The Effects of Class Size on Student Behavioral Outcomes: The Role of Teacher-Student Interactions

The Effects of Class Size on Student Behavioral Outcomes:

The Role of Teacher-Student Interactions

Qualifying Paper

Submitted by

Yi Xe Thng

July 2016

Acknowledgements

I would like to thank the Mathematica Policy Research Inc. and Research Connections for granting access to the Head Start FACES dataset, and the U.S. Department of Health & Services, Office of Planning, Research and Evaluation for funding the Head Start FACES study. I would also like to thank the members of my committee, Andrew Ho, Stephanie Jones, and Luke Miratrix for their thoughtful feedback on numerous drafts of this paper. Special thanks also to David Deming, for helping me develop this project in S-090 and providing feedback to the early versions of the paper. I am indebted to my advisor, Andrew Ho, for his unwavering support throughout this process.

Table of Contents

Abstract 1
Introduction 2
Background 5
Research Design 13
Empirical Strategy 23
Results 32
Discussion 36
Conclusion 43
Tables 48
References 63
Appendices 70

Tables

Table 1. Timing of data collected, by type of variable 48

Table 2. Maximum class size rules in Head Start. 49

Table 3. Sample means and sample size of the analytic sample and the rest of the sample. 50

Table 4. Details of outcome and hypothesized mediator variables. 52

Table 5. Covariate balance on baseline characteristics for treatment and comparison groups before and after propensity score matching by subclassification or full matching 53

Table 6. Effect size of smaller class size on student behavioral outcomes (RQ1) 55

Table 7. Effect size of smaller class sizes on quality of teacher-student interactions (RQ2a) 56

Table 8. Effect size of smaller class size on student behavioral outcomes after including mediator variable (RQ2b) 57

Table 9. Effect size of smaller class sizes on quality of teacher-student interactions by dimensions 59

Figures

Figure 1. Hypothesized mechanism of the effect of small class sizes on long-term outcomes. 4

Figure 2. Boxplots of absolute standardized bias before (unmatched) and after matching (subclassification or full matching). 60

Figure 3. Absolute standardized bias in means of treatment and comparison groups before and after matching by subclassification or full matching. 61

Figure 4. Propensity score distribution by treatment status. 62

Appendices

Appendix A: Student Behavioral outcome Index 70

Appendix B: CLASS Index 71

Appendix C: Variables List and Description 72

Abstract

Class size has a long history of research. To date, there is high quality evidence from causal studies suggesting that smaller class size yields short and long-term benefits for students. The understanding on how smaller class size achieve their benefits, i.e., the mechanisms, though, is less clear. Using data from the Head Start Family and Child Experiences Survey (FACES) 2009 cohort, I used propensity score techniques to investigate the effects of class size on behavioral outcomes for children who enrolled in Head Start for the first time in 2009, in full-day classrooms with predominantly 4 and 5-year olds. I also studied the role of teacher-student interactions in the classroom as a potential mediator of the above relationship. I found that smaller class sizes (17-18 children per class) had a very small but non-statistically significant effect (+0.10 *S.D.*) on student behavioral outcomes over comparison class sizes (19-20 children per class). I also found a statistically significant effect of smaller class sizes on the quality of teacher-student interactions in the classroom (+0.33 *S.D.*). This effect was driven mainly by a sub-component of the teacher-student interaction scale, namely, classroom organization (+0.42 *S.D.*). The findings did not rule out the hypothesis that the quality of teacher-student interactions in the classroom may be a potential mechanism by which smaller class size achieve their effects on students.

Introduction

Class size has been a much debated policy issue, with a long history of research (see Glass & Smith, 1979; Schanzenbach, 2014; Wilson, 2002). Prior to the 1970s, research on the effects of class size reduction was controversial, because studies yielded very different results (Mosteller, 1995). Recent studies using causal inference methods have found that smaller class sizes can improve student test scores (Angrist & Lavy, 1999; Fredriksson, Öckert, & Oosterbeek, 2013; Krueger, 1999) and provide long-term benefits (Chetty et al., 2011).

For example, using a randomized experimental design, the Tennessee Student-Teacher Achievement Ratio (STAR) study compared the effects of attending smaller class sizes (13 to 17 students) to that of regular class sizes (22 to 25 students) for four years from kindergarten through third grade (Finn & Achilles, 1990; Mosteller, 1995). The experiment found that smaller class sizes conferred short-term benefits for students' standardized test scores (Krueger, 1999), and long-term benefits in terms of high school completion (Finn, Gerber, & Boyd-Zaharias, 2005), higher earnings, college attendance, savings for retirement, as well as residence in higher-income neighborhoods (Chetty et al., 2011), and fewer arrests for crime (Krueger & Whitmore, 2001). Using a regression discontinuity approach that utilized maximum class-size rules, researchers found that after splitting classes that reached maximum class size in elementary schools, the smaller class sizes led to improvements in reading and math scores in Israel (Angrist & Lavy, 1999) and Sweden (Fredriksson, Öckert, & Oosterbeek, 2013) and benefits in areas such as motivation, self-confidence, and absenteeism for

students in Sweden (Fredriksson et al., 2013).

Despite the strength of evidence and increasing adoption of class size reduction policies at the state level in the U.S. (Education Commission of the States, 2010), debates on class size policy persist. Cost has often been cited as a barrier (Achilles, Finn, & Bain, 1998; Barnett, Schulman, & Shore, 2004; Biddle & Berliner, 2002) and has been raised as an argument in state election ballots (California Voter Guide, 1998; Washington 2014 Voters' Guide, 2014). Practical issues are also substantial when implementing class size reduction at scale, such as the difficulty of employing and training the necessary number of qualified teachers, and the challenges of creating extra classrooms (Biddle & Berliner, 2002). A few state-level studies of class size reduction programs, including California and Florida, have found little to no impact of reducing class size (Chingos, 2012; Jepsen & Rivkin, 2009). Others have acknowledged the benefits of class size reduction, but proposed that policy alternatives such as improving teacher quality are more effective given the costs (Ballotpedia, 2010; Odden, 1990).

These debates give rise to a question about mechanism: How does small class size achieve its impact on outcomes? The controversies about the effects of class size reduction could arise due to a poor understanding of the magnitude of the benefits over the costs, as well as a lack of clarity about the mechanisms at play, i.e., how smaller class sizes achieve their effects (Barnett, Schulman, & Shore, 2004; Goldstein & Blatchford, 1998). By identifying and then targeting these mechanisms, policymakers may achieve similar effects through less expensive interventions, or could undertake strategies to ensure those mechanisms are not undermined during scaled-up

implementation of the policy. In this study, I propose to explore a possible mechanism by which smaller class sizes improve student outcomes.

I hypothesize that smaller class sizes will improve student behavior directly, by increasing positive behavior and decreasing negative behavior, and also indirectly through improving the quality of teacher-student interactions, ultimately improving long-term outcomes (Figure 1).

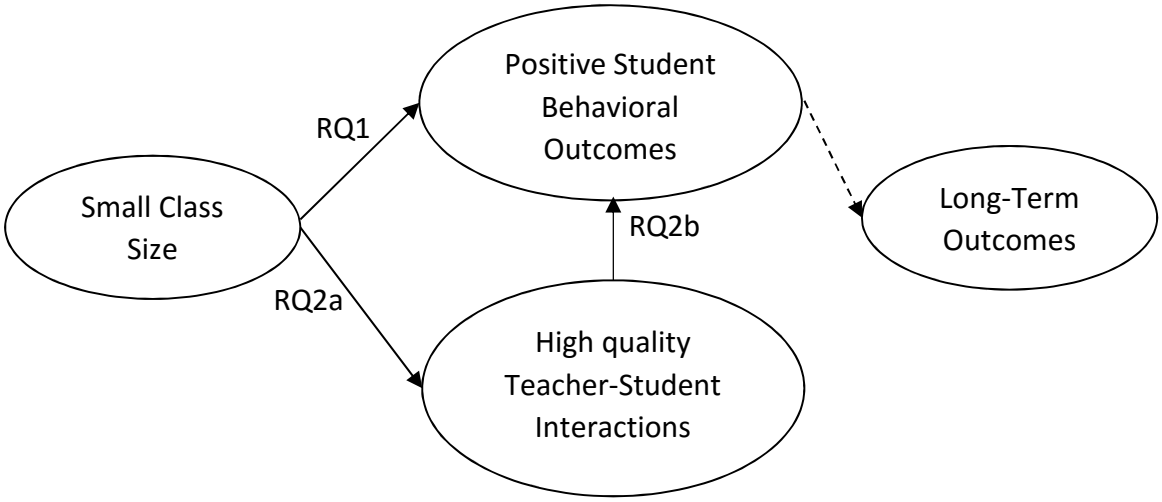


Figure 1. Hypothesized mechanism of the effect of small class size on long-term outcomes, with corresponding research questions (RQ) enumerated. The dotted line indicates links that this study does not address empirically.

I gather empirical evidence for this hypothesis in parts. First, I study the effects of class size on student behavioral outcomes, including social-emotional and problem behaviors (RQ1), which research has increasingly identified as a key predictor of school success and long-term outcomes (Raver, 2002; Duncan & Magnuson, 2011). Next, I study teacher-student interactions in the classroom as a potential mechanism by which smaller class sizes achieve their impact.

Specifically, I look at the (RQ2a) intermediary effects of class size on the quality of teacher-student interactions, and (RQ2b) the mediational role of the quality of teacher-student interactions on student behavioral outcomes.

Background

Class Size Mechanisms

Small class size is often perceived as a factor that influences student outcomes indirectly, rather than directly, by facilitating processes and conditions that increase effectiveness in teaching and learning for teachers and students (Molnar et al., 1999; Pedder, 2006; Wilson, 2002). This raises the question: How does small class size achieve its impact on outcomes? Theories on the mechanisms carrying the influence of small class size have focused on teacher behavior, student behavior, as well as teacher-student interactions. In this section, I review the literature on these potential mechanisms.

Changes in teacher behavior and teacher-student interactions

One set of theories focus on what teachers do in the classroom (Barnett, Schulman, & Shore, 2004; Biddle & Berliner, 2002; Finn, Pannozzo, & Achilles, 2003; Wilson, 2002). This set of theories focus on the proximal processes (Bronfenbrenner & Morris, 2006) in the classroom which reflect the nature and quality of children's interactions with teachers over extended periods of time. As summed up by Anderson (2002, p.52), "It is what teachers do in and with smaller classes that makes the difference, not simply being in smaller classes".

The question then is, what do teachers do differently in smaller size classrooms? A review of studies that focus on the kindergarten to lower elementary school years suggest that there are at least two aspects of teaching that have been postulated to be affected by class size. One aspect, teachers' teaching methods, involves how teachers organize the classroom and group students for instruction as well as instructional format (e.g., teacher-centered or student-centered formats). In the class size literature, teachers' teaching methods do not appear to differ very much regardless of class size. This teaching methods aspect has been the subject of earlier theories prior to, and including the Tennessee STAR experiment that postulated that reducing class size induces changes in teachers' teaching methods, such that they can provide more individualized and higher quality instruction (Finn, Pannozzo, & Achilles, 2003).

Contrary to this hypothesis, some studies have found that teachers did not change their teaching methods or beliefs as a result of reduction in class size (Evertson & Randolph, 1989; Johnston, 1990; Molnar et al., 1999). A study using observational data from STAR classrooms found that teachers did not change their teaching methods even though class size was reduced by about one-third of the original size (Evertson & Randolph, 1989). In the Evertson and Randolph (1989) study, the choice of teaching method appeared more greatly influenced by subject, rather than by class size. For example, for math, teachers in both small and large classrooms tended to use whole class instruction, followed by in-class assignments ("lecture-recitation-seatwork format" (Evertson & Randolph, 1989, p.96)). For reading, teachers tended to use reading circles for small-group reading, discussion, and in-class assignments regardless of class size.

In the SAGE (Student Achievement Guarantee in Education) class size reduction program in Wisconsin, Molnar et al. (1999) did not find evidence that teachers teaching smaller class sizes valued student-centered teaching over teacher-centered teaching more so than teachers in regular sized classrooms. Instead, Molnar et al. (1999) found that content coverage was valued more over student choice and interest. These studies suggest that class size reduction may not automatically induce teachers to change their teaching methods and beliefs.

There is a rich body of literature that examines the relationship between policies intended to change teachers' teaching practice and actual changes in their teaching practice (e.g., see Coburn, 2004; McLaughlin, 1987; Richardson, 1990). This literature suggests that teachers tend to be resistant to change even in the presence of specific policies directed at changing teaching practice. For class size reduction policies which are not direct interventions aimed at changing teaching methods, it seems even less likely that teachers would respond by voluntarily change their teaching practice.

A second aspect that seems more responsive to changes in class size is teacher-student interactions, such as answering students' questions and providing feedback, which is distinct from but may complement teachers' choice of methods. Time appears to be an important factor driving this responsiveness. For example, in interviews with 1,935 headteachers (i.e., principals), chairs of governors (i.e., heads of school board), teachers, and parents in primary schools in Britain, Bennett (1996) found that all the stakeholder groups rated time spent with individual students to be heavily influenced by class size. Presumably with fewer students in the class, teachers would have more time

to attend to each student.

This perception is supported by a number of studies in England which have found evidence for a link between smaller class size and greater quantity of teacher-student interactions. These studies tended to rely on classroom observations of the frequency and length of time spent in various types of interactions. In an observational study of 5-7 year olds in England, Blatchford et al. (2003) found that class size was negatively associated with percentage time spent teaching over class size ranges from 15 to 25 students. The frequency (number of 10-second time samples within a 5 minute observation period) of teacher-child interactions was also higher in smaller class sizes (below 20) compared to larger ones (above 30) while the frequency of not interacting was higher in the larger class sizes. In a related study of 4 and 5-year olds, Blatchford (2003) found higher frequency of occasions when children were the focus of teachers' attention in smaller class sizes (average of 19 children) than in larger class sizes (average of 33 children). In a separate study, Hargreaves, Galton, and Pell (1998) found a higher frequency (number of 25-second time samples) of feedback, both neutral and positive, as well as more sustained interactions between teachers and students in smaller class sizes.

Our understanding of the nature and quality of teacher-student interactions in smaller class sizes has come mainly through teacher interviews and self-reports. Teachers teaching smaller class sizes who were interviewed in the Tennessee STAR study (Johnston, 1990) and in the Wisconsin SAGE study (Graue et al., 2007; Graue & Oen, 2008; Molnar et al., 1999) indicated that they listened to their students more, and

developed better knowledge of their students and families. These teachers also indicated that they had more time to monitor and evaluate student learning, to provide feedback, and help in a timely manner. They could also spend more time with students who had difficulty with the material. Although these studies suggest a favorable relationship between smaller class size and the quality of teacher-student interactions, it should be noted that few studies have sought to replicate these findings through the use of independent observer ratings of the quality of teacher-student interactions. One exception is the Graue et al. (2007) study in which independent observer ratings of the classroom environment were conducted in a purposive sample to aid qualitative interpretations.

Whilst the above studies show the link between smaller class size and teacher-student interactions, other non-class size related studies have found that the quality of teacher-student interactions influence student behavioral engagement in the classroom (Downer, Rimm-Kaufman, & Pianta, 2007), as well as social skills (Moiduddin et al., 2012). Improved teacher-student interactions may lead to more sensitive and responsive relationships (Barnett, Schulman, & Shore, 2004), which in turn have been shown to be associated with better cognitive and language outcomes in the first three years of life (NICHD EECRN, 2000).

Changes in Student Behavior

Another set of theories focus on student behavior, which generally propose that students in smaller class sizes are more likely to be engaged socially and academically, and less likely to display problematic behavior, thus allowing teachers to focus more on

subject-matter instruction (Biddle & Berliner, 2002). Finn, Pannozzo, & Achilles (2003) draws on sociological and psychological perspectives to hypothesize that being in a small class size increases the "visibility of the individual" and the "sense of belonging" (p.346). With increased visibility, students cannot easily escape detection from teachers when they misbehave, and they also face more pressure to participate. In smaller class sizes, members also tend to feel greater affiliation with the group, which may influence behavior in positive ways.

Some evidence exists for improved student behavior in smaller class sizes, although the evidence generally hinges on teacher perceptions (Wilson, 2001). For example, in an interview of 28 teachers in the Wisconsin SAGE study, teachers indicated fewer disciplinary problems in classrooms with smaller class sizes (Molnar et al., 1999). They attributed reasons such as a "familylike atmosphere" (p.175) and their ability to notice and address disciplinary problems immediately, as well as more engaged students.

One study that included classroom observations and student interviews was conducted in the context of secondary schools in Hong Kong (Harfitt & Tsui, 2015). The observational study found that the students perceived a stronger sense of community in the smaller class sizes, and were more behaviorally engaged, for example, more frequently responding to teachers' questions and initiating interactions with teachers.

Stronger evidence between smaller class size and improved student behavior in the longer-term comes from the Tennessee STAR experiment. Finn and Achilles (1999) found that children assigned to smaller class sizes during *kindergarten to third grade*

scored 0.12 to 0.14 standard deviations higher on *fourth grade* teacher ratings of their learning, initiative-taking behaviors and 0.11 standard deviations lower on non-participatory behaviors (such as disruptive or withdrawn behavior). Chetty et al. (2011) replicated these results in an independent analysis of the same data for Grade 4 and extended them to Grade 8.

Improved student behavior is central to school readiness (see Raver, 2002, for a review), which has been shown to predict later performance on academic tests (Alexander & Entwisle, 1993; McLelland, Morrison & Holmes, 2000). However, children's behaviors, including social-emotional and problem behaviors, are also important as outcomes because they can affect how children interact with their peers and adults (Moiduddin et al., 2012). Researchers have also proposed that improved student behavior might be a link between smaller class size and its long-term benefits (Chetty et al., 2011; Finn & Achilles, 1999).

Examining improved student behavior as a mechanism for long-term benefits of small class size is beyond the scope of this paper. However, this paper assumes that student behavior is important as an outcome in itself. Hence, I propose to examine the relationship between smaller class size and student behavioral outcomes, specifically social-emotional and problem behaviors.

Head Start

This study is carried out within the context of Head Start classrooms. Head Start is a federally funded national program that seeks to promote school readiness for economically disadvantaged children under 5 years old (Office of the Administration for

Children and Families, 2015). The Office of Head Start administers grants to public and private, profit and non-profit agencies in local communities to provide services to young children and their families, through education, health, social and other services. Special emphasis is placed on helping preschoolers develop school readiness including the areas of reading and math, as well as social and emotional development.

Head Start classrooms provide a salient context for this study especially since smaller class size has been shown to have larger positive effects for children from low-income backgrounds than for children on average (Krueger, 1999). Moreover, studies that have documented positive benefits tended to study the effects of implementing smaller class size for younger children in kindergartens and/or elementary schools (Angrist & Lavy, 1999; Chetty et al., 2011; Finn, Gerber, & Boyd-Zaharias, 2005; Fredriksson, Öckert, & Oosterbeek, 2013; Krueger, 1999; Krueger & Whitmore, 2001). Studies have also shown that during the early childhood years, an interactive environment is important for children's learning (National Scientific Council on the Developing Child, 2004), which in theory, could be facilitated by smaller class sizes in an early childhood program such as Head Start. For reasons explained in the Research Design section, I focus on a particular segment of the Head Start population – children in full-day programs with predominantly 4 and 5-year olds in the classroom.

Summary

Researchers have hypothesized that changes in teacher and student behavior need not be mutually exclusive mechanisms of the effects of smaller class size (Biddle & Berliner, 2002). However, few studies have examined the interdependent links between

smaller class size, teacher-student interactions, and student behavioral outcomes (Finn, Pannozzo, & Achilles, 2003).

This paper contributes to literature by investigating the relationship between smaller class size and non-academic student outcomes, specifically student behavioral outcomes within an early childhood education context in Head Start. This paper also investigates the role of a potential mediator, teacher-student interactions, on this relationship. The research questions are:

RQ1: Does smaller class size predict student behavioral outcomes, including social-emotional and problem behaviors, using propensity score matching to account for selection into different class sizes in Head Start classrooms with predominantly 4 and 5-year olds in full-day programs?

RQ2: Does the quality of teacher-student interactions in the classroom mediate the effects of smaller class size on student behavioral outcomes in Head Start classrooms with predominantly 4 and 5-year olds in full-day programs? Specifically,

RQ2a: Is smaller class size associated with higher quality teacher-student interactions?

RQ2b: How well does class size explain student behavioral outcomes, including social-emotional and problem behaviors, once the quality of teacher-student interactions in the classroom is included as a covariate?

Research Design

Dataset

One of the key difficulties in studying the mechanisms of smaller class size is that the few experimental studies that had been conducted on the impact of class size did not set out to study the processes that might explain its effects (Goldstein & Blatchford, 1998). Hence, I turned to an observational dataset – the Head Start Family and Child Experiences Survey (FACES) (Malone et al., 2013). This is one of the few datasets that contains reliable and established measures of a potential mediator of smaller class size, teacher-student interactions, as well as students' behavioral outcomes (Goldstein & Blatchford, 1998). Moreover, in the Head Start FACES study, data were also collected on actual class size, i.e., the number of students and teachers in a class, as opposed to the average number of students per teacher in the school (Wilson, 2002). Furthermore, it is of interest to examine the effects of smaller class size on children in Head Start in particular since prior studies have found larger effects of smaller class size for children from low-income backgrounds than for children on average (Krueger, 1999).

The Head Start FACES is a periodic, longitudinal study of Head Start programs to provide descriptive information on a nationally representative sample of children aged 3 and 4-years old who were enrolled in the Head Start program for the first time in Fall 2009, their families, classrooms, and programs (Malone et al., 2013). Participants were selected through a multi-stage sampling design with four stages: "(1) Head Start programs, with programs defined as grantees or delegate agencies providing direct services; (2) centers within programs; (3) classrooms within centers; and (4) children within classrooms" (Malone et al., 2013, p. 28). A total of 3,718 children and families from 486 classrooms in 60 Head Start programs were sampled. Of these, 3,349 children

and their families participated in the study. I used data from the 2009 FACES cohort, that is, data on 3 and 4-year old children who enrolled in Head Start for the first time during fall 2009. The data that I used were collected in fall 2009 and spring 2010 (Table 1).

Sample

Class size in Head Start programs is guided by the Head Start Program Performance Standards (Head Start Bureau, 2005) (Table 2) which specify different class size ranges based on the predominant age of children in the classroom (3 year olds versus 4 and 5-year olds) and program type (full- versus partial-day) (henceforth termed as "class size categories"). Guided by preliminary analyses, I restricted attention to the variation in class sizes for the class size category with full-day programs serving predominantly 4 and 5-year olds (1,072 children)[1]. Table 3 compares the sample statistics for this analytic sample and the remaining sample.

My analysis by class size categories showed that some classrooms had class size beyond the range permissible by the Performance Standards. Since the characteristics that drive programs and centers to establish class sizes outside the permissible range, e.g., urbanicity, labor supply and available resources, may lead them to be substantively different from those which do so within the permissible range, defining smaller class size to be outside the permissible range and comparison classes to be within the permissible range may lead to estimates that include effects beyond smaller class size

[1] Propensity score matching was inappropriate for the other class size categories as satisfactory covariate balance could not be obtained.

alone. The dataset also did not contain sufficient covariates, e.g., demographic and socio-economic variables, at the program level to allow for matching. Hence I restricted my analytic sample to the classrooms that had class size within the range permissible by the Performance Standards. This limits the interpretation of my results to this specific group of students. The restriction eliminates a further 159 cases leaving 913 children across 135 classrooms. With my final analytic sample, I conducted a complete case analysis with 610 children across 115 classrooms.

Instruments and Measures

Question predictor (*SMALL*). I used a dichotomized indicator for smaller class size because this presents a simple case for estimating treatment effects. The alternative – to treat the different class sizes as multiple treatment doses – has been identified as an active research area (Stuart, 2010; see also, Imbens, 2000). I used the median class size within my analytic sample (19 children per class) to distinguish between smaller (17-18 children per class) and comparison (19-20 children per class) class sizes. The ensuing average class size was 17.6 and 19.9 children per class in the treatment and comparison group respectively.

Studies that have documented positive effects of class size on student outcomes have tended to have sizeable differences between one-third up to one-half the original class size (e.g., Angrist & Lavy, 1999; Chetty et al., 2011; Dee & West, 2011; Finn, Gerber, & Boyd-Zaharias, 2005; Fredriksson, Öckert, & Oosterbeek, 2013; Krueger, 1999; Krueger & Whitmore, 2001). However, studies that have documented positive relationships between class size and teacher and/or student behavior were more mixed

in their construction of the smaller class size variable. Some studies (e.g., Blatchford et al., 2003) have used class size as a continuous variable, and estimated an approximately linear relationship between class size and key variables such as percentage teaching time within the range of 15-25 children per class. Other studies (e.g., Blatchford, 2003; Blatchford et al., 2003; Hargreaves, Galton, & Pell, 1998) have grouped class sizes into small (e.g., below 20), large (e.g., 30), and sometimes various in-between categories.

The difference of an average of two students between smaller (average 17.6 children per class) and comparison (average 19.9 children per class) class sizes represents a very small variation in class size. Assuming a six-hour class day with one teacher who teaches continuously, the teacher could spend an extra 2.5 minutes per day, representing a 13% increase, with each child in the smaller class. Though seemingly inconsequential, it is the appropriate use of this short extra time, such as to provide an additional word of encouragement or a short feedback, accumulated over time (average 7.5 hours in a 36-week academic year) which together could have the potential to lead to general improvements in teacher-student relationships. While not ideal, this small variation presents an opportunity to test whether incremental small differences, for example in situations where only a limited budget is available, can make a difference.

Outcome variables. I used three measures of behavioral outcomes provided in the FACES 2009 dataset. Two of the measures were based on teacher reports on children's cooperative behavior and problem behavior in the classroom. To reduce the over-reliance on teacher reports (Finn et al., 2003), I used a third measure based on independent assessor ratings of children's social/cognitive behavior.

With multiple behavioral outcomes and subsequently mediator variables, the probability of a false rejection (Type I error) increases (see Deming, 2009). To address multiple inference, I created a composite index for children's behavioral outcomes based on the first component using principal components analysis (See Appendix A for details). The composite is constructed such that good outcomes, i.e., children's cooperative behavior and social/cognitive behavior, have a positive weighting, while the bad outcome, i.e. problem behavior, has a negative weighting. Overall, more positive values on the composite would indicate more of the good outcomes and/or less of the bad outcome.

The first measure was based on teachers' ratings of children's cooperative classroom behavior, such as following teacher's directions, and waiting for their turn during classroom and play activities. This measure was adapted from the Personal Maturity Scales developed by Alexander and Entwisle in 1988, and the Social Skills Rating Systems developed by Gresham and Elliott in 1990 (as cited in Malone et al., 2013). The Personal Maturity Scales was used by Zill and Daly (1993) in the 1976–1977 National Survey of Children, and modified by Alexander and Entwisle (Alexander, Entwisle, & Dauber, 1993; Alexander, Entwisle, & Horsey, 1997).

The second measure was based on teachers' ratings of children's problem behaviors such as being unable to pay attention, disrupting class activities, and fighting. This measure was modified from the Personal Maturity Scales developed by Alexander and Entwisle in 1998, and the Behavior Problems Index developed by Peterson and Zill in 1986 (as cited in Malone et al., 2013).

The third measure was based on independent assessors' ratings of children's behavior during testing sessions, on their level of activity, attention, organization/ impulse control, and sociability, using the Leiter International Performance Scale Revised (Leiter-R) Examiner Rating Scale. The Leiter-R examiner ratings were previously used in two large-scale studies – Administration for Children and Family's (2006) Early Head Start Transition to Prekindergarten, and Olds et al.'s (2004) Home Visiting 2000 (as cited by Malone et al., 2013). Table 4 provides further details for these three measures.

Hypothesized mediator variable (*CLASS*). I used the Classroom Assessment Scoring System (CLASS) (Pianta, La Paro, & Hamre, 2008) which measures quality in the classroom with respect to teacher-student interactions. The CLASS has been used in many studies that view teacher-student interactions as an important process measure for quality in classrooms (e.g., LoCaSale et al., 2007; Ponitz et al., 2009; Raver et al., 2008). The CLASS was developed based on "scales used in large-scale classroom observation studies in the National Institute of Child Health and Human Development (NICHD) Study of Early Care (NICHD Early Child Care Research Network [ECCRN], 2002; Pianta, La Paro, Payne, Cox, & Bradley, 2002) and the National Center for Early Development and Learning (NCEDL) MultiState Pre-K Study (Early et al., 2005)" (Pianta, La Paro, and Hamre, 2008, p.1).

The CLASS consists of the following domains: (a) Emotional Support (ES) which measures teachers' ability to support children socially and emotionally in the classroom, (b) Instructional Support (IS) which measures how well teachers use interactions such as feedback and language modeling to support student's cognitive and language

development, and (c) Classroom Organization (CO) which measures how well teachers manage classroom processes to create an environment that facilitates learning (Pianta, La Paro, & Hamre, 2008). See Table 4 for further details.

The Emotional Support domain is further made up of four dimensions, including positive climate, negative climate, teacher sensitivity, and regard for student perspectives (Pianta, La Paro, & Hamre, 2008). Positive climate reflects a warm, respectful environment of interactions marked by enjoyment between teachers and students. Negative climate reflects presence of negative interactions, such as anger, sarcasm and disrespect, and use of punishments. Teacher sensitivity reflects the degree of teachers' attentiveness towards students' needs, both academically and emotionally. Regard for student perspectives measures the extent to which teachers' interactions value students' points of view and ideas, and provide opportunities for development of student autonomy.

The Instructional Support domain is made up of three dimensions, including concept development, quality of feedback, and language modeling (Pianta, La Paro, & Hamre, 2008). Concept development measures the degree to which teachers engage in interactions with students that promote greater understanding and higher-order thinking skills among students. Quality of feedback measures the extent to which teachers provide comments and exchanges to students' work, ideas, and actions. Language modeling measures the degree to which teachers use language to motivate student learning, such as encouraging conversations, asking open-ended questions and using advanced language.

The Classroom Organization domain is made up of three dimensions, including behavior management, productivity, and instructional learning formats (Pianta, La Paro, & Hamre, 2008). Behavior management measures the teachers' ability to prevent and redirect student misbehavior. Productivity measures how well teachers manage instructional time through use of activities, routines, and transitions to maximize student learning time. Instructional learning formats measures how well teachers uses strategies such as learning objectives, facilitation techniques, and variety of materials to promote student interest and engagement.

The above hypothesized mediator variables were measured in spring 2010. With ten dimensions, the probability of false rejection of the null hypothesis increases (e.g., see Deming, 2009). To address multiple inference, I created a composite index based on the first component using principal components (see Appendix B for details), in addition to conducting analyses by domain. However, some studies have shown that each CLASS dimension may reflect a unique aspect of the classroom experience (see LoCaSale et al., 2007). Hence, I included analyses for each of the CLASS dimensions as a means of understanding the unique aspects of quality in teacher-student interactions that were driving the results at the domain and subsequently composite index level.

Selection variables. To model the selection process, I used variables at the child and program level that influence either selection into smaller class sizes or the outcome, or both (Austin, 2011; Harder, Stuart, & Anthony, 2010), but which are "not in the causal pathway between treatment and outcome" (Harder et al., 2010, p.237).

These variables are either time-invariant, or measured at baseline, but not after students have started in the smaller class sizes (Grindal, 2011).

Appendix C describes in detail the variables used in this study. Briefly, the child-level (or family-level) selection variables include demographic variables, socio-economic variables as well as factors that could influence parents' level of involvement in their children's education (e.g., single parent households and mother's employment status). The program-level selection variables include presence of program waitlists which might influence programs to adopt larger class sizes, and an index of program director's perception of program resource challenges which could influence investment in smaller class size.

To take into account the multistage sampling design of the original FACES 2009 sample, I included survey weights as a design covariate into the propensity score model. These weights would capture information about the probability of selection and response to the survey (DuGoff, Schuler, & Stuart, 2014). I did not include primary sampling unit and strata variables as it was not feasible to include a large number of them as categorical variables in my selection model (DuGoff et al., 2014).

Covariates. After creating the matched samples, I used regression adjustment to estimate the effects of smaller class size, by including child, teacher, classroom, and program covariates in the regression model (See Appendix C for details). Regression adjustment combined with matching has been shown to be more robust and efficient especially if the selection model is properly specified (Rosenbaum, 2005; Rubin, 1979).

Empirical Strategy

Selection Bias

One of the key challenges of using observational data to study the effects of smaller class size on student behavioral outcomes is that students in smaller class sizes may systematically differ from students in larger class sizes. Furthermore, the factors driving student selection into smaller class sizes are complex, and the direction of bias introduced may even contradict one another. For example, children requiring special education services, such as those with social-emotional disabilities, may be assigned to classrooms with smaller class sizes. This may possibly introduce a downward bias to student behavioral outcomes. Children whose parents are motivated to send their child to classrooms with smaller class sizes in hope of receiving a larger share of educational resources may introduce an upward bias. Under such circumstances, the overall direction and magnitude of bias is hard to predict. This motivates the use of quasi-experimental methods to address non-random selection of students into smaller class sizes.

In this study, I used propensity score techniques (Rosenbaum & Rubin, 1983) to address selection into smaller class sizes. Propensity score matching attempts to render the observed covariate distribution of the treatment and comparison groups comparable. A key aspect of this method is the modeling of the selection process into treatment. Propensity score techniques have the potential to mitigate the bias caused by confounders of the selection process and treatment outcome when the selection covariates are based on theory and knowledge of the selection process (Murnane &

Willett, 2011).

The assumption is that after applying propensity scores to balance the observed covariate distribution, children's enrolment in small and comparison Head Start class size would be as good as random. In more technical terms, "The … *assumption* … is that conditional on the measured covariates, there are no unmeasured confounders of the association between the treatment and the outcome" (Harder, Stuart, & Anthony, 2010, p.235, my italics). In reality, the unconfoundedness assumption cannot be tested, and there may still be unobserved confounders driving self-selection that remain unaccounted.

I conducted the quasi-experimental study in two stages: design and analysis stages (Rubin, 2007). In the first, or design, phase, I employed propensity score techniques to organize the data with the goal of reducing the bias between treatment and comparison groups. I first modeled the selection process by estimating the propensity score for selection into treatment status, followed by applying the propensity score to render the treatment and comparison groups more comparable using matching and subclassification methods (Harder, Stuart, & Anthony, 2010). After each application of the propensity scores, I evaluated the resulting covariate balance using criteria specified *a priori*, i.e. in the design stage before analyses were conducted (see Design Phase – Balance Diagnostics sub-section, p.29, for specific criteria). The steps in this phase were reiterated until the covariate balance between treatment and comparison groups was considered satisfactory according to the *a priori* criteria. No

outcome data were used at this stage to maintain the objectivity of the design phase (Rubin, 2007).

In the second, or analysis phase, I estimated treatment effects after conducting propensity score matching. To improve the precision of the estimate, I used covariate adjustment after propensity score matching. Through the use of a separate design and analysis phase, the quasi-experimental study attempts to approximate a randomized study in which subjects are randomly assigned to treatment and control groups without any outcome in sight (Rubin, 2007).

Mediation Analysis

Following the approach outlined in Baron and Kenny (1986), I fitted a series of three equations to address the mediational hypothesis:

(1) Regress the dependent variable (positive student behavioral outcome) on the independent variable (smaller class size) (RQ1);

(2) Regress the mediator variable (high quality teacher-student interactions) on the independent variable (smaller class size) (RQ2a); and

(3) Regress the dependent variable (positive student behavioral outcomes) on both the independent variable (smaller class size) and mediator variable (high quality teacher-student interactions) (RQ2b).

If the quality of teacher-student interactions mediate the link between smaller class size and positive student behavioral outcomes, then the following conditions must hold according to the Baron and Kenny (1986) formulation:

(1) The relationship between smaller class size and positive student behavioral outcomes is positive and statistically significant;

(2) The relationship between smaller class size and high quality teacher-student interactions is positive and statistically significant;

(3) The relationship between high quality teacher-student interactions and positive student behavioral outcomes continue to be positive and statistically significant even when smaller class size is included as an independent variable; and

(4) The magnitude of the relationship between smaller class size and positive student behavioral outcomes is smaller when the mediator variable, i.e., high quality teacher-student interactions, is included in the estimating equation than when the mediator variable is excluded.

I tested for conditions (1) to (3) using statistical inference tests with alpha of 5%, while I examined condition (4) by observation of the magnitude of the relationship of interest.

RQ1:

Design Phase: Achieving Covariate Balance

Overview. I first addressed selection into smaller class sizes by balancing the observed covariate distribution of treatment and comparison groups. The balancing was achieved through the use of exact matching on class size categories (children in full-day classrooms with predominantly 4 and 5-year olds) as well as two propensity score techniques: full matching and subclassification. In each iteration of the process, I first estimated the propensity score, applied the propensity score to the data via full

matching or subclassification, and evaluated the balance of the covariate distribution using *a priori* specified criterion, (see Design Phase – Balance Diagnostics sub-section, p.29, for specific criteria). The process of refining the propensity score model was reiterated until satisfactory balance was achieved.

Propensity score estimation. In the Head Start program guidelines, there are different class size categories which stipulate the class size based on the predominant age of children in the classroom, and the type of session (single or double session, which loosely translates to the number of hours spent in the program per day) (Head Start Bureau, 2005) (Table 2). Since both variables are explicit selection variables for class size and are likely to be associated with student behavioral outcomes, the effect of smaller class size may be substantively different for each of the class size categories. Green and Stuart (2014) found that exact matching on subgroups of substantive interest followed by estimating and matching on propensity scores separately within each subgroup resulted in the best balance among various options for propensity score estimation and matching.

Within the class size category for full-day classrooms with predominantly 4 and 5-year olds, I estimated the propensity score for being in a smaller class size using the logistic regression model:

$$P\{SMALL_{ij} = 1\} = \frac{1}{[1 + \mathrm{e}^{-(\delta_0 + \delta_1 S)}]} \qquad (1)$$

where $P(SMALL_{ij}=1)$ refers to the probability that child *i* is enrolled in a class *j* of smaller size and *S* refers to the vector of selection covariates at the child, classroom, and program level (See Appendix C for details).

Propensity score application. I used two separate propensity score techniques – full matching and subclassification – in order to check the sensitivity of results to the technique used. Compared to other common propensity score techniques, these techniques have the advantage of: (i) using all data, versus nearest neighbor matching in which data may be discarded if the controls are unmatched, and (ii) estimates not being sensitive to extreme weights, as inverse probability weighting may be (Stuart, 2010).

In full matching, every individual is grouped into a matched set consisting of at least one individual each from the treatment and comparison groups (Harder, Stuart, & Anthony, 2010; Stuart, 2010). The optimal matched sets are formed by minimizing the propensity score difference between all treatment-comparison group pairs within each matched set, and across all matched sets.

Subclassification is similar to full matching, in which individuals are grouped into subclasses containing individuals from both the treatment and comparison groups based on their propensity scores (Harder, Stuart, & Anthony, 2010; Rosenbaum & Rubin, 1984), but differs in that fewer subclasses are created. Some early work in subclassification suggests that creating five subclasses can remove "at least 90% of the bias in the estimated treatment effect due to all of the covariates that went into the propensity score" (Cochran & Rubin, 1973; Rosenbaum & Rubin, 1985 as cited in Stuart, 2010, p.9). However, depending on the sample size and the extent of propensity score

overlap between treatment and comparison groups, the optimal number of subclasses may differ (Harder et al., 2010).

Balance diagnostics. To evaluate the balance of the covariate distribution after application of propensity scores (i.e. full matching or subclassification), I used two balance diagnostics: standardized bias and region of common support. The standardized bias was calculated as the difference in means between the treatment and comparison groups for the covariate in question, divided by the standard deviation of the original treatment group:

$$Standardized\ Bias_{before} = \frac{\bar{X}_t - \bar{X}_c}{\sigma_t} \qquad (2)$$

I applied an *a priori* criterion of considering the covariate as balanced if the absolute standardized bias is less than 25.0% (Rubin, 2001). Although *t*-tests are also commonly used as balance diagnostics, Stuart (2010) cautions against its use since such hypothesis tests are an in-sample property and often reflect the power of the test to detect statistical differences rather than actual differences in means.

I also examined the region of common support for the estimated propensity scores of the treatment and comparison groups. A greater region of overlap between the two distributions would suggest that the treatment and comparison groups are similar in the observed covariate distribution, and that application of propensity score techniques might further improve the balance. Individuals with propensity scores outside the region of common support, however, are deemed to be substantively

different from those within the region, and it is common to remove them from the analysis (Harder, Stuart, & Anthony, 2010; Stuart, 2010).

Software. I used the MatchIt software developed by Ho, Imai, King, & Stuart (2011) to generate the propensity score, balance dianostics, and matching weights. After a propensity score model with satisfactory covariate balance was developed, I exported the dataset with the corresponding matching weights into Stata for the analysis of treatment effects.

Matching Weights. The MatchIt software generates weights that estimate the average effect of treatment on the treated (ATT) (Ho, Imai, King, & Stuart, 2011).To obtain the weights to estimate the average treatment effect (ATE) (Stuart, 2011), I calculated the following:

$$ATEweight_{ti} = \frac{n_i}{n_{ti}} \times \frac{n_t}{n} \quad (3a)$$

$$ATEweight_{ci} = \frac{n_i}{n_{ci}} \times \frac{n_c}{n} \quad (3b)$$

where $ATEweight_{ti}$ and $ATEweight_{ci}$ refers to the weight applied to each treatment unit *t* or comparison unit *c* for calculating the ATE, and *i* refers to the subclass each unit is assigned to by full matching or subclassification; n_i refers to the number of units in each subclass formed by matching, n_{ti} and n_{ci} refers to the number of treatment and comparison units respectively in each subclass *i*; *n* refers to the number of units in the sample (for a specific class size category), while n_t and n_i refers to the number of treatment and comparison units respectively in the sample. The first term in equation 3a scales the treatment units so that the number of treatment units are matched

equally within that subclass. The second term in equation 3a scales the weights generated by the first term to match the number of treatment units in the sample. The same reasoning holds in equation 3b for comparison units. Overall, this weighting scheme adjusts for the uneven numbers between treatment and comparison groups within and across subclasses, so that the treatment and comparison groups contribute their proportional weight to the average treatment effect.

Analysis Phase: Estimating Average Treatment Effect

For both the full matching and subclassification approaches, I used the following model with the corresponding matching weights to estimate the average treatment effect:

$$Y_{ij} = \beta_0 + \beta_1 SMALL_j + \lambda' Z + e_{ij} + u_j \qquad (4)$$

where Y_{ij} represents the outcome for child *i* in classroom *j*, $SMALL_j$ represents the treatment indicator for classroom *j*, vector *Z* represents the set of teacher and classroom covariates, and e_{ij} represents a mean-zero error term adjusted for clustering at the classroom level. The matching weights are calculated from the MatchIt software. β_1 represents the ATE of smaller class size, where a positive value indicates better behavioral outcomes for children in the treatment group (Mediation condition 1).

RQ2: I investigated whether the quality of teacher-student interactions in the classroom could be a potential mediator of the link between smaller class size and student behavioral outcomes in two stages.

RQ2a: In the first stage, I examined whether there is a relationship between smaller class size ($SMALL_j$) and quality of teacher-student interactions ($CLASS_j$), using

unmatched classroom-level data since the matching for RQ1 and RQ2b was performed on student-level data. I fitted OLS regression models with standard errors clustered at the classroom level:

$$CLASS_j = \alpha_0 + \alpha_1 SMALL_j + \lambda' Z + u_j \qquad (5)$$

where vector *Z* represents the set of teacher and classroom covariates. If the effect of smaller class size and children's outcomes acted through the quality of teacher-student interactions in the classroom, I would at least expect to find a statistically significant positive relationship (α_1) between smaller class size and the quality of teacher-student interactions (Mediation condition 2).

RQ2b: In the second stage, I added in the mediator variable, $CLASS_j$, to the estimating equation for RQ1:

$$Y_{ij} = \gamma_0 + \gamma_1 SMALL_j + \gamma_2 CLASS_j + \lambda' Z + \varepsilon_{ij} + u_j \qquad (6)$$

Following the Baron and Kenny (1986) formulation for studying mediation, a statistically significant positive relationship (γ_2) between student behavioral outcomes and the quality of teacher-student interactions (Mediation condition 3) as well as a smaller magnitude of γ_1 compared to β_1 (Mediation condition 4), in addition to meeting mediation conditions (1) and (2) in the previous research questions would suggest that the effect of smaller class size on student behavioral outcomes was mediated to some degree by high quality of teacher-student interactions.

Results

In Table 5, I show the covariate balance for covariates that could be associated with the outcome, or treatment status, or both, for children in full-day classrooms with

predominantly 4 and 5-year olds. Panels A and B show the means of each covariate for children in the treatment and comparison groups respectively in the unmatched dataset. Panels C and D show the means of each covariate for children in the comparison group after subclassification and full matching respectively. Covariates which have absolute standardized bias between treatment and comparison groups being greater than 25.0% are highlighted in the tables. The respective absolute standardized bias are shown visually in Figure 2 and Figure 3.

As shown in Table 5 Panels A and B, children in the smaller and comparison class sizes differed on a number of covariates, mainly at the program level. Unexpectedly, there was little difference in the percentage of children with IEP in both types of classrooms. It did not appear that there were selection effects into smaller class size based on children's home backgrounds. At the program and classroom level, however, there were a number of differences. As expected, classrooms with larger class sizes tended to be in programs with waitlists for children. The classrooms with smaller class sizes also tended to have program directors who had worked in the Head Start program for a longer time, and teachers with Bachelor's degree or above. The program directors of these classrooms also tended to perceive fewer challenges in running the program. These differences suggest that classrooms with smaller class sizes operated in different program environments from those with bigger class sizes. This motivates the need for matching to render the treatment and comparison groups more comparable.

In Figure 2, we see that after matching by subclassification and full matching respectively, the covariate balance generally improved. Specifically, in Figure 3, we see

that most covariates became more balanced using a yardstick of 25.0% in absolute standardized bias. However, a handful of covariates became more imbalanced after matching (but with absolute standardized bias still below 25.0%). One covariate, "expanded Head Start program in the past year", had its absolute standardized bias tilted above 25.0%.

An examination of the propensity score distribution for both treatment and comparison groups in Figure 4 shows substantial overlap for both matching methods used, although a number of cases had no direct overlap in the extreme ends of the propensity score distribution (14 individuals with propensity score < .054 all of whom were in the comparison group; 38 individuals with propensity score > 0.82 all of whom were in the treatment group). Almost all the individuals with propensity score below .054 were in programs that *had not* expanded Head Start in the past year, while almost all the individuals with propensity score above .82 were in programs that *had* expanded Head Start in the past year. I later removed these individuals beyond the region of common support in my analyses (Stuart, 2010), and checked for sensitivity of findings to the inclusion of these individuals.

In Table 6, I show the results for whether smaller class size predict student behavioral outcomes after propensity score matching using subclassification and full matching respectively (RQ1). The results show a very small effect size of smaller classes on positive student behavioral outcomes of around +0.10 standard deviations regardless of matching method and sample used (whether individuals beyond the region of common support were trimmed from the sample). These estimates were very noisy and

not statistically significant. In analyses not shown, I also did not find any statistically significant effects of smaller class size on each of the individual student outcome variables that made up my composite measure.

With this null finding for mediation condition 1, the question about mediation was no longer applicable. I show the rest of the results here for completeness. In Table 7, I show the results for whether smaller class size was associated with higher quality teacher-student interactions in the classroom. The results in Models 3 (no controls) and 4 (with controls) show that there was a small, positive, marginally statistically significant association between smaller class size and the quality of teacher-student interactions (+0.33 *S.D.*), i.e. positive evidence supporting mediation condition 2. Sub-analyses (Models 5-10) show that this association was driven primarily by the positive association of smaller class size with the CLASS Classroom Organization domain (+0.42 *S.D.*).

In Table 8, I show the results for the effect of smaller class size on student behavioral outcomes, after adding the quality of teacher-student interaction (CLASS composite variable, and each of the 3 CLASS domains) as a mediator variable, using full matching in the original analytic sample. If the mediation hypothesis were true, I would expect a statistically significant relationship between the quality of teacher-student interactions and student behavioral outcomes (mediation condition 3), and the magnitude of treatment effect to be smaller than that found for RQ1. I did not find a statistically significant relationship between the mediator and outcome. The magnitude of the effect of smaller class size on student behavioral outcomes generally remained unchanged after adding the mediator (Models 11-18) compared to before (Models 1 &

2). Hence both conditions did not hold in this case, and the results did not differ by propensity score matching method or whether individuals beyond the region of common support were trimmed from the sample.

Discussion

The research on smaller class size has yielded high quality evidence about its short- and long-term benefits. However, convincing decision-makers that the benefits are worth the costs continues to be a challenge. One key issue is that there is little understanding of how smaller class sizes achieve their outcomes, and few studies have addressed potential mechanisms (Barnett, Schulman, & Shore, 2004; Goldstein & Blatchford, 1998).

In this study, I investigated the hypothesis that higher quality teacher-student interactions could be a mechanism by which smaller class sizes achieve their effects on student behavior. I also investigated the effects of smaller class size on student behavioral outcomes, an important, but often neglected outcome in small class size research, but which has also been proposed to be a link to longer-term outcomes (Chetty et al., 2011; Finn & Achilles, 1999).

Limitations

Before discussing the findings, I note some limitations to my study. First, this study is based on observational data and cannot make ironclad causal claims for the main effects and mechanisms. I try to mitigate selection bias by using propensity score techniques to account for selection based on observed covariates.

Second, my study is carried out in the context of Head Start program and may not be generalizable to other contexts. In addition, the Head Start FACES 2009 study only included children aged 3 or 4-years old who attended Head Start for the first time in 2009 and not all children in Head Start during that year. Moreover, my analytic sample is sliced from the dataset according to maximum class size rules and is not a nationally representative sample of the Head Start population.

Next, the study focuses on specific facets of student behavior (social-emotional and problem behaviors) and student/teacher behavior (quality of teacher-student interactions) and is not generalizable to other possible mechanisms of small class size such as student motivation or teacher stress.

Furthermore, conclusions may only be drawn for the short-term effects (one academic year) of smaller class size on student behavioral outcomes and quality of teacher-student interactions. Also, conclusions may only be drawn for the range and particular definition of small class size adopted for the study.

Finally, the restriction of class size range may reduce statistical power to detect hypothesized effects.

Class size and student behavioral outcomes

Using propensity score matching methods, I did not find any statistically significant gains that class sizes of 17-18 students per class had over class sizes of 19-20 students per class for positive student behavioral outcomes in my sample of children attending Head Start classrooms with predominantly 4 and 5-year olds in full-day programs. The effect sizes were very small (+0.10 standard deviations).

To the best of my knowledge, there are no studies on the short-term relationship between smaller class size during the early childhood years and student behavioral outcomes to compare these results to. In the absence of a fairer comparison, I note the findings of the two closest studies.

In the first study, Dee and West (2011) studied the effects of smaller class sizes that arose when students experienced different class sizes in different subjects in *eighth grade*, on students' psychological and behavioral engagement. The authors found a very small effect size ranging from +0.05 to +0.09 standard deviations on the effect of smaller class sizes on students' psychological engagement, such as looking forward to the subject, seeing the subject as useful for their future, and not being afraid to ask questions. The authors however did not find any evidence for the effect of smaller class sizes on students' behavioral outcomes, such as disruptiveness or attentiveness. In the second study that analyzed Tennessee STAR data, researchers found *short- to middle-term* effects of being assigned to smaller class sizes during *kindergarten to third grade* on student behavioral outcomes during *fourth grade* (+0.12 to +0.14 standard deviations) (Chetty et al., 2011; Finn and Achilles, 1999).

With the above results in perspective, the effect sizes found in my study are comparable in magnitude to that found in the above two studies, except the estimates in my study are more imprecisely estimated. Dee and West (2011) noted that the effects of smaller class sizes on non-cognitive outcomes are generally smaller than that on academic outcomes. The larger effect sizes observed in the STAR study (Chetty et al., 2011; Finn and Achilles, 1999) compared to the Dee and West study, as well as my

study, could arise because students in the former study received a longer treatment duration.

It is possible that the small difference in class size, 17-18 versus 19-20 students per class, was too small to make substantial differences to the outcome studied. Previous studies in class size had a reduction in students by one-third (Tennessee STAR (Mosteller, 1995)) up to one-half (regression discontinuity studies in Israel and Sweden (Angrist & Lavy, 1999; Fredriksson, Öckert, & Oosterbeek, 2013)). Further research would be needed to look at short-term behavioral outcomes for a comparable reduction in class size.

Class size and teacher-student interactions

Using regression adjustment, I found that the quality of teacher-student interactions in the smaller class sizes of 17-18 students per class was statistically significantly higher than that in class sizes of 19-20 students per class (+0.33 standard deviations), and that this effect was driven primarily by Classroom Organization. This finding, however, was not derived via propensity score matching.

Classroom Organization domain. Even so, this finding seems to converge with previous findings that teachers spend less time managing classrooms and more time teaching when the class size is smaller. In Table 9, I present detailed results of the associations between class size and individual CLASS dimensions. I found that smaller class sizes of 17-18 versus 19-20 students per class was positively associated with the Productivity dimension, which looked at how well the teacher manages instructional routines and transitions to maximize learning time for students. It appears that having

fewer students in the classroom promotes greater productivity during lesson time. This also complements the finding that the percentage of time teachers spend teaching is higher in smaller class sizes (Blatchford, 2003; Blatchford et al., 2003).

Furthermore, smaller class size was positively associated with the Instructional Learning Formats dimension, which looks at how well the teacher uses a variety of learning modes and materials to facilitate and engage student interest. This seems to support the theory that there is greater individualization in smaller class sizes (Graue & Oen, 2008; Johnston, 1998; Molnar et al., 1999), i.e., with fewer students to manage, teachers could tailor their instruction to students' needs and learning styles and to actively engage them during class time.

However, it is puzzling that within the CLASS Classroom Organization domain, I did not find a statistically significant positive association between smaller class size and the Behavior Management dimension that looks at how well teachers set behavior expectations and redirect misbehavior. It is possible that teachers' method of classroom management is shaped by their training and prior beliefs (Kagan, 1992; Martin & Yin, 2006), hence a difference in class size alone did not change the way they manage the classroom.

Further research could be done to probe aspects of teacher behavior and practice that are amenable to changes in class size, and aspects that require further training or deeper changes in beliefs.

Instructional Support domain. It is also interesting to note the results of analysis by the other domains for which no statistically significant association was found with

smaller class size. Within the domain of Instructional Support, I found only a very small positive but statistically non-significant association between smaller class size and the Quality of Feedback dimension which included items such as whether teachers provide scaffolds to aid learning, and prompt students to explain their thinking. This result, taken together with past studies that found higher frequency of feedback in smaller class size (Hargreaves, Galton, & Pell, 1998), seems to suggest that quantity and high quality feedback may not always come together.

What was surprising though was that within the same Instructional Support domain, there was a statistically significant positive association between smaller class size and the Language Modeling dimension that measured the quantity and quality in "teachers' use of language-stimulation and language-facilitation techniques" (Pianta, La Paro, & Hamre, 2008, p. 75). This dimension included items such as whether conversations and open-ended questions occur frequently in the classroom, but also whether teachers use "advance language", "repeats" and "extends" students' responses, as well as "map" their own and "student actions with language". It is unclear whether the positive association was driven primarily through teachers of smaller class sizes allowing more conversations and asking more open-ended questions, which would agree with past findings of higher frequency of teacher-student interactions in classrooms (Blatchford, 2003; Blatchford et al., 2003), or driven by stronger performance over all items in the dimension, which would raise questions of why teachers in smaller class sizes could have higher quality of language modeling but not quality of feedback? Still within the Instructional Support domain, I did not find any

positive association between smaller class size and the Concept Development dimension, which measured how well teachers promoted higher-order thinking skills and understanding.

Emotional Support domain. The results by dimension within the Emotional Support domain, taken in the light of previous research findings, are puzzling. Past studies using Tennessee STAR and Wisconsin SAGE data found that teachers reported having greater personal and learning-related knowledge of their students such that they could better provide help and support to those who need it (Johnston, 1990; Molnar et al., 1999). Hence, we might expect a positive association between smaller class size with the Emotional Support domain, which includes dimensions that measure teacher sensitivity towards children's emotional and academic needs (Teacher Sensitivity), and warmth and respectfulness among teachers and students (Positive Climate). Instead, none of the associations were statistically significant, though there were small effect sizes (between +0.13 to +0.24 standard deviations). In terms of the Positive Climate and Negative Climate dimensions, the direction of association was opposite of what we might expect. It is unclear whether these results were due to a lack of power to detect statistically significant effects, or whether they were a case of misalignment between perceptions and actual practice.

Overall, these results, coupled with findings of previous observational studies, seems to paint a narrative that smaller class sizes might be associated with greater quantity of teacher-student interactions but that the quality of the interactions might vary. However, this study, as with the previous study, could not untangle these findings

to ascertain whether the smaller classes caused these observations in teacher-student interactions, or whether there were other factors associated with both selection into smaller class sizes and quality of teacher-student interactions that confounded the results.

Mediator Hypothesis of Teacher-Student Interactions

Since I could not establish a statistically significant relationship between smaller class size and positive student behavioral outcomes, this limited my ability to establish whether the quality of teacher-student interactions was a mediator of that relationship. The statistically significant results of the relationship between smaller class size and high quality teacher-student interactions did not close the door to the possibility of the mediator hypothesis. Future research should seek to achieve greater statistical power to ascertain this relationship.

Future research could also examine the sensitivity of findings to different definitions of “small”, and to missing data.

Conclusion

Class size research has a long history and there is strong evidence from credible research methods that smaller class sizes can improve student test scores and provide long-term benefits. Smaller class sizes are also popularly perceived by educators and parents to be beneficial for student learning. Still, debates persist over whether the benefits are worth the costs and whether there are more cost-effective policy alternatives to reducing class size. Moreover, recent studies on the large-scale

implementation of class size reduction policies in the United States have been inconsistent with results from those found in Israel and Sweden.

These debates and inconsistent large-scale implementation results give rise to a central question about mechanisms: How do class sizes achieve their results? By understanding the mechanisms of smaller class sizes, the policy debates do not have to boil down to a yes or no decision to implement class size reduction. Instead, the debates can move towards more conversations on how to utilize and optimize a policy that has been shown to work experimentally, and in the case of Israel and Sweden, on a large-scale basis.

A useful metaphor for this process is reverse engineering, the process of taking apart an object to see how it works, with the hope of re-producing it, enhancing it, or even use its critical components to create something new and better. With better understanding of the mechanisms of class size, questions could be raised, for example, on which are the critical components that should not be compromised – such as teacher quality – to ensure the success of the policy especially when implementation is at-scale? Are there policy complements to class size reduction – such as teacher professional development on strategies to enhance student learning in smaller class sizes – which if implemented could help teachers make the most use of smaller class sizes and stretch the benefits further? Can policy alternatives specifically targeting those mechanisms of smaller class sizes achieve similar benefits but at a lower cost?

Past research on class size has seldom focused on the mechanisms. My study addresses this issue by investigating relationships less often studied but which are

important for understanding the mechanisms by which smaller class sizes achieve their effects. Firstly, I examine the interdependent link between smaller class sizes, a potential mediator – high quality of teacher-student interactions, and positive student outcomes. Secondly, I examine a less often studied outcome in class size research – student behavioral outcomes – which in turn has been postulated as a mechanism for longer-term effects of class size reduction. In terms of research design, I utilized independent observations of the quality of teacher-student interactions, whereas most previous studies relied heavily on teacher reports of the quality of interactions, or independent observations of the quantitative aspect of interactions (frequency, amount of time etc.). As past experimental studies have not collected data on the quality of teacher-student interactions, I relied on an observational dataset and utilized propensity score matching to address selection into smaller class sizes.

My analysis found a very small positive but statistically non-significant relationship between smaller class sizes of 17-18 children versus 19-20 children per class and positive student behavioral outcomes in my sample through propensity score matching methods. As a result, I could not continue with mediational analysis to investigate the hypothesis that high quality of teacher-student interactions mediates the relationship between smaller class size and student behavioral outcomes. However, regression adjustment analyses found a positive association between smaller class size and the CLASS domain of Classroom Organization. This coheres with previous research findings based on observational data that smaller class size is associated with longer teaching times (Productivity dimension) and allow teachers to have greater

individualization in terms of task-related activities (Learning Formats dimension), and does not exclude Classroom Organization from the list of potential mechanisms of smaller class sizes. One surprising finding was that smaller class size was not statistically significantly associated with the CLASS domain of greater Emotional Support by teachers for students.

Future research on mechanisms of smaller class could consider combining experimental and quasi-experimental methods to study mediation. One possible experimental method to study mediation would be to couple smaller class size treatment with professional development on small class size interaction strategies. This would allow randomization into smaller class sizes, and also experimental manipulation of varying levels of the quality of teacher-student interactions.

A quasi-experimental way of studying mediation would be to combine experiments that randomly assign smaller class sizes to teachers and students with quasi-experimental methods such as propensity score matching to match classrooms with varying levels of quality of teacher-student interactions (e.g., Jo et al., 2011) or to use a counterfactual approach to mediational analysis (e.g., Vanderweele et al., 2013).

Future studies on mechanisms of class size should continue to focus on collecting reliable and valid measures of student behavioral outcomes as well as teaching processes and interactions in the classrooms, especially through independent observations (Goldstein & Blatchford, 1998). Future studies should also ensure greater variation in the range of class size studied to improve statistical power. With more

research on understanding the mechanisms of smaller class size, we can better design class size policies that can translate into benefits for students.

Tables

Table 1. Timing of data collected, by type of variable

	Type of Variable	Timing of Data Collected	
		Fall 2009	Spring 2010
1.	Question predictor	X	
2.	Outcome variables		X
3.	Hypothesized mediator variables		X
4.	Selection variables	X	
5.	Covariates		
	a. Age at outcome variable assessment		X
	b. Outcome variable baseline covariates	X	
	c. Others	X	

Table 2. Maximum class size rules in Head Start.

Ages	**Class Size**
4 and 5-year olds	Program average of 17-20 children enrolled per class. No more than 20 children enrolled in any class.
4 and 5-year olds in double session[1]	Program average of 15-17 children enrolled per class. No more than 17 children enrolled in any class.
3-year olds	Program average of 15-17 children enrolled per class. No more than 17 children enrolled in any class.
3-year olds in double session[1]	Program average of 13-15 children enrolled per class. No more than 15 children enrolled in any class.

Reference: *Head start design guide.* Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/tta-system/teaching/eecd/learning%20environments/planning%20and%20arranging%20spaces/edudev_art_00059_051606.html

[1] Author's interpretation: Each class in a double session is a partial-day program, thus allowing two partial-day programs in that classroom.

Table 3. Sample means[1] and sample size of the analytic sample (children in full-day programs with predominantly 4 and 5-year olds classrooms) and the rest of the sample.

	Analytic Sample		Rest of Sample	
	Mean	**n**	**Mean**	**n**
Child Characteristics				
Age at Start of School Term (Months)	***48.29***	***1071***	***44.27***	***1868***
Female	0.50	1072	0.49	1868
White	0.20	1072	0.22	1864
African American	***0.43***	***1072***	***0.25***	***1864***
Hispanic	***0.31***	***1072***	***0.45***	***1864***
Other race	***0.06***	***1072***	***0.09***	***1864***
Early Head Start	0.13	1063	0.13	1848
Other childcare	0.39	1069	0.37	1860
IEP/IFSP	0.05	1061	0.06	1843
Health insurance	0.96	1068	0.96	1856
Regular health provider	0.92	1063	0.91	1837
Low birth weight	***0.12***	***1061***	***0.09***	***1845***
Parent Characteristics				
Mother's education:				
Less than high school diploma	***0.33***	***978***	***0.37***	***1746***
High school/GED/vocational/Technical diploma	0.60	978	0.57	1746
Bachelor's degree or higher	0.06	978	0.05	1746
Mother's employment status:				
Employed full time	***0.31***	***981***	***0.22***	***1747***
Employed part time	0.21	981	0.21	1747
Looking for work	***0.25***	***981***	***0.21***	***1747***
Not working	***0.23***	***981***	***0.36***	***1747***
None of parents born in U.S.	***0.25***	***1036***	***0.31***	***1830***
Parent depression score	5.05	1063	4.90	1844
Household Characteristics				
Single parent household	***0.57***	***1039***	***0.47***	***1828***
Below 100% of income-poverty threshold	0.60	1072	0.63	1868
On multiple assistance programs	0.89	1072	0.89	1865
Household size	***4.51***	***1072***	***4.68***	***1868***
Moved multiple times in past year	0.12	1070	0.10	1863
English spoken at home	***0.25***	***1072***	***0.32***	***1868***
Neighborhood Characteristics				
Neighborhood crime	0.31	1067	0.28	1848
Program Characteristics				
Program waitlist	***0.94***	***1055***	***0.86***	***1860***

	Analytic Sample		Rest of Sample	
	Mean	**n**	**Mean**	**n**
Program challenges perception scale	-0.07	1055	-0.13	1860
Expanded HS program in past year	***0.18***	***1072***	***0.14***	***1868***
Director years in HS	***16.98***	***1055***	***19.16***	***1860***
Class hours per week	***34.24***	***1072***	***21.24***	***1868***
Teacher education: Bachelor's or above	***0.53***	***1072***	***0.46***	***1868***

[1] Variables that are statistically significantly different between the two groups (*p*-value < .05) are marked in italicized bold.

Table 4. Details of outcome and hypothesized mediator variables.

Name of metric	**Data Source**	**Interpretation**	**Citation**	**Original Source of Scale**	**No. of items**	**Scale**	**Reliability (**Cronbach's alpha internal consistency)
OUTCOME MEASURES							
Cooperative Classroom Behavior (*COOP*)	Teacher ratings	Children's cooperative behavior in the classroom (e.g., how often they follow teacher's directions and follow rules)	Malone et al., 2013	Personal Maturity Scale (Alexander & Entwisle, 1998) and the Social Skills Rating System (SSRS) (Gresham and Elliott 1990; Elliott et al. 1988).	12	3-point scale from 1 ("never") to 3 ("very often")	0.89
Problem Behavior (*PROB*)	Teacher ratings	Children's problem behaviors in terms of aggressive, hyperactive, and withdrawn behaviors	Malone et al., 2013	Personal Maturity Scales (Alexander & Entwisle, 1998) and the Behavior Problems Index (BPI) (Peterson and Zill 1986).	14	3-point scale from 1 ("never") to 3 ("very often").	0.87
Cognitive/Social (*LEITER*)	Assessor ratings	Children's behavior during testing sessions in areas of activity, organization/ impulse control, attention, and sociability	Malone et al., 2013	Leiter International Performance Scale Revised (Leiter-R) Examiner Rating Scale (Roid and Miller 1997)	Not published in Malone et al., 2013	4-point scale ("rarely/never," "sometimes," "often," or "usually/always")	0.90
HYPOTHESIZED MEDIATOR VARIABLES							
Classroom Assessment Scoring System (*CLASS*)	Observer ratings	Teacher-student interactions in the classroom, in three domains: Instructional Support; Emotional Support; and Classroom Organization	Malone et al., 2013	Classroom Assessment Scoring System (CLASS) (Pianta, La Paro, & Hamre, 2008)	10 dimensions rated in 4 observation cycles within the school day	7-point scale From 1 ("minimally characteristic") to 7 ("highly characteristic")	Instructional Support: 0.87; Emotional Support: 0.82; Classroom Organization: 0.77

Table 5. Covariate balance on baseline characteristics for treatment and comparison groups before (Panels A and B) and after propensity score matching by subclassification (Panel C) or full matching (Panel D) for full-day classrooms with predominantly 4 & 5-year olds (n = 610). Covariates where the absolute standardized bias between treatment and comparison groups (not shown) is greater than 0.25 are highlighted.

Covariates	**Treatment Means**	**Comparison Means**		
		Unmatched	**Subclassification**	**Full Matching**
	A	**B**	**C**	**D**
Age at Start of School Term (months)	48.68	48.61	49.71	49.94
Female	0.50	0.49	0.39	0.44
White	0.18	0.15	0.11	0.13
African American	0.40	0.49	0.45	0.46
Hispanic	0.36	0.32	0.39	0.37
Race (others)	0.06	0.04	0.04	0.04
Early Head Start	0.11	0.11	0.10	0.08
Non-HS Care Arrangements	0.43	0.40	0.43	0.39
Has IEP	0.03	0.04	0.03	0.03
Has health insurance	0.96	0.94	0.97	0.97
Has regular health provider	0.92	0.91	0.93	0.93
Low birthweight	0.09	0.12	0.09	0.08
Mother education: No high school	0.29	0.34	0.29	0.28
Mother education: High school or vocational education	0.68	0.57	0.68	0.68
Mother education: Bachelor's or above	0.04	**0.09**	0.03	0.04
Mother employed full-time	0.34	0.31	0.31	0.32
Mother employed part-time	0.23	0.21	0.27	0.27
Mother looking for work	0.21	0.24	0.19	0.21
Mother not working	0.22	0.23	0.23	0.21

Covariates	Treatment Means	Comparison Means		
		Unmatched	Subclassification	Full Matching
	A	B	C	D
None of parents born in USA	0.26	0.25	0.26	0.25
Parental depression score	4.29	4.73	3.92	4.19
Single parent household	0.52	0.55	0.53	0.54
Below 100% of income-poverty threshold	0.56	0.60	0.60	0.60
On multiple assistance programs	0.91	0.85	0.85	0.88
Household size	4.71	4.43	4.62	4.71
Moved multiple times in past year	0.13	0.13	0.14	0.13
English spoken at home	0.28	0.25	0.27	0.28
Neighborhood crime	0.30	0.27	0.35	0.29
Program waitlist	0.91	**0.99**	0.88	0.94
Program challenges perception scale	-0.24	-0.10	-0.14	-0.14
Expanded HS program in past year	0.19	0.18	**0.30**	**0.30**
Director years in HS	17.97	**14.47**	16.69	16.82
Class hours per week	32.92	**36.28**	33.42	33.24
Teacher education: Bachelor's or above	0.62	**0.43**	0.71	0.67
Classroom Weight	104.93	**138.45**	105.51	106.40

Table 6. Effect size of smaller class size on student behavioral outcomes (RQ1) in full-day classrooms with predominantly 4 & 5-year olds. Standard errors in parenthesis.

	Subclassification				Full Matching			
	Original		Trimmed[1]		Original		Trimmed[1]	
	M1	M2	M1	M2	M1	M2	M1	M2
Small class size (17-18 vs 19-20)	0.08	0.09	0.08	0.10	0.07	0.11	0.09	0.11
	(0.18)	(0.09)	(0.18)	(0.09)	(0.16)	(0.09)	(0.16)	(0.09)
Baseline behavior index		0.49***		0.49***		0.47***		0.47***
		(0.03)		(0.03)		(0.02)		(0.03)
No. of teachers in classroom		0.11		0.13		0.09		0.12
		(0.09)		(0.09)		(0.10)		(0.10)
Class hours per week		0.00		0.00		0.00		0.00
		(0.01)		(0.00)		(0.00)		(0.00)
Teacher depression score		-0.01†		-0.01†		-0.01*		-0.01*
		(0.01)		(0.01)		(0.01)		(0.01)
Teacher education: Bachelor's or above		0.03		0.05		0.00		0.02
		(0.08)		(0.07)		(0.07)		(0.07)
Age at assessment		0.01*		0.01*		0.01†		0.02*
		(0.01)		(0.01)		(0.01)		(0.01)
Female		0.14†		0.15*		0.15†		0.20**
		(0.08)		(0.08)		(0.08)		(0.08)
African American		-0.09		-0.14		-0.15		-0.19+
		(0.10)		(0.11)		(0.10)		(0.10)
Hispanic		-0.20†		-0.27*		-0.21†		-0.29**
		(0.10)		(0.10)		(0.11)		(0.11)
Has IEP		-0.24		-0.20		-0.40†		-0.37
		(0.15)		(0.16)		(0.23)		(0.24)
Single parent household		-0.33***		-0.30***		-0.30***		-0.30***
		(0.08)		(0.07)		(0.07)		(0.07)
Below 100% of income-poverty threshold		0.19*		0.19*		0.17*		0.14†
		(0.07)		(0.07)		(0.08)		(0.08)
Constant	-0.02	-0.79†	-0.03	-0.94*	-0.01	-0.72	-0.03	-1.06*
	(0.09)	(0.42)	(0.07)	(0.41)	(0.08)	(0.48)	(0.09)	(0.47)
Adjusted R-square	0.00	0.52	0.00	0.51	0.00	0.49	0.00	0.50
N	610	610	558	558	610	610	558	558

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .005$

[1] Cases where propensity scores (PS) were beyond the region of common support (PS < 0.054 or PS > 0.82) were trimmed from the analytic sample.

Table 7. Effect size of smaller class sizes on quality of teacher-student interactions (RQ2a) in full-day classrooms with predominantly 4 & 5-year olds (*n* = 115). Standard errors in parenthesis.

	Dependent variable							
	CLASS		**Instructional Support**		**Emotional Support**		**Classroom Organization**	
Independent variable	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**
Small class size	0.38*	0.33†	0.25	0.20	0.23	0.18	0.45*	0.42*
(17-18 Vs 19-20)	(0.18)	(0.19)	(0.18)	(0.19)	(0.19)	(0.19)	(0.18)	(0.19)
No. of teachers		0.39*		0.45*		0.61**		0.24
		(0.19)		(0.19)		(0.20)		(0.19)
Class hours per week		0.01		0.01		0.01		0.00
		(0.01)		(0.01)		(0.01)		(0.01)
Teacher depression score		0.01		0.01		0.01		0.00
		(0.02)		(0.02)		(0.02)		(0.02)
Teacher education:		0.42*		0.44*		0.36†		0.30
bachelor's or above		(0.18)		(0.18)		(0.19)		(0.19)
Program challenges		0.05		0.00		0.07		0.04
perception scale		(0.09)		(0.09)		(0.09)		(0.09)
Constant	-0.36**	-1.66**	3.22***	1.58**	10.31***	8.61***	8.83***	5.66***
	(0.12)	(0.57)	(0.12)	(0.57)	(0.12)	(0.58)	(0.16)	(0.57)
Adjusted R-square	.03	0.07	.00	0.04	.00	0.06	.03	0.02

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

Table 8. Effect size of smaller class size on student behavioral outcomes after including mediator variable (RQ2b) for full-day classrooms with predominantly 4 & 5-year olds, using full matching (*n* = 610). Standard errors in parenthesis.

	M11	M12	M13	M14	M15	M16	M17	M18
Small class size	0.07	0.11	0.07	0.11	0.07	0.11	0.07	0.11
(17-18 vs 19-20)	(0.16)	(0.09)	(0.21)	(0.09)	(0.21)	(0.12)	(0.21)	(0.12)
CLASS		0.01						
		(0.05)						
Instructional Support				0.03				
				(0.07)				
Emotional Support						0.05		
						(0.09)		
Classroom Organization								0.00
								(0.07)
Baseline behavior		0.46***		0.46***		0.46***		0.46***
index		(0.03)		(0.03)		(0.03)		(0.03)
No. of teachers in		0.08		0.08		0.07		0.08
classroom		(0.10)		(0.11)		(0.11)		(0.10)
Class hours per week		0.00		0.00		0.00		0.00
		(0.00)		(0.00)		(0.00)		(0.00)
Teacher depression		-0.01*		-0.01*		-0.02*		-0.01*
score		(0.01)		(0.01)		(0.01)		(0.01)
Teacher education:		0.00		-0.01		-0.01		0.00
Bachelor's or above		(0.08)		(0.08)		(0.08)		(0.07)
Age at assessment		-0.04		-0.03		-0.04		-0.04
		(0.05)		(0.05)		(0.05)		(0.05)
Program challenges		0.01†		0.01†		0.01†		0.01†
perception scale		(0.01)		(0.01)		(0.01)		(0.01)
Female		0.15†		0.15†		0.15†		0.15†
		(0.08)		(0.08)		(0.08)		(0.08)
African American		-0.15		-0.15		-0.16		-0.16
		(0.10)		(0.10)		(0.10)		(0.10)
Hispanic		-0.21†		-0.20†		-0.21†		-0.20†
		(0.11)		(0.11)		(0.11)		(0.11)
Has IEP		-0.38		-0.39		-0.39		-0.38
		(0.24)		(0.24)		(0.24)		(0.24)
Single parent		-0.30***		-0.30***		-0.30***		-0.30***
household		(0.07)		(0.07)		(0.07)		(0.07)
Below 100% of income-		0.18*		0.18*		0.18*		0.18*
poverty threshold		(0.09)		(0.09)		(0.09)		(0.08)
Constant	-0.01	-0.73	-0.01	-0.79	-0.01	-0.95	-0.01	-0.76
	(0.08)	(0.48)	(0.08)	(0.54)	(0.08)	(0.65)	(0.08)	(0.60)
Adjusted R-square	0.00	0.48	0.00	0.49	0.00	0.49	0.00	0.48

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .005$

Table 9. Effect size of smaller class sizes (Small = 17 to 18, Comparison = 19 to 20 children per class) on quality of teacher-student interactions by dimensions[1] in full-day classrooms with predominantly 4 & 5-year olds (*n* = 115). Standard errors in parenthesis.

	Dependent Variable		
	Instructional Support Dimension		
Independent Variable	**Concept Development**	**Quality of Feedback**	**Language Modeling**
Small class size	-0.03	0.13	0.39*
	(0.20)	(0.19)	(0.18)

	Emotional Support Dimension			
	Positive Climate	**Negative Climate**	**Teacher Sensitivity**	**Regard for Student Perspectives**
Small class size	-0.13	-0.28	0.23	0.24
	(0.19)	(0.20)	(0.18)	(0.19)

	Classroom Organization Domain		
	Behavior Management	**Productivity**	**Instructional Learning Formats**
Small class size	0.21	0.45*	0.36*
	(0.20)	(0.19)	(0.17)

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .005$

[1] Controls included in model were the same as those used for RQ2a: Number of teachers, Class hours per week, Teacher depression score, Teacher education: Bachelor's degree or above, Program challenges perception scale.

Figure 2. Boxplots of absolute standardized bias before (unmatched) and after matching (subclassification or full matching). Dotted line refers to absolute standardized bias of 0.25.

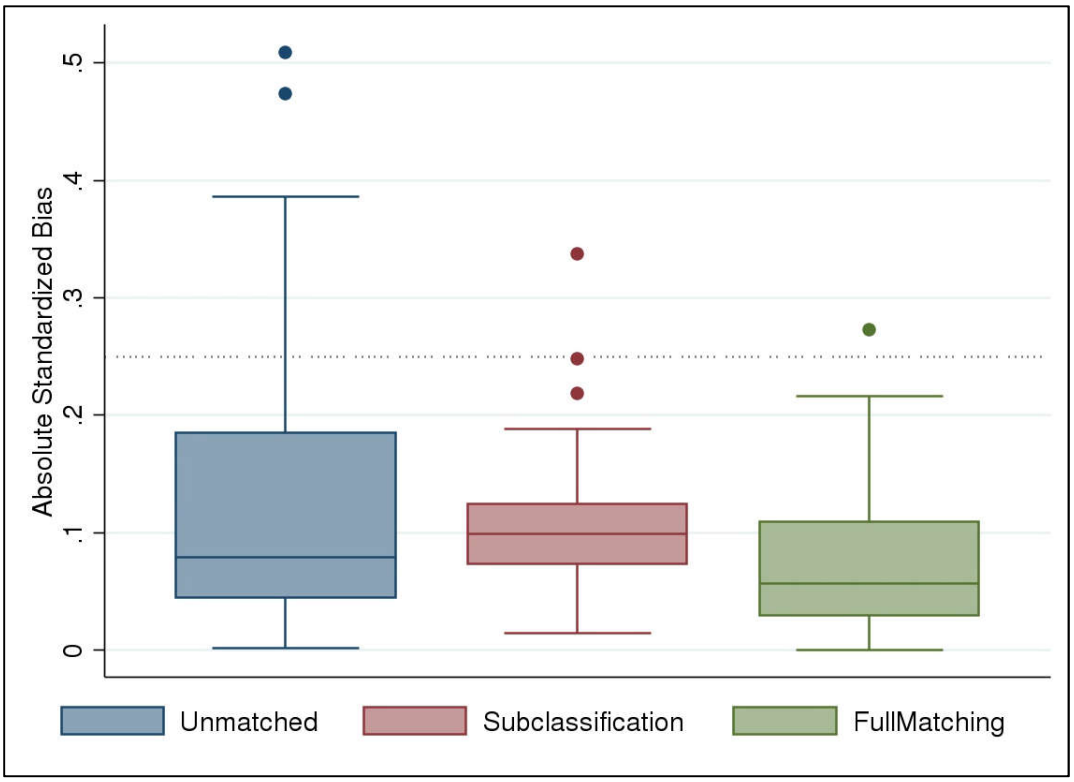


Figure 3. Absolute standardized bias in means of treatment and comparison groups before and after matching by subclassification or full matching.

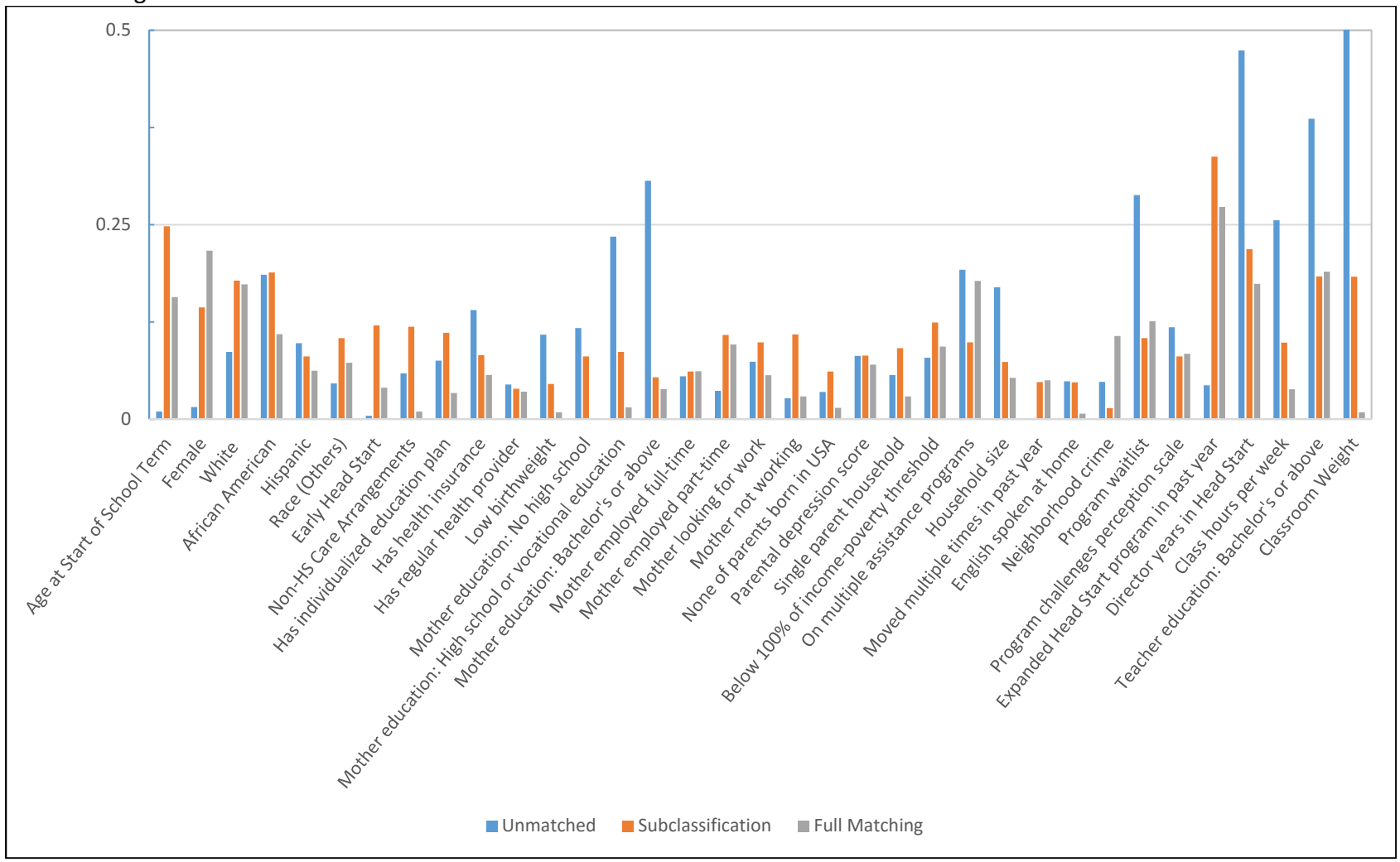
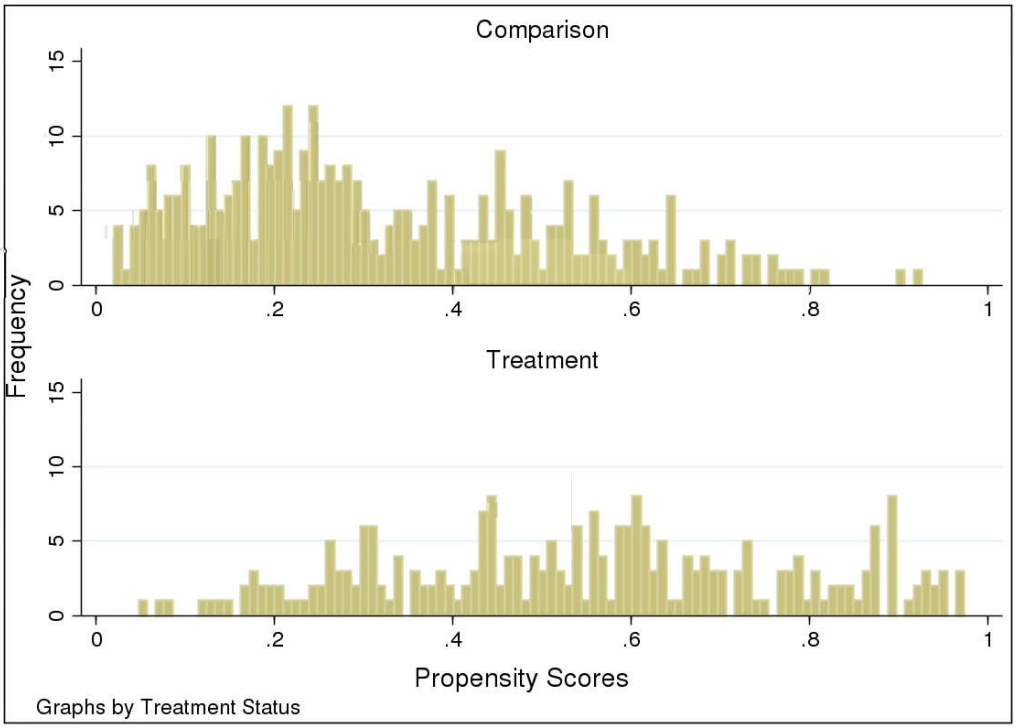


Figure 4. Propensity score distribution by treatment status.



References

Achilles, C. M., Finn, J. D., & Bain, H. P. (1998). Using Class Size to Reduce the Equity Gap. *Educational Leadership*, *55*(4), 40-43.

Alexander, K. L., Entwisle, D. R., & Dauber, S. L. (1993). First-grade classroom behavior: Its short-and long-term consequences for school performance. *Child Development*, *64*(3), 801-814.

Alexander, K. L., Entwisle, D. R., Blyth, D. A., & McAdoo, H. P. (1988). Achievement in the first 2 years of school: Patterns and processes. *Monographs of the Society for Research in Child Development*, i-157.

Anderson, L.W. (2002). Balancing breadth and depth of content coverage: Taking advantage of the opportunities provided by smaller classes. In J.D. Finn & M.C. Wang (Eds.), *Taking small classes one step further* (pp. 51-61). Greenwich, CT: Information Age Publishing Inc.

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

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399-424.

Ballotpedia (2010). *Florida Class Size Amendment 8 (2010).*Retrieved from http://ballotpedia.org/Florida_Class_Size,_Amendment_8_(2010)

Barnett, W. S., Schulman, K., & Shore, R. (2004). Class size: What's the best fit? *NIEER Policy Matters, 9, 1-11.* Retrieved from http://nieer.org/resources/policybriefs/9.pdf

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, *51*(6), 1173.

Bennett, N. (1996). Class size in primary schools: Perceptions of headteachers, chairs of governors, teachers and parents. *British Educational Research Journal*, *22*(1), 33-55.

Biddle, B.J., & Berliner, D.C. (2002). Small class size and its effects. *Educational Leadership, 59*(5), 12-23.

Blatchford, P. (2003). A systematic observational study of teachers' and pupils' behaviour in large and small classes. *Learning and Instruction*, *13*(6), 569-595.

Blatchford, P., Bassett, P., Goldstein, H., & Martin, C. (2003). Are class size differences related to pupils' educational progress and classroom processes? Findings from the institute of education class size study of children aged 5–7 years. *British Educational Research Journal*, *29*(5), 709-730.

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. *Handbook of child psychology*.

California Voter Guide (1998). *Proposition 8: Class Size Reduction Funding.* California: Secretary of State. Retrieved from http://vote98.sos.ca.gov/VoterGuide/Propositions/8noarg.htm

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

Chingos, M. M. (2012). The impact of a universal class-size reduction policy: Evidence from Florida's statewide mandate. *Economics of Education Review*, *31*(5), 543-562.

Coburn, C. E. (2004). Beyond decoupling: Rethinking the relationship between the institutional environment and the classroom. Sociology of Education, 77(3), 211-244.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 417-446.

Dee, T. S., & West, M. R. (2011). The non-cognitive returns to class size. Educational Evaluation and Policy Analysis, 33(1), 23-46.

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. American Economic Journal: Applied Economics, 1(3), 111-134.

Downer, J. T., Rimm-Kaufman, S. E., & Pianta, R. C. (2007). How do classroom conditions and children's risk for school problems contribute to children's behavioral engagement in learning? *School Psychology Review*, *36*(3), 413.

DuGoff, E. H., Schuler, M., & Stuart, E. A. (2014). Generalizing observational study results: Applying propensity score methods to complex surveys. *Health Services Research*, *49*(1), 284-303.

Duncan, G. J., & Magnuson, K. (2011). The nature and impact of early achievement skills, attention skills, and behavior problems. In G.J. Duncan & J. Murnane (Eds.), *Whither opportunity (*47-70). New York: Russell Sage Foundation.

Education Commission of the States (2010). Class Size Policies. Retrieved from http://www.ecs.org/clearinghouse/85/21/8521.pdf

Evertson, C. M., & Randolph, C. H. (1989). Teaching practices and class size: A new look at an old issue. Peabody Journal of Education, 67(1), 85-105.

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

Finn, J. D., Gerber, S. B., & Boyd-Zaharias, J. (2005). Small classes in the early grades, academic achievement, and graduating from high school. *Journal of educational Psychology*, *97*(2), 214.

Finn, J. D., Pannozzo, G. M., & Achilles, C. M. (2003). The "why's" of class size: Student behavior in small classes. *Review of Educational Research*,*73*(3), 321-368.

Fredriksson, P., Öckert, B., & Oosterbeek, H. (2013). Long-term effects of class size. *The Quarterly Journal of Economics, 128*(1), 249-285.

Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. Educational evaluation and policy analysis, 1(1), 2-16.

Goldstein, H., & Blatchford, P. (1998). Class size and educational achievement: A review of methodology with particular reference to study design. *British Educational Research Journal*, *24*(3), 255-268.

Graue, E., Hatch, K., Rao, K., & Oen, D. (2007). The wisdom of class-size reduction. *American Educational Research Journal*, *44*(3), 670-700.

Graue, M. E., & Oen, D. (2008). You just feed them with a long-handled spoon: Families evaluate their experiences in a class size reduction reform. *Educational Policy, 23*(5), 685-713.

Green, K. M., & Stuart, E. A. (2014). Examining moderation analyses in propensity score methods: Application to depression and substance use. *Journal of Consulting and Clinical Psychology*, *82*(5), 773.

Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system (SSRS)*. American Guidance Service.

Grindal, T. (2011). The effects of preschool setting on young children's cognitive skills, social behavior and approaches to learning: A propensity score analysis

(Unpublished qualifying paper). Harvard Graduate School of Education, Cambridge, Massachusetts.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*(3), 234.

Harfitt, G. J., & Tsui, A. (2015). An examination of class size reduction on teaching and learning processes: A theoretical perspective. *British Educational Research Journal*, *41*(5), 845-865.

Hargreaves, L., Galton, M., & Pell, A. (1998). The effects of changes in class size on teacher–pupil interaction. *International Journal of Educational Research*, *29*(8), 779-795.

Head Start Bureau (2005). Head Start centers and use of space. Head Start Design Guide. HHS/ACF/ACYF/HSB. Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/tta-system/teaching/eecd/learning%20environments/planning%20and%20arranging%20spaces/edudev_art_00059_051606.html Last updated: October 2014

Ho, D., Imai, K., King, G., & Stuart, E. (2011). MatchIt: Nonparametric preprocessing for parametric casual inference. *Journal of Statistical Software, 42*(8), 1-28.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706-710.

Jepsen, C., & Rivkin, S. (2009). Class size reduction and student achievement the potential tradeoff between teacher quality and class size. *Journal of human resources*, *44*(1), 223-250.

Jo, B., Stuart, E. A., MacKinnon, D. P., & Vinokur, A. D. (2011). The use of propensity scores in mediation analysis. *Multivariate Behavioral Research*, *46*(3), 425-452.

Johnston, J. M. (1990). Effects of Class Size on Classroom Processes and Teacher Behaviors in Kindergarten through Third Grade. Retrieved from http://files.eric.ed.gov/fulltext/ED321848.pdf

Kagan, D. M. (1992). Implication of research on teacher belief. *Educational Psychologist*, *27*(1), 65-90.

Krueger, A. B. (1999). Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics, 114*(2), 497-532.

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

LoCasale-Crouch, J., Konold, T., Pianta, R., Howes, C., Burchinal, M., Bryant, D., ... & Barbarin, O. (2007). Observed classroom quality profiles in state-funded pre-kindergarten programs and associations with teacher, program, and classroom characteristics. *Early Childhood Research Quarterly, 22*(1), 3-17.

Malone, L.; Carlson, B.L.; Aikens, N.; Moiduddin, E.; Klein, A.K.; West, J., … & Rall, K. (2013). Head Start Family and Children Experiences Survey: 2009 User's Manual. Ann Arbor, MI: Child Care & Early Education Research Connections.

Martin, N. K., Yin, Z., & Mayall, H. (2006). Classroom management training, teaching experience and gender: Do these variables impact teachers' attitudes and beliefs toward classroom management style? Proceedings from *Annual Conference of the Southwest Educational Research Association, 2006*. Austin, TX: ERIC.

McClelland, M. M., Morrison, F. J., & Holmes, D. L. (2000). Children at risk for early academic problems: The role of learning-related social skills. *Early Childhood Research Quarterly*, *15*(3), 307-329.

Moiduddin, E., Aikens, N., Tarullo, L., West, J., Xue, Y. (2012). Child Outcomes and Classroom Quality in FACES 2009. OPRE Report 2012-37a. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Molnar, A., Smith, P., Zahorik, J., Palmer, A., Halbach, A., & Ehrle, K. (1999). Evaluating the SAGE program: A pilot program in targeted pupil-teacher reduction in Wisconsin. *Educational Evaluation and Policy Analysis*, *21*(2), 165-177.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children*, 113-127.

Murnane, R. J., & Willett, J. B. (2010). Methods matter: Improving causal inference in educational and social science research. New York: Oxford University Press.

National Institute of Child Health and Human Development Early Child Care Research Network (2000). The relation of child care to cognitive and language development. *Child Development, 71* (4), 960-980.

National Scientific Council on the Developing Child. (2004). Young children develop in an environment of relationships. Working Paper No. 1. Retrieved from

http://developingchild.harvard.edu/wp-content/uploads/2004/04/Young-Children-Develop-in-an-Environment-of-Relationships.pdf

Odden, A. (1990). Class size and student achievement: Research-based policy alternatives. *Educational evaluation and policy analysis*, *12*(2), 213-227.

Office of the Administration for Children and Families (2015). About Office of Head Start. Retrieved from http://www.acf.hhs.gov/programs/ohs

Pedder, D. (2006). Are small classes better? Understanding relationships between class size, classroom processes and pupils' learning. *Oxford Review of Education*, *32*(2), 213-234.

Peterson, J. L., & Zill, N. (1986). Marital disruption, parent-child relationships, and behavior problems in children. *Journal of Marriage and the Family, 48*(2), 295-307.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System Manual, Pre-K*. Baltimore, MD: Paul H Brookes Pub Co.

Ponitz, C. C., Rimm-Kaufman, S. E., Grimm, K. J., & Curby, T. W. (2009). Kindergarten classroom quality, behavioral engagement, and reading achievement. *School Psychology Review*, *38*(1), 102-120.

Raver, C. (2002). Young children's emotional development and school readiness. *Social Policy Report*, *16*(3), 3-19.

Raver, C. C., Jones, S. M., Li-Grining, C. P., Metzger, M., Champion, K. M., & Sardin, L. (2008). Improving preschool classroom processes: Preliminary findings from a randomized trial implemented in Head Start settings. *Early Childhood Research Quarterly, 23*(1), 10-26.

Richardson, V. (1990). Significant and worthwhile change in teaching practice. *Educational researcher, 19*(7), 10-18.

Roid, G.H., & L.J. Miller (1997). *Leiter International Performance Scale Revised, Examiner Rating Scale (Leiter-R).* Lutz, FL: Psychological Assessment Resources, Inc.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33-38.

Rosenbaum, P.R. (2005). Observational study. In B.S. Everitt & D.C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Volume 3) (pp. 1451-1462). Chichester, United Kingdom: John Wiley & Sons.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, *74*(366a), 318-328.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*(3-4), 169-188.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20-36.

Schanzenbach, D. W. (2014). Does Class Size Matter? *Policy Briefs, National Education Policy Center, School of Education, University of Colorado, Boulder*.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1-27.

Vanderweele, T. J., Hong, G., Jones, S. M., & Brown, J. L. (2013). Mediation and spillover effects in group-randomized trials: a case study of the 4Rs educational intervention. *Journal of the American Statistical Association*, *108*(502), 469-482.

Washington 2014 Voters' Guide (2014). *Washington Class Size Reduction Measure Initiative 1351.* Washington: Office of the Secretary of State. Retrieved from https://wei.sos.wa.gov/agency/osos/en/press_and_research/PreviousElections/2014/General-Election/Pages/Online-Voters-Guide.aspx

Wilson, V. (2002). *Does small really make a difference?* (SCRE Research Report No. 107) *The Scottish Council for Research in Education.* Retrieved from http://www.classsizematters.org/wp-content/uploads/2012/11/107.pdf

Appendices

Appendix A: Student Behavioral Outcome Index

A summary index for student behavioral outcomes was constructed using the following measures: Teacher ratings of children's cooperative classroom behavior, problem behavior, and assessor ratings. Table 10 shows a strong but negative correlation between children's cooperative classroom behavior and problem behavior. The correlation between assessor ratings and the other two teacher ratings was moderate. Table 11 shows the principal components loading for the index. The positive outcomes were loaded positively while the negative outcome was loaded negatively such that a positive value on the index represents good outcomes. The summary index was formed for the outcomes variable, measured in spring 2010, and the outcome baseline covariate, measured in fall 2009, respectively.

Table 10. Correlations among measures of children's behavioral outcomes in spring 2010: Children's Cooperative Classroom Behavior (*COOP*), problem behavior (*PROB*), and assessor ratings (*LEITER*).

	COOP	***PROB***	***LEITER***
COOP	1.00		
PROB	-0.65	1.00	
LEITER	0.28	-0.28	1.00

Table 11. Principal components analysis of the measures of children's behavioral outcomes in spring 2010: Children's Cooperative Classroom Behavior (*COOP*), problem behavior (*PROB*), and assessor ratings (*LEITER*).

Component 1	
Eigenvalue	1.81
Proportion of variance explained	0.60
Principal component loadings	
COOP	0.65
PROB	-0.65
LEITER	0.40

Appendix B: CLASS Index

Table 12. Principal components analysis of the measures of the CLASS domains measured in spring 2010: Instructional Support (*IS*), Emotional Support (*ES*), and Classroom Organization (*CO*).

Component 1	
Eigenvalue	1.84
Proportion of variance explained	0.61
Principal component loadings	
IS	0.52
ES	0.52
CO	0.60

Appendix C: Variables List and Description

Selection Variables

The selection variables used in this study are described below.

A. Child Characteristics.

Age at start of school term. A continuous variable indicating the child's age in months at the start of the school term in Fall 2009.

Child gender. A dichotomous variable coded 1 if the child was female and 0 if the child was male.

Child race. A categorical variable indicating whether the child was white, African American, Hispanic or other race. These were coded as a series of dummy variables, with the category of interest coded 1, 0 otherwise. White was set as the reference category.

Early Head Start. A dichotomous variable coded 1 if the child participated in Early Head Start.

Non-Head Start care arrangements. A dichotomous variable coded 1 if the child received childcare before or after Head Start classes, 0 otherwise.

IEP. A dichotomous variable coded 1 if the child had an Individualized Education Plan or Individual Family Service Plan in Fall 2009, 0 otherwise.

Health insurance. A dichotomous variable coded 1 if the child had a Health Insurance Plan, 0 otherwise.

Regular healthcare provider. A dichotomous variable coded 1 if the child had a regular healthcare provider.

Low birth weight. A dichotomous variable coded 1 if the child had low birth weight (below 5.5 lbs), 0 otherwise.

B. Parent Characteristics.

Mother's education. A categorical variable indicating whether the mother's highest level of education was (i) less than a high school diploma; (ii) high school diploma, GED, or vocational/technical diploma; or (iii) bachelor's degree or higher. These were coded as a series of dummy variables, with the category of interest coded 1, 0 otherwise. "Less than a high school diploma" was set as the reference category.

Mother's employment status. A categorical variable indicating whether the child's mother was employed full-time, part-time, looking for work, or not in the labor force. These were coded as a series of dummy variables, with the category of interest coded 1, 0 otherwise. "Not in the labor force" was set as the reference category.

Parents' county of origin. A dichotomous variable coded 1 if both parents were not born in the U.S., 0 otherwise.

Parent's depression score. A continuous measure of the interviewed parent's depression score. This was calculated based on the parent's response to 12 items on the interview, each scored on a scale of 0 to 3 for a total score range of 0 points (not depressed) to 36 points (severely depressed). The FACES drew the items from the Center for Epidemiologic Studies – Depression scale [CES-D] (Malone et al., 2013).

C. Household Characteristics.

Single parent household. A dichotomous variable coded 1 if the child was in a "not two-parent household", 0 otherwise (Malone et al., 2013).

Ratio of income to poverty. A dichotomous variable coded 1 if the household's ratio of income to poverty was below 100% of federal poverty threshold.

On multiple assistance program. A dichotomous variable coded 1 if the household received multiple assistance such as welfare, TANF, general assistance, food stamps, energy assistance etc., 0 otherwise.

Household size. A discrete continuous measure of the total number of household members.

Multiple moves. A dichotomous variable coded 1 if the child's family moved twice or more in the past year, 0 otherwise.

English spoken at home. A dichotomous variable coded 1 if the primary language spoken to the child at home was English, 0 otherwise.

D. Neighborhood Characteristics.

Neighborhood crime. A dichotomous variable coded 1 if the household member had witnessed and/or was a victim of violent/non-violent crime in the neighborhood.

E. Program Characteristics.

Program wait list. A dichotomous variable coded 1 if the Head Start Program had a wait list, 0 otherwise.

Program challenges index. A standardized index measure of the challenges faced by the program. This was formed by performing a Principal Components Analysis on 11 items where the program director indicated whether each item made it harder for him/her to do his/her job well in areas such as time constraints, lack of funds, lack of qualified staff, staff turnover, lack of parental support, challenging population etc. The

index was formed from the first component, which was then standardized to have a mean of 0 and standard deviation of 1 in the dataset.

Expanded Head Start program. A dichotomous variable coded 1 if the Head Start program was expanded in the past year, 0 otherwise.

Director Head Start years. A continuous variable indicating the number of years the director had been working with the Head Start program.

F. Classroom Characteristics.

Class hours per week. A discrete continuous measure of the number of hours the class meets per week.

Teacher education (Bachelor's or above). A dichotomous indicator coded 1 if the teacher had an associate or bachelor's degree, or higher, 0 otherwise.

Design Covariates.

Class weight. The Fall 2009 class weight provided in the Head Start FACES dataset.

Covariates

For all three research questions, I used the classroom, teacher, and program covariates described below. For RQ1 and RQ2b which involved children in the analysis, I included the child covariates described below. Except for the child's age at assessment of outcomes which was measured in spring 2010, all other covariates were measured in fall 2009.

The child covariates were obtained from parent interviews or direct child assessments, teacher and classroom covariates were obtained from teacher interviews,

while the program selection variables were obtained from Head Start program director interviews.

A. Classroom Covariates.

Number of teachers. A discrete continuous measure of the number of teachers in the classroom, recorded during the classroom observation.

Class hours per week. A discrete continuous measure of the number of hours the class meets per week.

B. Teacher Covariates.

Teacher depression score. A continuous measure of the interviewed teacher's depression score. This was calculated based on the teacher's response to 12 items on the interview, each scored on a scale of 0 to 3 for a total score range of 0 points (not depressed) to 36 points (severely depressed). The FACES drew the items from the Center for Epidemiologic Studies – Depression scale [CES-D] (Malone et al., 2013).

Teacher education. A dichotomous indicator coded 1 if the teacher had an associate or bachelor's degree, or higher, 0 otherwise.

C. Program Covariates.

Program challenges index. A standardized index measure of the challenges faced by the program. This was formed by performing a Principal Components Analysis on 11 items where the program director indicated whether each item made it harder for him/her to do his/her job well in areas such as time constraints, lack of funds, lack of qualified staff, staff turnover, lack of parental support, challenging population etc. The

index was formed from the first component, which was then standardized to have a mean of 0 and standard deviation of 1 in the dataset.

D. Child Covariates.

Assessment age. A continuous measure of the child's age in months at assessment during Spring 2010.

Baseline score. The child's baseline score for the behavioral outcomes index, obtained in Fall 2009. The index is formed by taking the principal components of the standardized outcome variables (Cooperative classroom behavior, problem behaviors, and Leiter-R assessor ratings), as described in the main paper.

Child gender. A dichotomous variable coded 1 if the child was female and 0 if the child was male.

Child race. A categorical variable indicating whether the child was African American, Hispanic or other race. These were coded as a series of dummy variables, with the category of interest coded 1, 0 otherwise. Other race was set as the reference category.

IEP. A dichotomous variable coded 1 if the child had an Individualized Education Plan or Individual Family Service Plan in Fall 2009, 0 otherwise.

Single parent household. A dichotomous variable coded 1 if the child was in a "not two-parent household", 0 otherwise (Malone et al., 2013).

Ratio of income to poverty. A dichotomous variable coded 1 if the child lived in a household reported to have a household income-to-poverty ratio below 100% of poverty threshold, 0 otherwise.