424.
- [53] Todd, P.E. and K.I. Wolpin (2003), "On the Specification and Estimation of the Production Function for Cognitive Achievement," *Economic Journal*, 113, F3-F33.

- [54] Wößmann, L. (2003), “Schooling Resources, Educational Institutions and Student Performance: the International Evidence,” *Oxford Bulletin of Economics and Statistics*, 65, 117-170.

## A Appendix: Technical Details

### A.1 Computation of $\hat{d}$ and $\hat{s}$

The test for FSD requires:

- (i) computing the values of  $\hat{F}(z_j)$  and  $\hat{G}(z_j)$  for  $z_j, j = 1, \dots, J$ , where  $J$  denotes the number of points in the support  $\mathcal{Z}$  that are utilized ( $J = 500$  in the application, where the points are equally spaced beginning at the first percentile and ending at the 99<sup>th</sup> percentile of the empirical support,  $\mathcal{Z}$ , to focus attention away from extreme values),
- (ii) computing the differences  $d_1(z_j) = \hat{F}(z_j) - \hat{G}(z_j)$  and  $d_2(z_j) = \hat{G}(z_j) - \hat{F}(z_j)$ , and
- (iii) finding  $\hat{d} = \sqrt{\frac{NM}{N+M}} \min \{\max\{d_1\}, \max\{d_2\}\}$ .

If  $\hat{d} \leq 0$  (to a degree of statistical certainty), then the null of FSD is not rejected. Furthermore, if  $\hat{d} \leq 0$  and  $\max\{d_1\} < 0$ , then  $X$  FSD  $Y$ . On the other hand, if  $\hat{d} \leq 0$  and  $\max\{d_2\} < 0$ , then  $Y$  FSD  $X$ . If  $\hat{d} = \max\{d_1\} = \max\{d_2\} = 0$ , then the (estimated) distributions of  $X$  and  $Y$  are identical. The test for SSD requires the following additional steps:

- (i) calculating the sums  $s_{1j} = \sum_{k=1}^j d_1(z_k)$  and  $s_{2j} = \sum_{k=1}^j d_2(z_k)$ ,  $j = 1, \dots, J$ , and
- (ii) finding  $\hat{s} = \sqrt{\frac{NM}{N+M}} \min \{\max\{s_{1j}\}, \max\{s_{2j}\}\}$ .

If  $\hat{s} \leq 0$  (to a degree of statistical certainty), then the null of SSD is not rejected. Moreover, if  $\hat{s} \leq 0$  and  $\max\{s_{1j}\} < 0$ , then  $X$  SSD  $Y$ ; otherwise, if  $\max\{s_{2j}\} < 0$ , then  $Y$  SSD  $X$ .

### A.2 The Recentered Bootstrap

To obtain recentered bootstrap p-values, we compute the relative frequency of  $\{\hat{d}^{**} > \hat{d}\}$ , where  $\hat{d}^{**}$  is the recentered bootstrap estimate of  $d$ . The recentering algorithm requires:

- (i) generating bootstrap samples of size  $N$  ( $M$ ) from  $X$  ( $Y$ ),
- (ii) computing the values of  $\hat{F}^*(z_j)$  and  $\hat{G}^*(z_j)$  for  $z_j, j = 1, \dots, J$ , where the values of  $z_j$  used to analyze the original sample are utilized,
- (iii) computing the differences  $d_1^c(z_j) = \left[ \hat{F}^*(z_j) - \hat{G}^*(z_j) \right] - \left[ \hat{F}(z_j) - \hat{G}(z_j) \right]$  and  $d_2^c(z_j) = \left[ \hat{G}^*(z_j) - \hat{F}^*(z_j) \right] - \left[ \hat{G}(z_j) - \hat{F}(z_j) \right]$ , and
- (iv) finding  $\hat{d}^{**} = \sqrt{\frac{NM}{N+M}} \min \{\max\{d_1^c\}, \max\{d_2^c\}\}$ .

We then compute the relative frequency of  $\{\hat{d}^{**} > \hat{d}\}$ , where  $\hat{d}$  is the sample estimate of  $d$ .

**Table 1. Unconditional Stochastic Dominance Tests.**

Distributions		Observed	First Order Dominance									Second Order Dominance									
$X$	$Y$	Ranking	$d_{1,MAX}$	$d_{2,MAX}$	$d$	$Pr\{d_1^* \leq 0\}$	$Pr\{d_2^* \leq 0\}$	$Pr\{d^* \leq 0\}$	$Pr\{d_1^* \geq d_1\}$	$Pr\{d_2^* \geq d_2\}$	$Pr\{d^* \geq d\}$	$s_{1,MAX}$	$s_{2,MAX}$	$s$	$Pr\{s_1^* \leq 0\}$	$Pr\{s_2^* \leq 0\}$	$Pr\{s^* \leq 0\}$	$Pr\{s_1^* \geq s_1\}$	$Pr\{s_2^* \geq s_2\}$	$Pr\{s^* \geq s\}$	
<b>A. 8<sup>th</sup> Grade Class Size: 10<sup>th</sup> Grade Test Scores</b>																					
Small	Medium	None	2.150	1.533	1.533	0.000	0.000	0.000	0.622	0.598	0.514	246.826	9.049	9.049	0.000	0.266	0.266	0.480	0.530	0.528	
						0.016	0.012	0.028	0.030	0.090	0.000				0.148	0.130	0.278	0.092	0.666	0.336	
Small	Large	None	1.636	0.804	0.804	0.002	0.000	0.002	0.658	0.814	0.752	153.163	4.385	4.385	0.096	0.122	0.218	0.538	0.634	0.522	
						0.014	0.028	0.042	0.136	0.486	0.126				0.220	0.328	0.548	0.242	0.602	0.326	
Medium	Large	None	0.714	0.648	0.648	0.000	0.002	0.002	0.774	0.818	0.640	26.211	41.308	26.211	0.354	0.100	0.454	0.538	0.594	0.198	
						0.020	0.018	0.038	0.608	0.618	0.308				0.266	0.328	0.594	0.586	0.456	0.114	
<b>B. 10<sup>th</sup> Grade Class Size: 12<sup>th</sup> Grade Test Scores</b>																					
Small	Medium	M SSD S	2.720	1.169	1.169	0.000	0.000	0.000	0.658	0.658	0.658	373.404	-0.439	-0.439	0.000	0.882	0.882	0.524	0.598	0.598	
						0.008	0.010	0.018	0.000	0.132	0.000				0.198	0.184	0.382	0.004	0.992	0.986	
Small	Large	L SSD S	4.149	0.333	0.333	0.000	0.012	0.012	0.608	0.728	0.728	547.169	-0.174	-0.174	0.000	0.870	0.870	0.542	0.496	0.496	
						0.010	0.004	0.014	0.000	0.838	0.704				0.152	0.190	0.342	0.000	0.940	0.892	
Medium	Large	L FSD M	2.935	0.044	0.044	0.000	0.134	0.134	0.578	0.804	0.804	477.344	0.112	0.112	0.000	0.408	0.408	0.518	0.532	0.532	
						0.008	0.008	0.016	0.000	0.984	0.968				0.184	0.176	0.360	0.000	0.808	0.602	

NOTES: Small classes have fewer than 20 students; medium classes have between 20 and 30 students; large classes have 31 or more students. All results use appropriate panel weights. Probabilities obtained via 500 bootstrap repetitions (first row: simple bootstrap; second row: recentered bootstrap). No observed ranking implies only that the distributions are not rankable in the first- or second-degree sense. See text for further details.

**Table 2. Conditional Stochastic Dominance Tests.**

Distributions		Observed Ranking	First Order Dominance									Second Order Dominance								
X	Y		d <sub>1,MAX</sub>	d <sub>2,MAX</sub>	d	Pr{d <sub>1</sub> *≤0}	Pr{d <sub>2</sub> *≤0}	Pr{d*≤0}	Pr{d <sub>1</sub> *≥d <sub>1</sub> }	Pr{d <sub>2</sub> *≥d <sub>2</sub> }	Pr{d*≥d}	S <sub>1,MAX</sub>	S <sub>2,MAX</sub>	s	Pr{s <sub>1</sub> *≤0}	Pr{s <sub>2</sub> *≤0}	Pr{s*≤0}	Pr{s <sub>1</sub> *≥s <sub>1</sub> }	Pr{s <sub>2</sub> *≥s <sub>2</sub> }	Pr{s*≥s}
<b>A. 8<sup>th</sup> Grade Class Size: 10<sup>th</sup> Grade Test Scores (Partial Residual)</b>																				
Small	Medium	M FSD S	15.571	-0.902	-0.902	0.148	0.734	0.882	0.594	0.466	0.438	3697.722	-0.985	-0.985	0.196	0.742	0.938	0.576	0.440	0.426
						0.324	0.358	0.682	0.366	0.780	0.604				0.358	0.486	0.844	0.334	0.598	0.368
Small	Large	L FSD S	30.681	-1.084	-1.084	0.008	0.978	0.986	0.552	0.584	0.584	7391.206	-1.084	-1.084	0.008	0.986	0.994	0.518	0.582	0.582
						0.252	0.148	0.400	0.000	0.930	0.922				0.378	0.370	0.748	0.002	0.736	0.728
Medium	Large	L FSD M	29.703	-0.945	-0.945	0.042	0.882	0.924	0.420	0.504	0.498	7014.720	-0.945	-0.945	0.052	0.920	0.972	0.404	0.502	0.484
						0.380	0.158	0.538	0.002	0.934	0.874				0.462	0.316	0.778	0.014	0.818	0.754
<b>B. 8<sup>th</sup> Grade Class Size: 10<sup>th</sup> Grade Test Scores (Full Residual)</b>																				
Small	Medium	S SSD M	1.256	1.615	1.256	0.000	0.000	0.000	0.978	0.896	0.962	-0.422	194.296	-0.422	0.526	0.000	0.526	0.620	0.890	0.620
						0.000	0.000	0.000	0.816	0.534	0.584				0.156	0.026	0.182	0.918	0.212	0.914
Small	Large	None	1.220	1.484	1.220	0.000	0.000	0.000	0.902	0.828	0.800	9.341	144.512	9.341	0.398	0.002	0.400	0.536	0.788	0.530
						0.004	0.000	0.004	0.796	0.636	0.558				0.256	0.056	0.312	0.672	0.422	0.588
Medium	Large	M SSD L	0.212	3.084	0.212	0.052	0.000	0.052	0.702	0.442	0.702	-0.076	461.819	-0.076	0.586	0.000	0.586	0.474	0.428	0.474
						0.000	0.004	0.004	1.000	0.052	0.988				0.122	0.182	0.304	0.880	0.006	0.738
<b>C. 10<sup>th</sup> Grade Class Size: 12<sup>th</sup> Grade Test Scores (Partial Residual)</b>																				
Small	Medium	S FSD M	-0.108	1.760	-0.108	0.574	0.216	0.790	0.450	0.590	0.260	-0.362	394.168	-0.362	0.620	0.226	0.846	0.410	0.580	0.272
						0.272	0.450	0.722	0.908	0.372	0.606				0.418	0.480	0.898	0.740	0.332	0.438
Small	Large	L FSD S	4.854	-0.165	-0.165	0.174	0.650	0.824	0.660	0.400	0.250	1121.477	-0.291	-0.291	0.188	0.746	0.934	0.630	0.314	0.178
						0.272	0.414	0.686	0.504	0.662	0.430				0.312	0.518	0.830	0.466	0.596	0.396
Medium	Large	L FSD M	8.381	-0.573	-0.573	0.122	0.644	0.766	0.586	0.464	0.420	1916.552	-0.599	-0.599	0.132	0.758	0.890	0.574	0.402	0.350
						0.300	0.298	0.598	0.384	0.874	0.774				0.356	0.396	0.752	0.334	0.818	0.720
<b>D. 10<sup>th</sup> Grade Class Size: 12<sup>th</sup> Grade Test Scores (Full Residual)</b>																				
Small	Medium	None	1.523	0.582	0.582	0.000	0.000	0.000	0.894	0.976	0.976	231.479	48.165	48.165	0.086	0.000	0.086	0.442	0.968	0.818
						0.002	0.000	0.002	0.450	0.946	0.886				0.392	0.006	0.398	0.056	0.834	0.218
Small	Large	None	1.898	0.366	0.366	0.000	0.006	0.006	0.722	0.864	0.864	346.350	19.359	19.359	0.074	0.032	0.106	0.430	0.792	0.706
						0.008	0.002	0.010	0.200	0.966	0.894				0.294	0.086	0.380	0.038	0.820	0.242
Medium	Large	None	0.867	0.452	0.452	0.000	0.004	0.004	0.840	0.746	0.720	123.885	2.296	2.296	0.110	0.270	0.380	0.514	0.670	0.532
						0.000	0.000	0.000	0.672	0.948	0.872				0.154	0.090	0.244	0.192	0.834	0.604

NOTES: Small classes have fewer than 20 students; medium classes have between 20 and 30 students; large classes have 31 or more students. All results use appropriate panel weights. Probabilities obtained via 500 bootstrap repetitions (first row: simple bootstrap; second row: recentered bootstrap). No observed ranking implies only that the distributions are not rankable in the first- or second-degree sense. First-stage regressions include: race dummies, gender dummy, limited English proficiency (LEP) dummy, father's education dummies, mother's education dummies, home computer dummy, family composition dummies, family socio-economic status, number of siblings dummies, dummies for teacher race, teacher gender dummy, teacher experience dummies, teacher education dummies, average class ability dummies, dummies indicating student and teacher are of the same race, dummy indicating student and teacher are of the same gender, dummies for amount of student disruptions in class, regional dummies, urban and rural dummies, school enrollment dummies, grade-level enrollment dummies, dummies for the percentage of minority students in school, dummies for number of total full-time teachers as well as by race, teacher salary dummies, percentage of students in school in remedial reading, percentage of students in school in remedial math, percentage of students in school in bilingual education, length of school year dummies, lagged test score, dummies indicating if family composition, number of siblings, teacher salaries, school year length, number of full-time teachers by race, and percentage of students in remedial reading, remedial math, and bilingual education are missing. Partial residual test comparisons based on equation (7); full residual comparisons based on equation (8). See text for further details.

**Table 3. Oaxaca-Blinder Decompositions of Mean Test Score Gaps.**

Class Size		Observed Gap	Portion of Observed Gap Due to Differences in:		
<i>X</i>	<i>Y</i>		Endowments	Intercepts	Coefficients
<b>A. 8th Grade Class Size: 10th Grade Test Scores</b>					
Small	Medium	0.067	0.254	-3.067	2.880
Small	Large	0.029	0.863	-10.047	9.213
Medium	Large	-0.038	0.032	-6.979	6.910
<b>B. 10th Grade Class Size: 12th Grade Test Scores</b>					
Small	Medium	-0.385	-0.267	1.272	-1.391
Small	Large	-1.939	-1.742	-2.085	1.888
Medium	Large	-1.554	-1.380	-3.357	3.183

NOTES: Positive numbers in columns 4-6 indicate advantage to *X*; negative numbers indicate advantage to *Y*. See Table 2 for further details.

**Table 4. Unconditional and Conditional Stochastic Dominance Tests: Detailed Comparison of Small Classes.**

Distributions		Observed	First Order Dominance									Second Order Dominance									
X	Y	Ranking	d <sub>1,MAX</sub>	d <sub>2,MAX</sub>	d	Pr{d <sub>1</sub> *≤0}	Pr{d <sub>2</sub> *≤0}	Pr{d*≤0}	Pr{d <sub>1</sub> *≥d <sub>1</sub> }	Pr{d <sub>2</sub> *≥d <sub>2</sub> }	Pr{d*≥d}	S <sub>1,MAX</sub>	S <sub>2,MAX</sub>	s	Pr{s <sub>1</sub> *≤0}	Pr{s <sub>2</sub> *≤0}	Pr{s*≤0}	Pr{s <sub>1</sub> *≥s <sub>1</sub> }	Pr{s <sub>2</sub> *≥s <sub>2</sub> }	Pr{s*≥s}	
<b>A. 8<sup>th</sup> Grade Class Size: 10<sup>th</sup> Grade Test Scores (Unconditional)</b>																					
Small I	Small II	II SSD I	1.943	0.278	0.278	0.000	0.114	0.114	0.564	0.692	0.692	282.157	-0.285	-0.285	0.002	0.694	0.696	0.536	0.578	0.576	
						0.016	0.008	0.024	0.016	0.898	0.802				0.172	0.200	0.372	0.042	0.946	0.926	
<b>B. 8<sup>th</sup> Grade Class Size: 10<sup>th</sup> Grade Test Scores (Partial Residual)</b>																					
Small I	Small II	I FSD II	-0.360	9.592	-0.360	0.794	0.122	0.916	0.330	0.640	0.262	-0.369	2263.625	-0.369	0.824	0.136	0.960	0.316	0.626	0.234	
						0.284	0.260	0.544	0.800	0.416	0.664				0.472	0.330	0.802	0.606	0.330	0.462	
<b>C. 8<sup>th</sup> Grade Class Size: 10<sup>th</sup> Grade Test Scores (Full Residual)</b>																					
Small I	Small II	None	2.296	1.191	1.191	0.000	0.000	0.000	0.898	0.904	0.904	239.626	96.319	96.319	0.084	0.000	0.084	0.486	0.944	0.724	
						0.000	0.000	0.000	0.448	0.890	0.792				0.238	0.012	0.250	0.128	0.786	0.130	
<b>D. 10<sup>th</sup> Grade Class Size: 12<sup>th</sup> Grade Test Scores (Unconditional)</b>																					
Small I	Small II	None	0.966	0.755	0.755	0.000	0.000	0.000	0.752	0.684	0.564	103.464	11.552	11.552	0.016	0.400	0.416	0.550	0.500	0.426	
						0.018	0.016	0.034	0.250	0.378	0.056				0.176	0.200	0.376	0.270	0.596	0.248	
<b>E. 10<sup>th</sup> Grade Class Size: 12<sup>th</sup> Grade Test Scores (Partial Residual)</b>																					
Small I	Small II	I FSD II	-0.558	8.300	-0.558	0.794	0.088	0.882	0.516	0.608	0.508	-0.585	2025.631	-0.585	0.840	0.100	0.940	0.528	0.576	0.498	
						0.206	0.304	0.510	0.932	0.272	0.802				0.312	0.408	0.720	0.870	0.220	0.718	
<b>F. 10<sup>th</sup> Grade Class Size: 12<sup>th</sup> Grade Test Scores (Full Residual)</b>																					
Small I	Small II	None	2.068	1.104	1.104	0.000	0.000	0.000	0.978	0.986	0.986	150.144	132.353	132.353	0.146	0.000	0.146	0.556	0.976	0.580	
						0.000	0.000	0.000	0.586	0.922	0.880				0.202	0.000	0.202	0.202	0.722	0.034	

NOTES: Small I classes have 10-15 students; Small II classes have 16-19 students. See Table 1 for further details.

**Table 5. Oaxaca-Blinder Decompositions of Mean Test Score Gaps.**

Class Size		Observed Gap	Portion of Observed Gap Due to Differences in:		
<i>X</i>	<i>Y</i>		Endowments	Intercepts	Coefficients
<i>A. 8th Grade Class Size: 10th Grade Test Scores</i>					
Small I	Small II	1.430	1.332	1.079	-0.982
<i>B. 10th Grade Class Size: 12th Grade Test Scores</i>					
Small I	Small II	-0.409	-0.023	6.061	-6.447

NOTES: See Table 3.

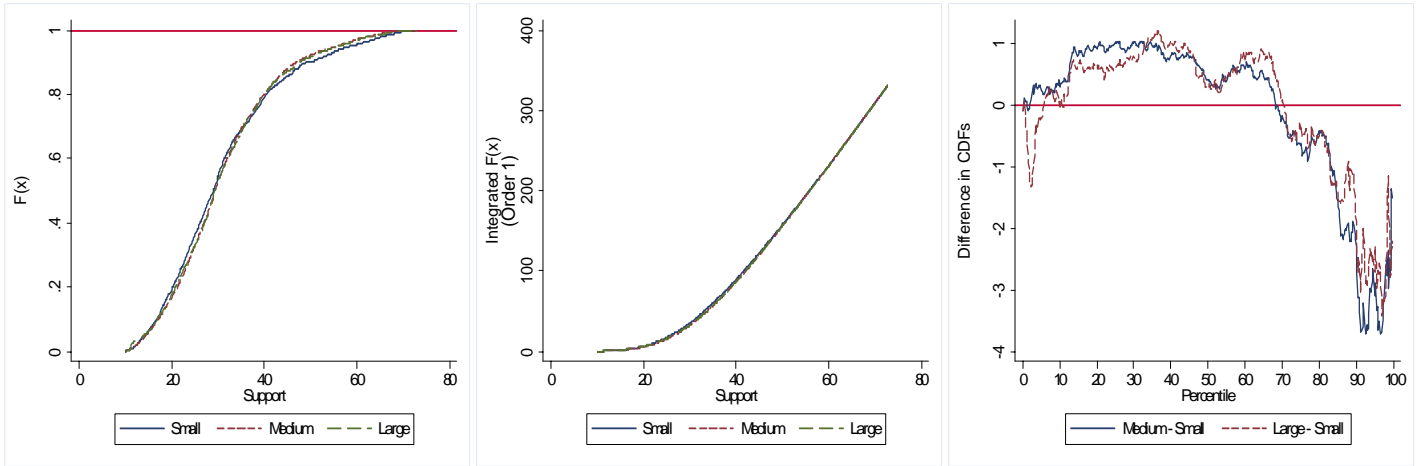


Figure 1. CDFs and Integrated CDFs: Unconditional 10<sup>th</sup> Grade Test Scores by 8<sup>th</sup> Grade Class Size.<sup>†</sup>  
<sup>†</sup>NOTE: Small  $\implies$  < 20 students, Medium  $\implies$  20 – 30 students, and Large  $\implies$  > 30 students.

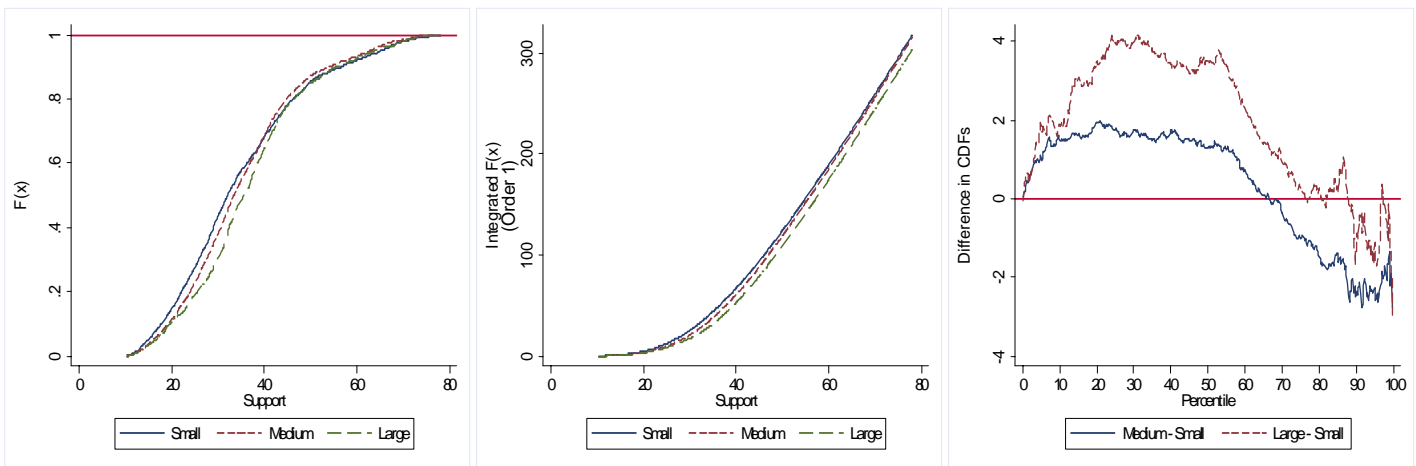


Figure 2. CDFs and Integrated CDFs: Unconditional 12<sup>th</sup> Grade Test Scores by 10<sup>th</sup> Grade Class Size.<sup>†</sup>  
<sup>†</sup>NOTE: Small  $\implies$  < 20 students, Medium  $\implies$  20 – 30 students, and Large  $\implies$  > 30 students.

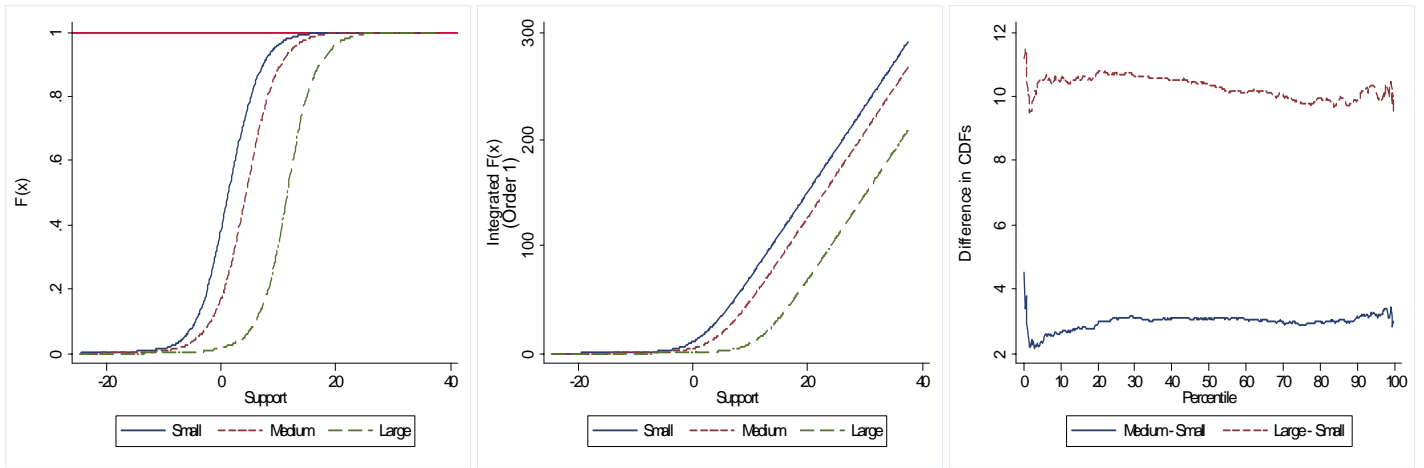


Figure 3. CDFs and Integrated CDFs: Partial Residual 10<sup>th</sup> Grade Test Scores by 8<sup>th</sup> Grade Class Size.<sup>†</sup>

<sup>†</sup>NOTE: Small  $\implies$  < 20 students, Medium  $\implies$  20 – 30 students, and Large  $\implies$  > 30 students.

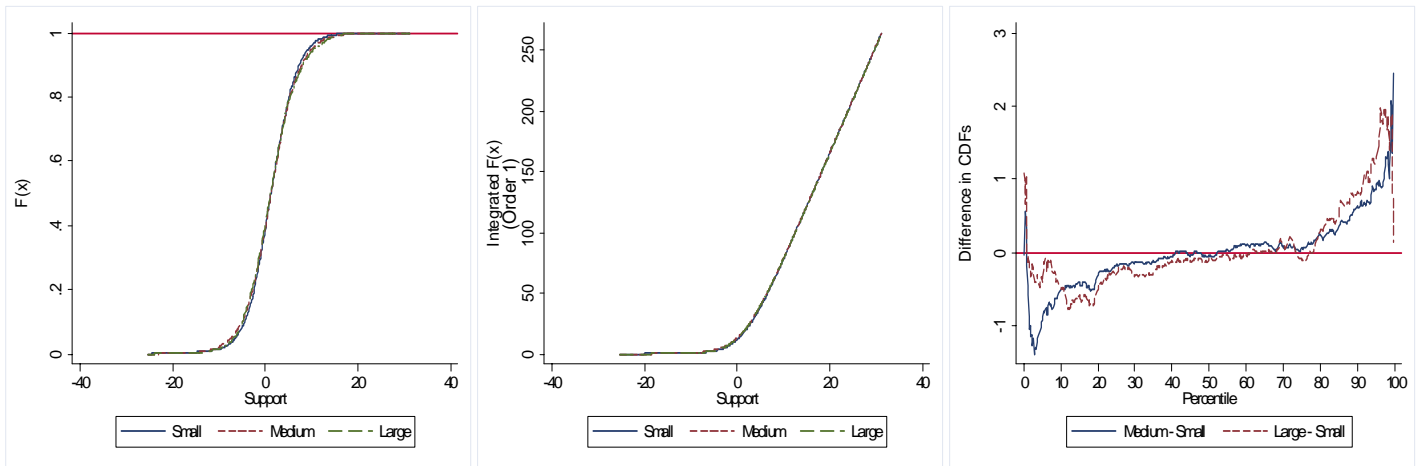


Figure 4. CDFs and Integrated CDFs: Full Residual 10<sup>th</sup> Grade Test Scores by 8<sup>th</sup> Grade Class Size.<sup>†</sup>

<sup>†</sup>NOTE: Small class size is ‘dominant’ category Small  $\implies$  < 20 students, Medium  $\implies$  20 – 30 students, and Large  $\implies$  > 30 students.

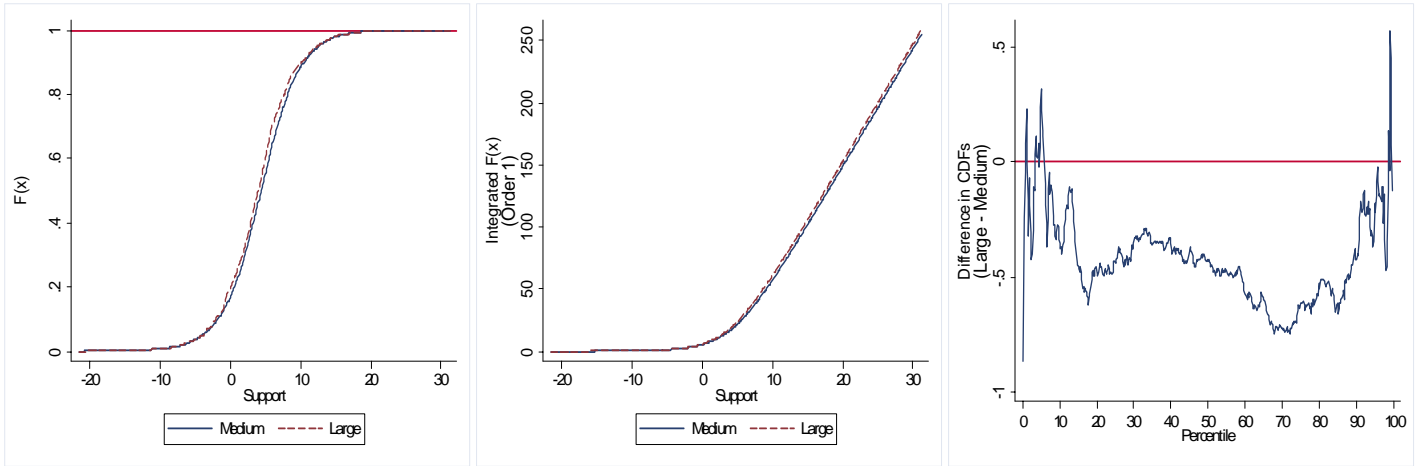


Figure 5. CDFs and Integrated CDFs: Full Residual 10<sup>th</sup> Grade Test Scores by 8<sup>th</sup> Grade Class Size.<sup>†</sup>

<sup>†</sup>NOTE: Medium class size is 'dominant' category Medium  $\implies$  20 – 30 students and Large  $\implies$  > 30 students.

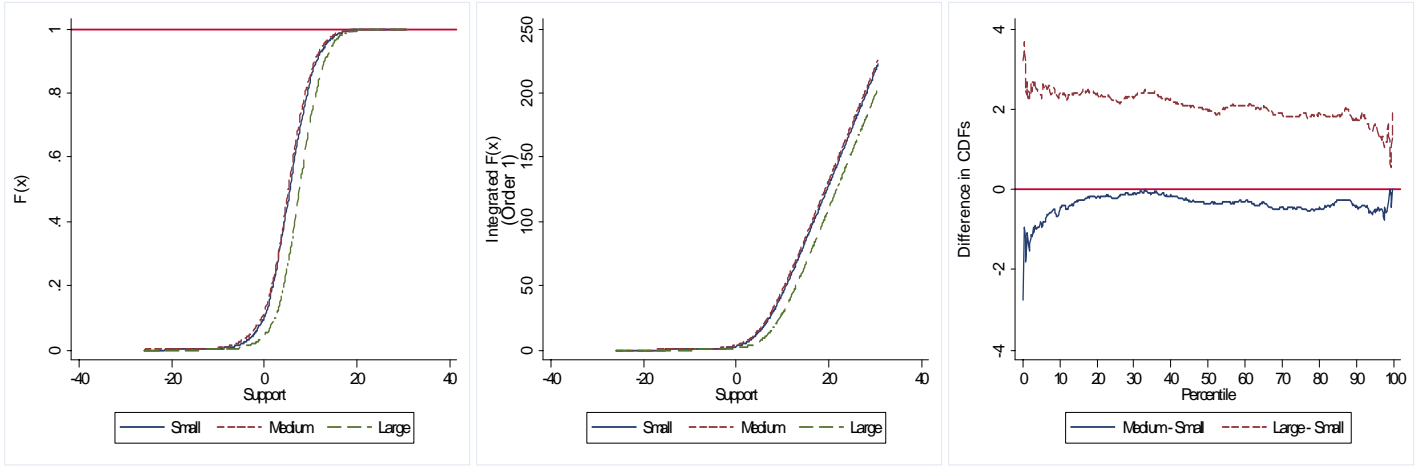


Figure 6. CDFs and Integrated CDFs: Partial Residual 12<sup>th</sup> Grade Test Scores by 10<sup>th</sup> Grade Class Size.<sup>†</sup>

<sup>†</sup>NOTE: Small  $\implies$  < 20 students, Medium  $\implies$  20 – 30 students, and Large  $\implies$  > 30 students.

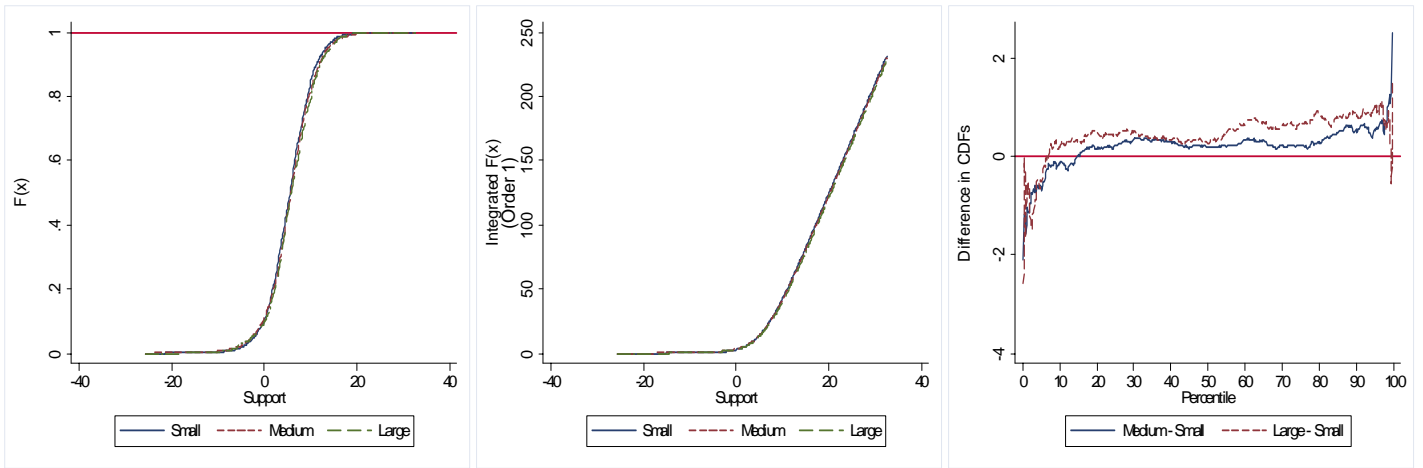


Figure 7. CDFs and Integrated CDFs: Full Residual 12<sup>th</sup> Grade Test Scores by 10<sup>th</sup> Grade Class Size.<sup>†</sup>

<sup>†</sup>NOTE: Small class size is ‘dominant’ category Small  $\implies$  < 20 students, Medium  $\implies$  20 – 30 students, and Large  $\implies$  > 30 students.

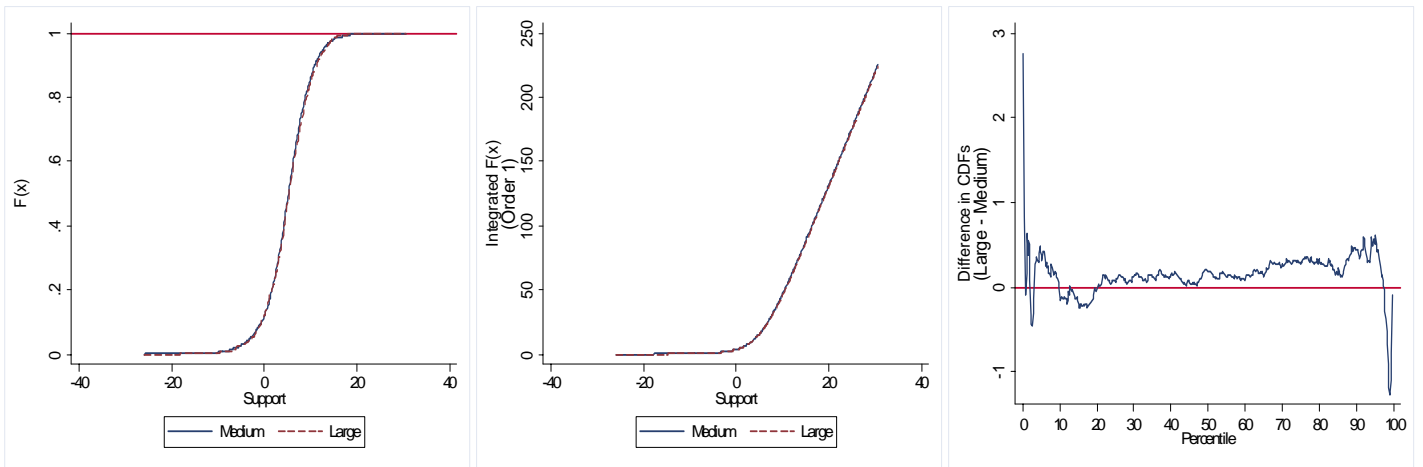


Figure 8. CDFs and Integrated CDFs: Full Residual 12<sup>th</sup> Grade Test Scores by 10<sup>th</sup> Grade Class Size.<sup>†</sup>

<sup>†</sup>NOTE: Medium class size is ‘dominant’ category Medium  $\implies$  20 – 30 students and Large  $\implies$  > 30 students.

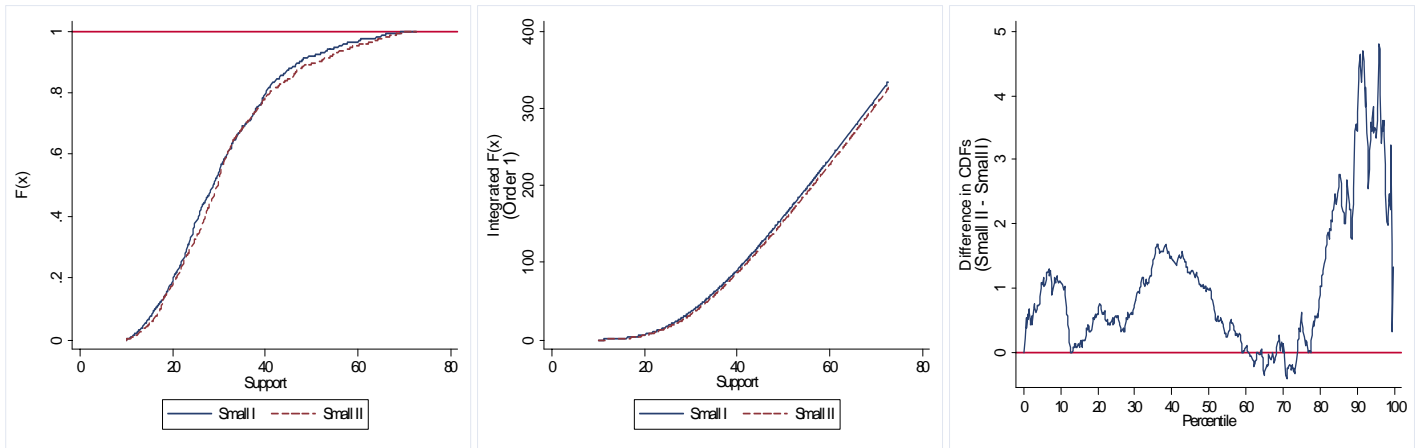


Figure 9. CDFs and Integrated CDFs: Unconditional 10<sup>th</sup> Grade Test Scores by 8<sup>th</sup> Grade Class Sizes.<sup>†</sup>

<sup>†</sup>NOTE: Small I  $\implies$  10 – 15 students and Small II  $\implies$  16 – 19 students.

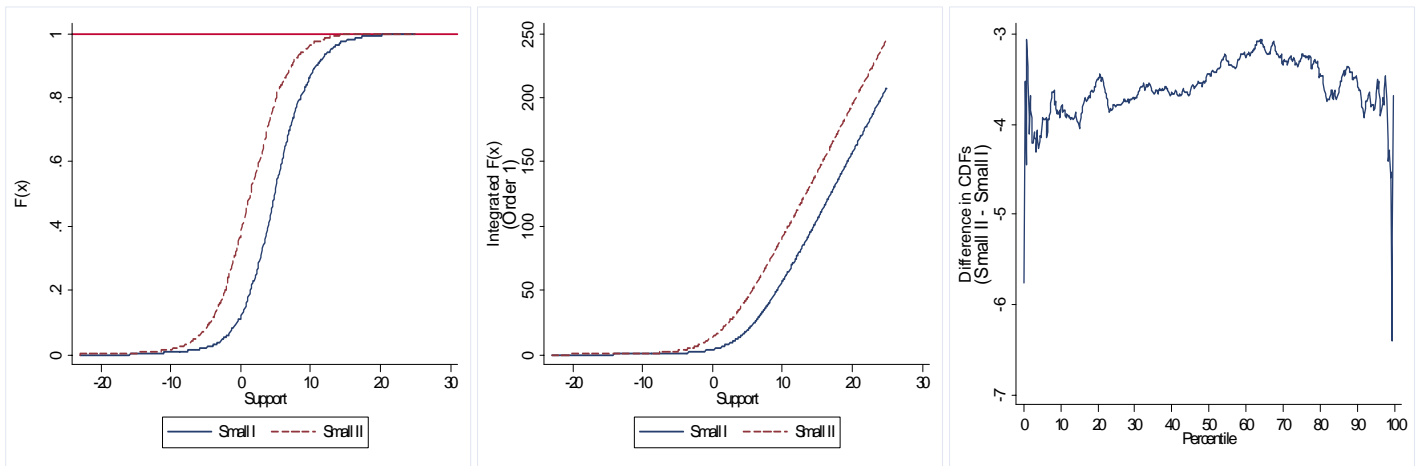


Figure 10. CDFs and Integrated CDFs: Partial Residual 10<sup>th</sup> Grade Test Scores by 8<sup>th</sup> Grade Class Sizes.<sup>†</sup>

<sup>†</sup>NOTE: Small I  $\implies$  10 – 15 students and Small II  $\implies$  16 – 19 students.

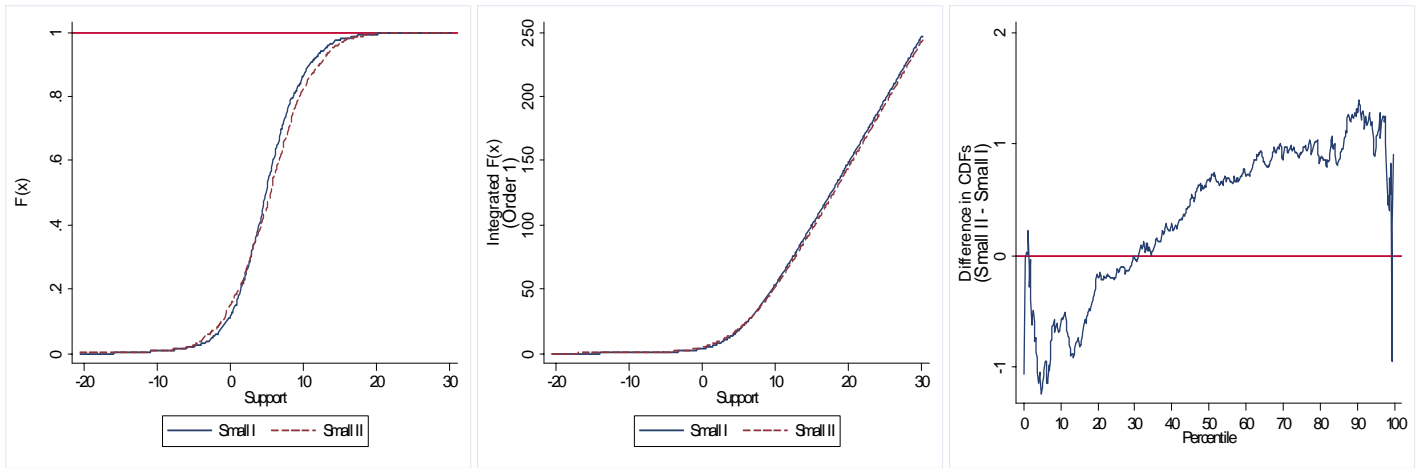


Figure 11. CDFs and Integrated CDFs: Full Residual 10<sup>th</sup> Grade Test Scores by 8<sup>th</sup> Grade Class Sizes.<sup>†</sup>

<sup>†</sup>NOTE: Small I class size as ‘dominant’ category. Small I  $\implies$  10 – 15 students and Small II  $\implies$  16 – 19 students.

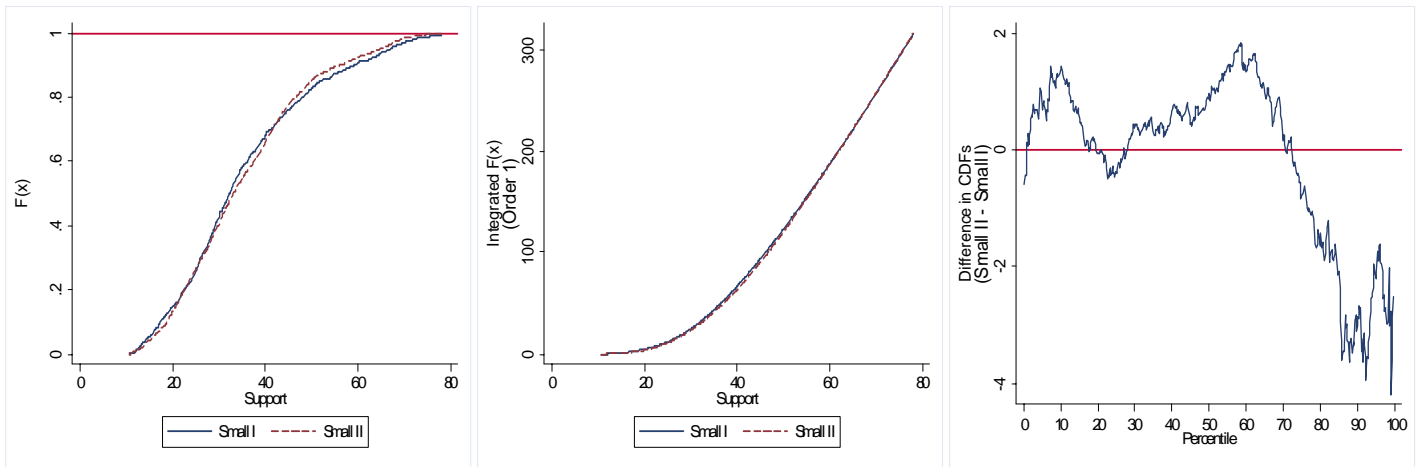


Figure 12. CDFs and Integrated CDFs: Unconditional 12<sup>th</sup> Grade Test Scores by 10<sup>th</sup> Grade Class Sizes.<sup>†</sup>

<sup>†</sup>NOTE: Small I  $\implies$  10 – 15 students and Small II  $\implies$  16 – 19 students.

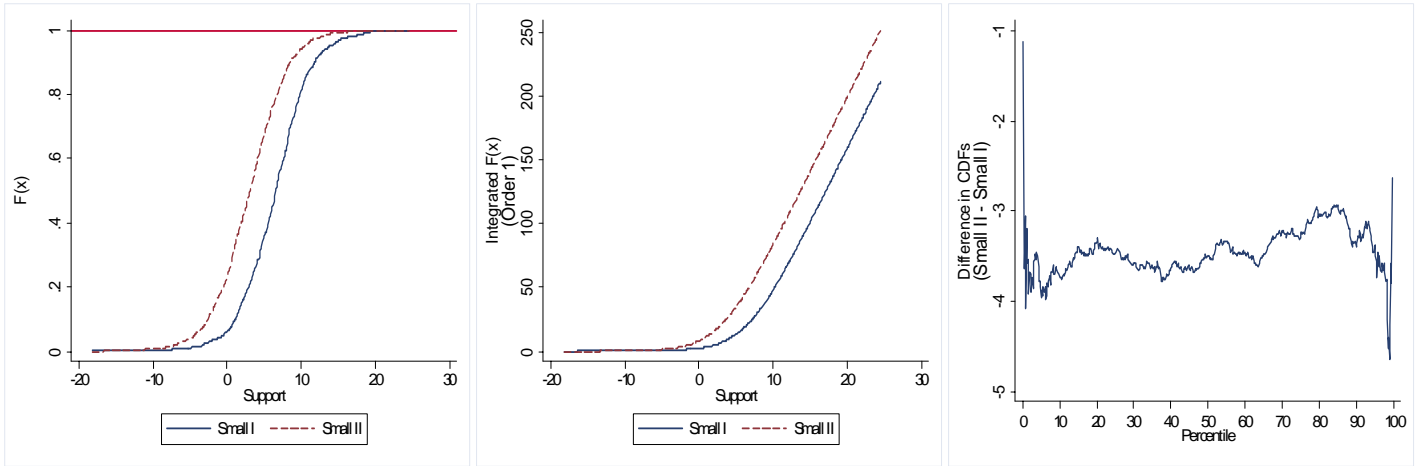


Figure 13. CDFs and Integrated CDFs: Partial Residual 12<sup>th</sup> Grade Test Scores by 10<sup>th</sup> Grade Class Sizes.<sup>†</sup>

<sup>†</sup>NOTE: Small I  $\implies$  10 – 15 students and Small II  $\implies$  16 – 19 students.

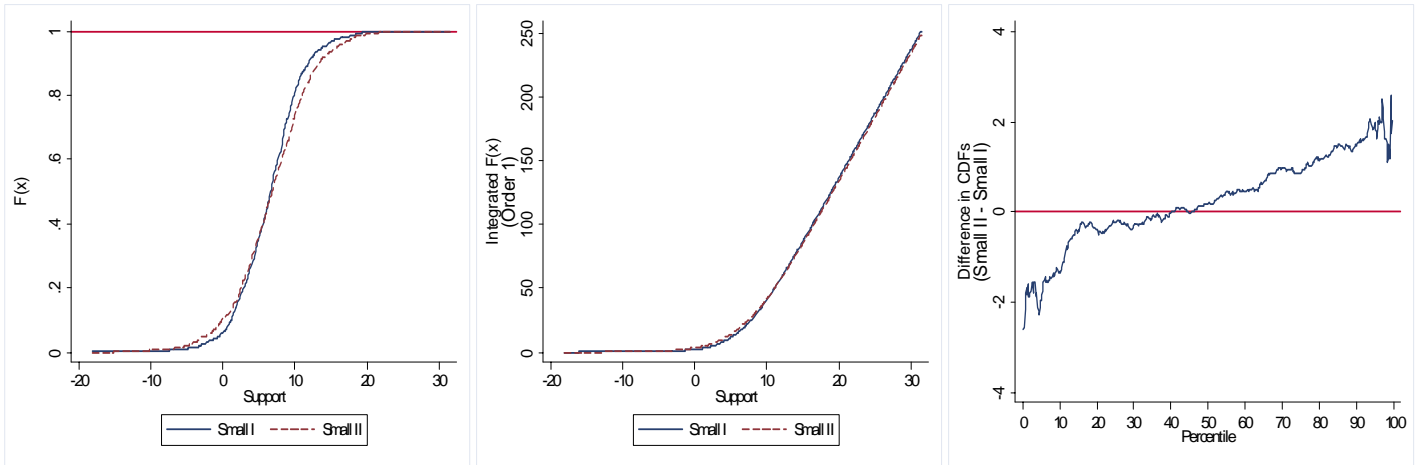


Figure 14. CDFs and Integrated CDFs: Full Residual 12<sup>th</sup> Grade Test Scores by 10<sup>th</sup> Grade Class Sizes.<sup>†</sup>

<sup>†</sup>NOTE: Small I class size as 'dominant' category. Small I  $\implies$  10 – 15 students and Small II  $\implies$  16 – 19 students.