Class Size and Relationships that Occur during Instruction

by

Kolby Gadd

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Educational Studies)
in the University of Michigan
2020

Doctoral Committee:

Professor Deborah Loewenberg Ball, Chair
Professor Stephen L. DesJardins
Professor Brian M. McCall
Dr. Mark White, Universitetet i Oslo

Kolby Gadd

kjgadd@umich.edu

ORCID iD: 0000-0002-7125-5030

ACKNOWLEDGMENTS

I, like others who have completed a dissertation, have followed a long road to reach this point. For me, the clearest beginning to this road was over 20 years ago when I spent a summer in middle school doing manual labor for pay in my neighborhood. I was basically unsuccessful doing this labor, but I resolved at that time to obtain an education and do work that required the use of my brain more so than my body. I am grateful for the experiences I have had and those who have supported me in my pursuit of formal education. These acknowledgments are mostly about those who have supported me and constitute an effort on my part to express appreciation.

My parents expressed confidence in my capabilities and never pushed me firmly in any particular direction. I think this relationship helped me be curious and believe I can more or less do what I want to do and trust my own decision making. They also raised me in a faith that endorsed obtaining as much education as possible. I am grateful to have had that foundation to grow into an adult.

As I have pursued my education, no person has provided more support than Jamie, my spouse since sophomore year of undergraduate studies. You have always respected my decision to be a full-time student despite the inconvenience and lack of luxury that goes along with it. I am amazed we have been able to prioritize and manage our lives in order to accomplish family goals during the years I have been a student. We have always had everything we need and many things we want. Your dedication and disposition have made that possible. I am grateful to share life with you.

I have been a parent since my time as an undergraduate student, and I am grateful to have studied education while raising children. My children have helped me learn about myself in ways that have been challenging and valuable. Thank you kids for being playful, forgiving, thoughtful, and you.

Many relationships outside my family have also made it possible for me to obtain my education. Christine Walker, my teacher for methods of teaching mathematics, expressed enthusiasm in my potential as a teacher and helped me develop my ideas about teaching as an undergraduate student. Christine also facilitated first contact with both people who became my graduate school advisors. First, Doug Corey visited Dr. Walker's class and talked about the potential to do academic work in mathematics education. Second, Dr. Walker encouraged me to attend a talk to be given by Deborah Ball. Both of these experiences planted in my mind the potential to work on improving mathematics education in ways that could reach beyond the students I directly teach. I am grateful for exposure to big ideas that helped me imagine paths of interest I might pursue.

I had wonderful mentors—Travis Lemon, Dawn Barson, and Pam Dallon—as a young teacher who helped me think about how instruction might cultivate students' sense making and brilliance. Their example of being thoughtful practitioners who are dedicated to incorporating student thinking into their instruction has offered me concrete examples for a frame of reference as I have ventured into the academic world.

Inasmuch as I was well prepared for my doctoral studies at the University of Michigan, it was in large part due to the rigorous program I completed as a master's student. My preparation to engage with academic research is thanks to the high expectations and patience of faculty members such as Doug Corey, Keith Leatham, and Dan Siebert. I am especially grateful for the time Dr. Corey took with me to have informal conversations about my budding ideas about education research.

Thank you to Deborah Ball for serving as my advisor throughout my time at U of M. I have faced a set of challenges during my time as a doctoral student that most likely would have prevented me completing the degree without your compassion. I came to Michigan to because I wanted to learn from your experience and perspective regarding mathematics instruction and policy. I have had success at Michigan because of who you are as a person. Thank you.

I had gracious committee members who surprised me with their willingness to help. I knew I had selected people I wanted to work with, and the level of attentiveness you gave to my requests exceeded anything I had anticipated. Your support made the dissertation a tremendous learning opportunity for me.

I found an office home in the School of Education 1600 Suite. Special thanks to Merrie, Aileen, and Meghan for providing a pleasant place to work all these years.

In my first year as a doctoral student, I learned the School of Education planned to open the Center for Education Design, Evaluation, and Research (CEDER). I thought the work to be done at CEDER sounded interesting and am grateful to have connected with Vicki Bigelow to pick up work on an hourly basis over the years. Time at CEDER shaped my ideas about what I would be interested in doing after completing the degree, which provided a measure of motivation to complete the degree.

I have benefited from many peer relationships during my time as a graduate student. Most prominent among these is my good fortune to work with Joy Johnson. You have shown great empathy to me and so often made the drudgery of graduate student work tolerable. Graduate work is isolating, but you consistently helped me not feel alone.

Reflecting on relationships and other aspects of life that have come together to make it possible for me to finish the dissertation inspires a degree of awe. I am thankful for the support I have had and the work I have been able to do. As I acknowledge these opportunities I have had, I find it appropriate to also acknowledge that many people do not have comparable opportunities. Disparities in opportunities exist in our society along lines of race, sex, ethnicity, SES, sexual orientation, and numerous other factors outside anybody's control. These disparities have given me advantages relative to other people of equal capacity and desire.

This acknowledgement of my privileged position leads to a concluding acknowledgment of responsibility. I feel a sense of responsibility to use whatever I have gained to contribute to the dream of creating an inclusive, just, and equitable society for all people. To all who have supported me, you have helped me desire for everyone to have the support they need, too.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Class size has long been a concern among education stakeholders. Although assumptions about the value of smaller class sizes abound, existing research does not offer clear conclusions about the effects of class size on learning. With respect to instruction, researchers have principally focused on differences in teachers' practice or the enacted curriculum as a function of class size. In this study, I focused on whether and how class size impacts relationships between teachers and students as well as relationships among students. Relationships are an integral part of instruction. This study sought to analyze empirical evidence about the relationship between class size and aspects of instruction that education stakeholders might naturally wonder about.

To study this aspect of the relationship between class size and instruction, I used data from the Early Childhood Longitudinal Study (ECLS). This dataset comprises a representative sample of children who attended elementary school in the United States. These data thus provide information about class size on a national level for an age at which many researchers suspect the effects of class size might be the most prominent (see, e.g., Finn and Achilles (1990)). Measures of relationships between teachers' and students and among students available in the ECLS data include surveys of both teachers and students. Based on the nature of these surveys, I selected analytic methods to, as much as possible, control for relevant factors that are unobserved in the data. Specifically, I used linear regression and Bayesian Additive Regression Trees for cross-sectional data, and I used event history analysis and fixed-effects linear regression for longitudinal data. Using these models, I examine the average relationship between class size and instruction as well as the possibility of heterogeneous relationships based on race, ethnicity, sex, and family SES.

Results from this analysis reveal some significant relationships between class size and instruction. These significant results include differences in the relationship between class size and the relationship between teachers' and students among males and females. Although these statistically significant relationships are small enough to be of questionable practical relevance, they are suggestive of areas that merit deliberate research in the future to understand the relationship between class size and instruction. Future research could build on the findings of this study in multiple ways. Qualitative researchers might, for example, use findings from this study to find areas where it is valuable to uncover details about the experiences of students and teachers in larger and smaller classrooms. Quantitative researchers might learn from the results of this study to design scales that are more likely to detect the effects of class size than the scales I used in this study.

CHAPTER 1

Introduction

When I dropped my oldest child off at school on the first day of kindergarten,[1] there was lots of talk among parents about the size of the class. There were two kindergarten teachers at the school, and each class had more than 25 students. Many parents wondered if the classes were too large. About half way through the school year, a rumor spread among parents that the classes were indeed too large and that the school was considering hiring another kindergarten teacher in order to make three classes with less than 20 students in each class. While another teacher did not begin mid-year, a third kindergarten teacher was present at the school for the following year. When my next oldest child attended kindergarten at the same school, each of the three classes had less than 20 students.

Though arguably unjustified by research, people in the United States have consistently supported class size reduction. My preceding story is an illustration of the role class size can play in the minds of education stakeholders. Perhaps because class size is a visible and concrete feature of classrooms, it seems to appear often as a policy lever to improve classrooms. Indeed, a set of nationally representative surveys (*Education Next/PEPG Surveys*, 2020), for example, reveals that a plurality of respondents favored using funds to reduce class size over increasing teacher salaries or purchasing educational resources. These surveys provide evidence that class size reduction is a popular policy among stakeholders to implement in American schools.

Support for class size reduction may be driven by the ease with which one can imagine instructional improvements that could follow. One way instruction could change is related to teacher practices such as individual attention for students. For students who flounder with academics and remain practically unnoticed in large classes, class size reduction offers a common sense strategy to deliver needed attention to individual

[1] Incidentally, this was also my first day as a PhD student at the University of Michigan

students and improve academic achievement. A second example of teacher practices that could differ in smaller rather than larger classes pertains to the content itself. Teachers in smaller classes may be able to more efficiently teach basic content and use the balance of the time to help students work on more complex ideas.

Another way instruction might improve in connection with class size reduction could be more or less separate from teacher practices. Specifically, teachers and students could engage in similar activities in all classes, but teachers and students in smaller classes might have stronger interpersonal relationships than students in larger classes. Students in smaller classes might, for example, be more comfortable sharing their thinking during whole class discussion compared to students in larger classes. In this example, the activities in larger and smaller classrooms is the same, but students' experience is different.

Nevertheless, instruction may remain the same regardless of class size. Cohen, Raudenbush, and Ball (2003) described class size as a resource that, like other resources, teachers and students may or may not use during instruction. Teachers, for example, might employ the same instructional strategies in small and large classes. Indeed, Hoxby (2000), reflecting on finding no relationship between class size and student achievement, hypothesized that teachers' practice is robust to transient changes in class size. In order for changes to occur in instruction, teachers may need both training and motivation to take advantage of class size reduction.

Changes in instruction that occur as a result of class size reduction can also be negative. Smaller classes are ostensibly more likely than larger classes to produce homogeneous thinking among students. As such, smaller rather than larger classes may be undesirable for teachers who rely on diversity of student thinking to facilitate engaging and productive discussions. Another potential negative effect of class size reduction on instruction is associated with the risk of diluting teacher quality (Hanushek, 1998). Because class size reduction requires more teachers, policy implementation that creates a shortage in the supply of teachers can force schools to hire new teachers who are less capable than previously hired teachers to support meaningful instruction (ibid.).

For now, the ways in which instruction may or may not systematically change in response to class size reduction are unknown. Researchers who have studied the effects of class size reduction have primarily focused on student achievement on standardized tests and other academic or life outcomes (Dee & West, 2011). As a whole, these studies raise questions about the effectiveness and cost efficiency of class size reduction. Nonetheless, enthusiasm for class size reduction persists among many education stakeholders in the United States. For this reason, I focus this study on an examination of the relationship between class size and instruction.

CHAPTER 2

Framework

Theory

In this section, I provide an overview of the major elements of this study and the theoretical relationship among them. Specifically, I outline a framework for instruction and provide a definition of class size. Then I review theories about the relationship between class size and instruction and situate my work within these theories.

Instruction

My framework for instruction in this study is based on the instructional triangle (Ball, 2018; Cohen et al., 2003) as represented in Figure 2.1. The authors of this framework posit that instruction is not the work of the teacher alone. Rather, instruction is co-created by teachers and students as they work together on content. Teachers, students, and content constitute the principal actors in instruction and are represented by the vertices of triangle in the figure. During instruction, these actors interact with each other to create several concurrent relationships. These relationships are between (1) teacher and stuff (often academic content), (2) teacher and students, (3) students and stuff, and (4) students and students.

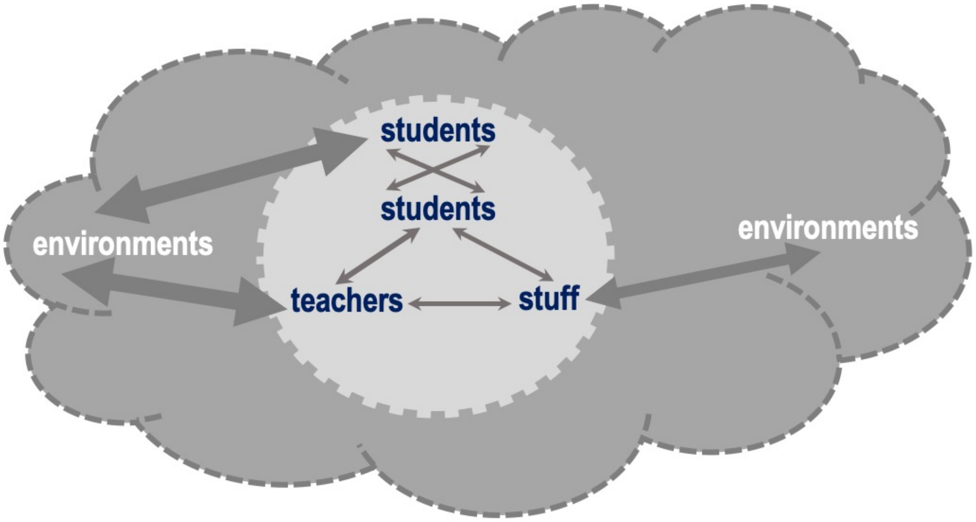


Figure 2.1: The instructional triangle as revised from Cohen, Raudenbush, and Ball (2003) by Ball (2018).

Given the instructional triangle as a framework, I view the study of instruction as the study of actors and relationships within the instructional triangle. As such, what students and teachers do as well as what stuff they work on are examples of relevant topics for the study of instruction. Beyond exclusive focus on students, teachers, and content, other relevant topics are the nature of (1) students' engagement with stuff, (2) their relationship and interactions with the teacher and other students, and (3) teachers' interactions with stuff. Further, the relationship between any of these topics is relevant for research on instruction.

Ball's (2018) revisions to the instructional triangle highlight a few points relevant to my study. First, students and their relationships with each other are positioned at the top vertex of the triangle. This positions students as a central focus within the framework. After all, schools and instruction exist for the benefit of students. Second, the environment is expansive and the boundary between the environment is penetrable. These features of the diagram emphasize that instruction occurs in a social historical, cultural, and political environment that unavoidably influences all happenings in the classroom. Third, teachers and students interact regarding 'stuff.' This term is more inclusive than a term such as 'content' in that it suggests that teachers and students interactions matter for instruction even when they are not explicitly working on academic content. Using this more inclusive term does not diminish the need for careful thinking about content. Instead, it elevates non-academic interactions teachers and students have about any topic (i.e., “stuff”) to an appropriate level of influence on instruction within the framework.

Ball and Forzani (2007) argued for the study of instruction when they proposed a distinction between research *related* to education and research *in* education. Whereas all that is necessary for research to be related to education is an educational setting, research in education requires a focus on the interaction between teachers, students, and content. To provide a couple concrete examples, a study of the relationship between teacher credentials and student achievement would be research related to education, and a study of the relationship between teacher credentials and the content taught in the classroom would be research in education. Because instruction is an integral part of education, research *related* to education is ancillary to research *in* education.

Ideas about important areas of study within instruction come from a variety of authors. To begin, studying instruction can reveal the extent to which experiences in the classroom connect with students knowledge, background, and culture. Cohen (2011) explained how teachers who are more acquainted with student knowledge are able to tailor classroom activities in a way that engages students in learning. Ladson-Billings (2009) described culturally relevant pedagogy as a way to embrace differences and create learning experiences for students that are relevant to their lives. Efforts in the classroom to understand who students are and what motivates them are an important way to promote participation and engagement in academic learning among all students.

Another aspect of instruction researchers might study is the sense of community and belonging among students and teachers in the classroom. Ladson-Billings (2009) referred to culture in successful classrooms with African-American children as a family. Inherent in this culture is responsibility not only for individual work but also for the work of the entire group. Community motivates students to both offer and receive support when necessary to ensure the success of the group. Sense of community and the value of each student is also apparent in Shalaby’s (2017) observations of four students. Shalaby argued that the struggles some students have in school can be attributed to adverse conditions within schools rather than solely to the students themselves. That is, emphasis in schools on conformity might position some students as “troublemakers” because classroom instruction is, in many ways, not for them. Inasmuch as instruction is more inclusive of personality and culture and promotes notions of supporting and caring for one another, it is more likely to help students thrive both academically and as humans.

A commonality among these ideas about important aspects of instruction is attention to the relationships among students and between the teacher and students. That is, acquaintance with students' knowledge, culturally relevant pedagogy, and broadening the group of students who benefit from classroom instruction all have a foundation in relationships among people. Accordingly, I apply my focus in this study to relationships within the instructional triangle and the activity of students in the classroom.

Conceptualizing instruction in terms of the relationships that occur during instruction makes it possible to unpack the complex work involved in leading instruction as a teacher. To illustrate this complex work in terms of relationships found in the instructional triangle, I close this subsection with a description of the work teachers do and an argument for why a deep understanding of instruction in research is needed.

Teachers must establish and maintain a productive relationship with a group of young people with whom they spend at most several hours most weekdays for roughly nine months. This endeavor is fraught with challenges. As Cohen (2011) explains, teachers and students generally do not choose to work with one another, so they may not naturally desire to enter a working relationship. The temporary nature of the relationship is another barrier teacher and students must overcome, for one party may determine that the effort necessary to cultivate a productive relationship is not worthwhile given the limited time that the relationship will exist. Moreover, a teacher's effort to cultivate a relationship with one student may undermine the teacher's relationship with another student. Distributing attention among students may be, at times, a zero sum game, and taking time to meet the needs of one student may marginalize another. Further complicating matters, students have relationships with one another, and these relationships can influence the teacher's relationship with individual students. A teacher, for example, may grant certain privileges to some students but not others because of students' choices in the classroom. These other students may become jealous or resentful of the students who received privileges, and, in turn, they may develop negative feelings toward the teacher for withholding privileges from them. Regardless of the teacher's justification for extending privileges to some students and not others, he or she strengthens his or her relationship with some students and simultaneously weakens it with others. Each other aspect of the work of teaching (managing content and managing how students interact with the content and one another) is similarly fraught with extensive challenges.

As challenging as each aspect of the work of teaching is, the demands of these separate aspects of the work of teaching are not the greatest difficulty during instruction. Rather, the greater difficulty in carrying out the work of teaching during instruction is to coordinate work across every aspect of the work of teaching simultaneously and in real time without compromising any of these aspects. Common classroom situations often create tensions for aspects of the work of teaching that seemingly require teachers to privilege one aspect over another. Herbst (2003) situates these tensions within the didactical contract, which teachers and students continuously negotiate as they work together on mathematics.

Consider, as an example of the tension on multiple aspects of the work of teaching teachers may experience, a teacher who endeavors to teach students to set up and solve single-digit subtraction problems arising within a story problem. Students may experience frustration identifying the minuend and subtrahend in the story problem. Students may project their frustration onto the teacher, who they may believe is content for them to be unsuccessful. This situation could erode the relationship between the teacher and students. In an effort to preserve this relationship, the teacher may tell the students to use the larger number in the story problem as the starting value and to take the smaller number away from the larger number. This strategy may enable the students to successfully set up and solve the problem at hand, but students will encounter single digit subtraction problems that arise within a story problem in the future to which they cannot correctly apply this strategy. Thus, the teacher has allayed students' frustration to build the teacher-student relationship, but the teacher has also compromised the mathematical content of the task. If, however, the teacher had maintained the integrity of the mathematical task, the teacher-student relationship would have been subject to further damage.

The work of teaching consists of many dilemmas that require specialized skill and knowledge to resolve. Mathematical Knowledge for Teaching (MKT) is one way researchers have described this specialized knowledge (Ball, Thames, Phelps, & others, 2008; Silverman & Thompson, 2008). Even teachers who have specialized knowledge may not deploy that knowledge in their instruction (Hill et al., 2008). To deploy this knowledge in instruction, teachers need the skill to communicate with students in way that helps students develop understanding of mathematical ideas.

In the preceding example, the teacher may have been able to maintain both a productive relationship with students and the integrity of the mathematical task with knowledge of a variety of problem solving strategies and a way to communicate these strategies to students. Of course, there is no strategy the teacher could suggest that will certainly help students understand how to set up and solve single digit subtraction problems that arise in word problems. Nonetheless, a teacher who can suggest a robust set of strategies for students to consider is more likely to carry out the work of teaching in a way that honors each aspect of the work than a teacher who resorts to suggesting mnemonics or tricks.

Specialized knowledge and skill, however, cannot solve all dilemmas that are inherent to the work of teaching. McDonald et al. (2014) underscore the role of judgment in the work of teaching. Teachers use judgment, for example, to balance the needs of one student with the needs of the entire class. A teacher may recognize that a particular student needs his or her attention. The teacher could provide this attention by working with the students individually. While meeting with the student individually, however, other students or the entire class may also need the teacher's attention. In this case, the teacher may choose to circulate among all students individually or to address students' needs in a whole group setting. To do the work of teaching in either of these instructional formats will require knowledge and skill of the teacher, but the work of teaching to select the instructional format is mostly a matter of judgment.

Finally, teachers carry out the work of teaching within an environment in which they are prone to feel a sense of responsibility for doing the work in particular ways. Lortie (1975) explains how teachers enter the profession with more or less clear ideas about the nature of classroom instruction based on their lifelong experiences as students in classrooms. Herbst, Nachlieli, and Chazan (2011) offer evidence that teachers have a sense for norms and expectations regarding instructional choices. Although teachers are not necessarily obliged to act on their inclinations to reproduce the types of instruction they experienced as students and are normative for the profession, they may need to act in deliberate ways to disrupt this reproduction.

Reproduction of instruction that has historically existed in American schools is problematic for students of color and economically disadvantaged students who have historically been marginalized in American schools. Teachers in the United States are most often white middle-class women. If these teachers are inclined to do the work of teaching in a way that reproduces their own classroom experiences and expectations, they are unlikely to meet the needs of students who are in the greatest need of valuable school experiences.

In order to disrupt the reproduction of pedagogy that is biased against students of color and economically disadvantaged students, researchers must develop a deep knowledge of the work of teaching. Unless researchers can identify and describe the work teachers need to do to meet the needs of marginalized students, researchers will be unlikely to identify the knowledge and skill teachers need to do this work. Without this knowledge and skill, teachers will not be prepared to adequately resolve tensions that arise within the didactical contract (Herbst, 2003). If teachers cannot adequately resolve these tensions, they will necessarily compromise the mathematics, their relationship with students, or their ability to help students interact with the content and one another productively. For this reason, I argue that researchers have a critical responsibility to develop knowledge about the work of teaching involved in leading the instruction students experience in schools.

Class Size

In this study, I attend to class size rather than teacher-student ratio. I focus on class size because prior research suggests it has a stronger relationship with changes in instruction and educational outcomes (see, e.g., Finn & Achilles (1990)). Class size is defined as number of students in a classroom without consideration of the number of adults who work in the class; teacher-student ratio, on the other hand, is sensitive to the number of adults such as teacher aides who might work in the classroom in addition to the regular classroom teacher. Classrooms with one adult have a *class size* equal to the *student-teacher ratio*. For classrooms where at least two adults work, however, *class size* is greater than *student-teacher ratio*.

One reason that class size might have a stronger relationship than student-teacher ratio with outcomes of interest is fidelity of implementation. Graue, Hatch, Rao, and Oen (2007) documented some of the ways policies that focus on student-teacher ratio can be compromised. Specifically, when multiple teachers work in a classroom together, they might rotate responsibility for leading instruction and completing clerical work. School administrators might also ask the "extra" teacher to cover for the absence of another teacher in the building. These sorts of actions can effectively double the true student-teacher ratio in a classroom. Because reduction in class size is less exposed than student-teacher ratio to issues that compromise implementation, researchers can identify the treatment effects of class size reduction with greater ease.

Matters relating to how multiple adults interact in the classroom point to an aspect of instruction that is not made visible in the instructional triangle (Cohen et al., 2003). Specifically, when more than one adult is present in the classroom the instructional triangle could be revised to represent an interaction between adults in a manner similar to the representation of interactions among students in Figure 2.1. When two adults are present in the classroom, as is the case when the student-teacher ratio is less than the class size, the adults can interact in ways that affect other components of the instructional triangle. Considering this interaction as part of the instructional triangle helps elucidate the complexity associated with understanding the relationship between student-teacher ratio and instruction. While that relationship is outside the scope of this study, it offers an avenue of valuable research to pursue.

An important consideration pertaining to class size is the extent to which it is distributed (un)evenly among children from diverse backgrounds. For a variety of reasons including population density and economic investment, some groups of students might systematically experience classes of larger or smaller size than their peers along lines of race, ethnicity, geographic location, and family SES. Thus, I contextualize my analysis of the relationship between class size and instruction by first conducting an analysis of which groups of students might be more or less likely to enroll in larger and smaller classes. Such context makes it possible to know which groups of students, if any, are disproportionately exposed to the consequences of class size.

Theorized Relationships Between Class Size and Instruction

Having established a framework for class size and instruction separately, I now outline theory to link them. To begin thinking about the relationship between class size and instruction, an important point to bear in mind is that they need not necessarily be related. Cohen, Raudenbush, and Ball (2003) described class size as a resource that may or may not influence how students and teachers interact with each other. In other words, like any other resource, teachers and students have no obligation to interact in particular ways because of class size. Many instructional strategies (e.g., large group lecture) can happen in large or small classes, so familiarity and habit might lead teachers and students to interact in the same ways regardless of class size. Hoxby (2000) offered a hypothesis along these lines after finding no evidence of a relationship between student achievement and exogenous year-to-year variation in class size.

Nonetheless, there are many ways in which class size and instruction might be related. To that end, researchers have developed multiple theories about how class size might affect instruction. One theory is that small classes reduce the probability that a student disrupts the class (see, e.g., Lazear (2001)). Limiting disruptions has at least two potential benefits. First, fewer disruptions in the classroom preserves more time for learning, which can lead to higher academic achievement for students. Second, fewer disruptions can render the classroom and environment in which teachers work less stressful. Reducing stress for teachers may lead to greater retention of teachers and yield an overall improvement in the education system.

Smaller rather than larger classes might also make fundamentally different forms of instruction possible. Pedagogies that require a small class may be superior for learning some academic content and lead to better achievement for students. If, for example, one-on-one instruction proved important for learning a specific piece of content, smaller rather than larger classes might be necessary to facilitate this instructional format. Beyond a focus on academic learning, some pedagogies that require smaller class size may be necessary to make formal education empowering and relevant for students. hooks (1994), for example, described engaged pedagogy as an approach that emphasizes well being and striving for self-actualization of both the teacher and students. In order to enact engaged pedagogy, hooks explained that a limited class size is necessary because some students can avoid full participation if the class is too large. More than diminishing their own experience, hooks argued that students who avoid full participation compromise learning experiences for the entire class.

Similar to researchers' ideas about important aspects of instruction, these theorized relationships between class size and instruction have a basis in relationships among people in the classroom. That is, student behavior, teacher stress, individual attention, and well being in the classroom are all connected to human relationships that occur in the classroom as a fundamental part of instruction. This commonality motivates my hypothesis for this study, which is that relationships and student activity in the classroom differ among smaller and larger classes. To investigate this hypothesis, I set aside considerations about how these changes might subsequently affect other matters such as the teacher labor market and student achievement. In other words, this project is an endeavor to understand the relationship between class size and the educational experience itself. I leave how such changes might influence other aspects of the educational system for further research.

Background and the Need for Further Research

Researchers have adopted a variety of perspectives to study class size, and I turn now to a review of this research. I begin with research on the relationship between class size and instruction, and outline the knowledge that remains to be created in this area. Then I review research that examines the effects of class size on educational outcomes such as student achievement on end-of-year standardized tests. Although this research is not directly related to the relationship between class size and instruction, I include it in my review for two reasons. First, doing so allows me to address a potential concern about the need to study the relationship between class size and instruction. Second, I use this research to further develop my hypotheses about the relationship between class size and instruction.

After reviewing research relevant to class size, I argue for a somewhat different perspective than is commonly used for studying the relationship between class size and instruction as the motivation for this study. This perspective is based on the idea from the instructional triangle (Cohen et al., 2003) that instruction comprises a set of simultaneously occurring relationships. From the perspective of the instructional triangle, these relationships are at least implied in any study of instruction. For this study, however, these relationships are the explicit motivation and focus.

Prior Research on the Relationship Between Class Size and Instruction

Prior research suggests some ways in which class size might be associated with instructional practices and participation. Stasz and Stecher (2000), for example, found that teachers in classrooms with fewer students spent less time on discipline and more time with individual students compared to their counterparts in classrooms with more students. This finding is consistent with theory that suggests smaller classes have fewer behavior problems and more individual attention than larger classes. Nonetheless, Stasz and Stecher relied on teacher survey data that was representative of schools in California. Because schooling in the United States is largely decentralized, findings from California may or may not be relevant for other states for a variety of reasons and knowledge about what happens on average in the United States remains unknown. Despite these limitations, Stasz and Stecher's work provides some credence to the theory underlying the relationship between class size and instruction.

Other research calls into question the extent to which teachers actually change their practice in larger and smaller classes. Using national, though not representative, data of middle school and high school mathematics teachers, Betts & Shkolnik (1999) found limited evidence that teachers reallocate how they spend their time in instructional activities such as covering new material, discipline, or reviewing material. Moreover, in areas where the authors found evidence of time reallocation, the magnitude was small.

Student activity is another aspect of the instructional triangle that may have a relationship with class size. Dee and West (2011) studied the effect of class size on students' non-cognitive skills in middle school with nationally representative data. Among the measures they used for these skills, a few were relevant to the process of instruction. Namely, these researchers included a student rating of fear to ask questions in particular subjects and teacher ratings of student disruption and attentiveness in class. Dee and West found students in smaller classes reported less fear of asking questions than students in larger classes, but they did not find a relationship with teacher ratings of students.

Qualitative research is consistent with the mixed findings from quantitative research about the relationship between class size and instructional practices and participation. Englehart (2007) conducted observations and interviews with students who were were each enrolled in two classes of varying size. This author found that students were able to return their attention to instruction after distractions in the classroom with greater ease in small classes. Evertson and Randolph (1989), on the other hand, conducted classroom observations in connection with Project STAR and found no evidence of differences in instruction among smaller and larger classes.

Research outside the United States also provides evidence of the relationship between class size and instruction. Peter Blatchford and colleagues have studied the effects of class size on grouping of students within the classroom (Blatchford, Baines, Kutnick, & Martin, 2001) as well as interactions with the teacher and on task behavior (Blatchford, Bassett, & Brown, 2011). Results of this work indicate that students engage in instruction more in smaller rather than larger classes, and this difference in engagement is more pronounced for students with lower academic achievement than their peers. Also, students in smaller rather than larger classes received more individual attention from the teacher. Like other research findings, this empirical evidence is consistent with theoretical ideas about the relationship between class size and instruction. Moving into less theorized territory, Blatchford, et al. (2001) found that larger classes were associated with more and larger groupings of students within a class. Further, these researchers found that classes with less than 25 students were more likely than classes with more than 25 students to work with the whole group. Although there is less theory to guide thinking about decisions teachers make regarding how to group students in smaller and larger classes, one implication of this finding could be that teachers find some classes too large to sustain work with the whole group, so they split the class into smaller groups to simulate a smaller class.

Other international research call into question the relationship between class size and instruction. In an analysis of data from the Third International Math and Science Study (TIMSS), Pong & Pallas (2001) found matters relevant to instruction such as academic content and teacher practices were unrelated to class size. This analysis, however, was hindered by the limited measures of instruction available in the TIMSS data.

Findings from these and other international studies provide more evidence of potential differences in instruction associated with class size, but they provide limited information about class size and instruction in the United States. International research on the relationship between class size and instruction has questionable generalizability to the United States because of significant differences in context. Indeed, Pong & Pallas (2001) found evidence that class size seems to have a unique role in the United States. Even among locales with many salient similarities such as the United Kingdom and the United States, differences in curriculum, teacher training, and locus of administrative control of schools raise questions about external validity of international results.

In existing literature on class size, studies of its association with relationships that occur during instruction is, to my knowledge, limited at best. Indeed, one could describe much of the existing research of class size as *related* to education rather than *in* education (Ball & Forzani, 2007). For the portion of research *in* education, it often pertains to teacher practices such as time spent on particular activities. This research thus not only offers an opportunity to add to knowledge about class size *in* education, it also offers a perspective on instruction often not taken up by researchers.

To this point, I have established that prior research leaves unresolved important questions about the relationship between class size and instruction in the United States. This gap in the literature, however, is insufficient to justify further study, for a substantial amount of research on class size suggests that the effect of class size on educational outcomes is uncertain and probably inefficient with respect to cost. A possible implication of this research is that any effects of class size on instruction are irrelevant. Therefore, I pause to consider the merits of this research and a potential argument against studying the relationship between class size and instruction.

Class Size from the Perspective of the Education Production Function

A common perspective researchers have used to study class size is known as the education production function. From this perspective, education is a process that takes a set of inputs and produces a set of outputs. Researchers typically operationalize inputs as quantities such as teacher's years of experience, per pupil school expenditures, and curriculum materials; they operationalize outputs as quantities such as students' end-of-year test scores, college going, and lifetime earnings (see, e.g., Greenwald, Hedges, & Laine (1996)). In this subsection, my goal is to provide a summary of research on class size from the perspective of the education production function, which establishes an argument that the effect of class size on educational outputs is uncertain and perhaps inefficient with respect to cost. I reserve discussion of the limitations of this research and its conclusions for the following subsection in which I complete an argument for studying the relationship between class size and instruction.

Within the perspective of the education production function, researchers have used experimental and quasi-experimental methods of analysis to isolate exogenous variation in class size and measure the effect of class size on various educational outcomes. Throughout this study, I use the word 'outcome' in a special way. Specifically, I use 'outcome' to refer to the result or benefit that comes *after* an experience. As an example, researchers might conceptualize student achievement on an end-of-year standardized test as an educational 'outcome' that is the result of an experience–participating in classroom instruction for an academic year. Although the perspective of the education production function does not preclude researchers from defining output as the experience itself, such a definition is, to my knowledge, rare. With that definition established, I use the remainder of this subsection to present the equivocal evidence from education production function literature regarding a link between class size reductions and improved educational outcomes.

A prominent experiment that showed a positive relationship between class size reductions and educational outcomes was project STAR (Student-Teacher Achievement Ratio), which took place in Tennessee in the 1980s (Finn & Achilles, 1990). For project STAR, researchers randomly assigned kindergarten students to a small class, a regular class, or a regular class with an aide. Students remained in this class structure through third grade. As class size study that successfully implement a randomized control trial research design at a large scale, findings from project STAR have a high degree of validity and reliability. In the short run, results of the STAR project indicated that students who were randomly assigned to smaller rather than larger classes had, on average, higher achievement in reading and mathematics as measured by end of year exams (ibid.).

Subsequent research utilizing data from project STAR provides evidence of long run effects of class size on educational outcomes. Dynarski, Hyman, & Schanzenbach (2013), for example, analyzed the effect of small classes in early elementary school on post-secondary educational outcomes. These researchers found that students who attended smaller classes during project STAR were more likely than their counterparts in larger classes to attend and graduate from college. Further, students who attended smaller classes also earned degrees in high-earning fields at a higher rate than others who participated in the study.

Of particular note, findings from project STAR indicate that the effect of class size may be greatest for students who have traditionally been under served in the United States' educational system. Specifically, students of color and low-income students who attend smaller rather than larger classes benefit more from class size reduction than their white and more economically advantaged peers. Although the mechanisms of these heterogeneous results remain unknown, these findings suggest that class size reduction policy could potentially be implemented in such a way as to make schooling in the United States more equitable.

Other researchers have used quasi-experimental methods to find evidence for an inverse relationship between class size and educational outcomes. Angrist and Lavy ((1999)), for instance, took advantage of class size limits in Israeli schools to estimate the effect of class size on student achievement in reading and math using instrumental variables. These authors found that smaller classes led to increases in achievement in fourth and fifth grade, though the effects of class size were smaller in third grade.

In contrast, other researchers have found no evidence of a relationship between class size and educational outcomes. Hoxby (2000) used variation in births to isolate plausibly exogenous variation in early elementary school class sizes. Using this variation, Hoxby found that small, transient changes in class size did not lead to changes in student achievement. Comparing this finding to those from project STAR, Hoxby suggested that STAR findings might suffer from a Hawthorne effect. That is, teachers who participated in project STAR might have felt incentivized by the potential for smaller classes in the future. Thus incentivized, teachers may have increased their effort in smaller classes, and this increase in effort may have been the true cause of increases in student achievement. Given Hoxby's findings and the threat to the validity of findings from project STAR, a null effect of class size on student achievement is plausible.

In addition to possible increased teacher effort, problems with research implementation raise concerns about the reliability of the results of project STAR. Hanushek (1999) described the attrition and treatment cross over observed in the STAR sample. Specifically, less than half of all students who began kindergarten in the experiment remained through third grade, and movement from a regular classes to small classes mid-experiment was more common than movement in the other direction. Such phenomena are common in complex social research and renders assignment to treatment non-random to some extent. Project STAR also had no way to know if teaching quality was randomly distributed among small and regular classes. Hanushek (2003) suggested that, without the ability to check for balance in teaching quality after randomization, variation in teaching quality could help explain why only 40 of 79 small classes outperformed regular classes in project STAR. Altogether, challenges associated with implementing a complex social experiment preclude placing full confidence in project STAR research findings.

Beyond questions about the results of project STAR, other research also calls into question previous findings about the relationship between class size and educational outcomes. In fact, Angrist, Lavy, Leder-Luis, and Shany (2017) replicated the research design of Angrist and Lavy (1999) and found no relationship between class size and achievement. These authors suggest a change in the education production function could be responsible for null effects in later research. They also mention the possibility of a chance finding in the previous research.

Although many individual studies contain evidence that class size reductions are associated with improved educational outcomes, the literature as a whole may show weak or null effects. Hanushek (1997) conducted an expansive literature review and found that teacher-pupil ratio is largely unrelated to educational outcomes for students. One possible reason Hanushek raises for this weak relationship is consistent with Cohen, Raudenbush, and Ball's (2003) ideas about the use of resources in instruction. Specifically, that school inputs such as teacher-pupil ratio may or may not be put to good instructional use. Because of variation in the use of resources in classrooms, Hanushek argues that policy decisions to lower teacher-pupil ratio or adjust any other input to schools are unlikely to improve educational experiences for students.

Other researchers have criticized Hanushek's methods and argued for positive effects of class size on student achievement. Krueger (2002), proposed revisions to Hanushek's method, and these revisions yielded results indicating an inverse relationship between class size and educational outcomes. Greenwald, Hedges, & Laine (1996) claim that Hanushek's "vote counting" approach to meta-analysis is problematic. These authors used a set of stochastically independent samples to analyze the effect of resources on achievement and found an inverse relationship between student-teacher ratio and achievement on standardized tests. In spite of differences in thought regarding the overall relationship between class size and students' educational outcomes, researchers have agreed that class size reduction can benefit some subgroups of students and that the potentially heterogeneous effects of class size need further investigation.

Beyond the question of whether or not class size is associated with improvement in educational outcomes, class size reduction policy may not be a wise use of funds for that purpose. Class size reduction is expensive because, for one reason, it requires more teachers to serve the same number of students. Multiple researchers have argued that alternative interventions could yield similar improvement in educational outcomes for lower cost than class size reduction. Rivkin, Hanushek, & Kain (2005), for example, suggested the benefits of a one standard deviation increase in teacher quality exceed a ten student class size reduction. Next, Dynarski, Hyman, and Schazenbach (2013) explained that the cost to induce a child into college through class size reduction is much greater than the cost to do so through alternative means such as increased access to financial aid for college. Finally, Bowne, Magnuson, Schindler, Duncan, & Yoshikawa (2017) concluded that class size reduction is not a cost effective way to improve learning in early childhood education because gains in learning due to class size reduction only occur at the lowest end of the class size distribution. None of these groups of researchers disputed that class size reduction improves student outcomes. Rather, they proposed that alternative interventions may yield similar improvement in educational outcomes for less money.

One possible reason for mixed findings related to the efficacy and efficiency of class size reduction policy to improve educational outcomes relates to my framework for class size as a resource for instruction. A possible conclusion from research using the education production function perspective is that the effects of class size on instruction are irrelevant because whatever those effects are, they are not, on average, associated with improvement in educational outcomes. Considering class size as a resource teachers and students may or may not use during instruction, however, introduces the possibility that class size has no effect on instruction and, in turn, educational outcomes. Some researchers who have found a lack of a relationship between class size and educational outcomes have noted the mediating role instruction plays in this relationship (Hanushek, 2003; Hoxby, 2000). If class size is a resource that does not affect outcomes in a deterministic fashion, methods designed to measure a single, "true" effect of class on outcomes will necessarily lead to mixed findings depending on how teachers and students used the resource. Because I conceive of class size as a resource rather than an input with a well defined output, I argue for the need to understand how teachers and students use it during instruction. To finish this subsection, I describe in further detail the need to understand the relationship between class size and instruction.

Motivation for this Study of the Relationship Between Class Size and Instruction

My motivation for studying the relationship between class size and instruction relies on the idea that students and teachers derive contemporaneous benefits from instruction that may be related to class size. In other words, I posit that teachers and students can derive meaningful benefits from the educational experience itself and not solely from the outcomes of education. Specifically, I refer to the benefits students and teachers receive from having positive relationships during instruction. Granted, the nature of relationships that occur during instruction most likely influences educational outcomes, but such cascading relationships lie outside the scope of this study as I mentioned in the theoretical framing of the relationship between class size and instruction. For this study, I focus attention on the relationships that occur during instruction, which offers a different perspective than a focus on instructional practices and outcomes that is common in research on class size.

Focusing on the association between class size and relationships that occur instruction brings attention to a rather visceral aspect of schooling. This aspect is that children's experience in school comprises a tremendous portion of their formative years as human beings. Jackson (1990) noted that children in elementary school spend more waking hours in the classroom than in any other space. A reasonable desire for this space is for children to have positive experiences there. My study of the possible association of class size with relationships that occur during instruction is motivated in part by a need to know the extent to which class size might be related to a positive school experience for children.

Though a focus on experience is less common than a focus on outcomes in research relating to class size, focus on experience aligns with priorities of every day life. While children grow as human beings, I posit that parents, caregivers, and society as a whole care about more than the extent to which school prepares children to live full, productive lives in the future. They also care if children live full, productive lives at school. They care if the experience itself is enriching.

Attention to the quality of experiences is apparent in many aspects of life. Research on, for example, spending habits and happiness offers advice such as “buy experiences instead of things,” and “buy less insurance” (Dunn, Gilbert, & Wilson, 2011). Both of these pieces of advice rely on the idea of living in the present. That is, experiences to live and remember generally provide greater contentment than taking possession of items, and protection from potential future adverse events does not carry the benefits people are often inclined to anticipate.

Connecting the idea of living in the present to an educational context, Dewey (1923) warned that “remote aims” provide unsatisfactory motivation for education. That is, educating children in order for them to be productive adults is untenable. Instead, the value in education, according to Dewey, resides in the experience itself. Following this reasoning, a good starting point to understand the importance of class size for education is to focus on the experiences children have during instruction rather than outcomes such as end of year achievement, say, children’s scores on standardized tests or the lifetime earnings those children eventually accrue.

When considering the experiences children have in school, an important point to bear in mind is the uneven distribution of experiences among groups of students. That is, students in American schools often have systematically different experiences based on race, ethnicity, sex, and family SES (e.g., Morton & Riegle-Crumb, 2019). For this reason, a primary motivation for this study is to understand the potentially heterogeneous relationship between class size and instruction.

In this section, I have presented an argument for studying the relationship between class size and instruction. Specifically, I have argued there are many open questions about the relationship between class size and instruction in the United States, and among these open questions is the possible association between class size and relationships that occur during instruction. Further, because of potentially heterogeneous effects of class size on instruction that are strongest for traditionally marginalized students, greater understanding of these effects has potential to reduce inequality and injustice in schools. In the following section, I outline and explain the set of questions I use to study the relationship between class size and instruction.

Research Questions and Hypotheses

To investigate the relationship between class size and instruction, I designed a study based on the following research questions:

1. During early elementary school in the United States, who is enrolled in larger and smaller classes?
2. For students in the United States, what is the observed relationship between class size and instruction?
3. To what extent does this relationship differ along lines of student race, ethnicity, sex, and family SES?

Responding to this set of research questions will build knowledge about who receives the benefits of smaller classes (if such benefits exist) in the United States, what these benefits are, and whether these benefits are more or less pronounced for some students. My purpose for question one is to understand the relationship between socio-demographic characteristics and class size in the United States. Because of the high cost associated with class size reductions, a hypothesis related to this question is that students who have characteristics consistent with affluence and privilege in terms of race, ethnicty, and family SES are more likely than their counterparts to enroll in small classes and less likely to enroll in large classes.

Question two addresses the relationship between class size and instruction and is the central focus of this research. As elaborated in the framework for this study, my view of instruction is grounded in the instructional triangle (Cohen et al., 2003). As such my hypotheses for this research question correspond to components of the instructional triangle. Specifically, my hypotheses pertain to students' relationships with the teacher and each other.

My hypotheses are further grounded in literature that relationships between teachers and students are foundational for building community in the classroom and providing students with positive experience at school (hooks, 1994; Ladson-Billings, 2009; Shalaby, 2017). In focusing on the quality of experiences students have in school, I hold that connectedness with teachers and other students corresponds to a enriching experience. Further, I hold that being seen as competent and productive by teachers and peers is associated with an enriching, positive experience at school.

To articulate specific hypotheses, I separate my ideas into students' relationship with their peers and students' relationship with their teacher. With respect to students' relationships with each other, I hypothesize that class size is related to sense of communmity and shared responsibility in classrooms. For example, depending on the size of the class, students might be more or less comfortable and able to help their peers both with academic work and in other ways. I also hypothesize that class size is related to students basic social interactions–i.e., processing emotions and getting along with others in the classroom. These sorts of interpersonal exchanges, while they may not be academic content, they fall under the category "stuff" as it appears in Ball's (2018) revised instructional triangle. This stuff is an inextricable component of the experience children have in classrooms.

With respect to children's relationship with their teacher, I hypothesize that differences in feelings of affection and warmth exist in relationship with class size. While these feelings may be related to the attention students receive from their teacher, but my hypothesis is distinct from the teacher practice of individual attention. Some researchers who have studied class size have examined the extent to which teachers spend time with individual students. I do not make an effort to study those outcomes in this study. Instead of focusing on the teacher practices associated with class size, I focus on the relationships that occur during instruction. Feelings of closeness and conflict in the relationship between teacher and student, like the relationship among students, is integral to the experience children have in schools.

I use research question number three to focus attention on potential differences that exist in the effect of class size on instruction with respect to student characteristics. One common point of agreement in research on class size from the perspective of the education production function is that the effects of class size differ for students based on their socio-demographic characteristics (see, e.g., Dee & West (2011)). Specifically, research suggests that students who are members of historically marginalized groups derive greater benefits, in terms of educational outcomes, from class size reduction. Inasmuch as these heterogeneous effects exist, the effect of class size on instruction is plausibly different for these students relative to their peers. Therefore, as part of this study, I consider how the relationships between class size and instruction examined in question two might differ for students based on sex, race, ethnicity, and income as well as for students whose identities lie at the intersection of these groups.

CHAPTER 3

Methods

In this section, I explain how I gather empirical evidence to respond to my research questions. I begin this section by establishing the data and measures for this study. Having established these component parts, I then present the analytic methods whereby I use the components to draw inference about the phenomenon of interest. I finish the section by outlining some limitations of the analysis that are important to bear in mind for interpretation of the results.

Data

To respond to my research questions, I use data from the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS) (Tourangeau et al., 2012) collected by the National Center for Education Statistics (NCES). NCES identified a stratified random sample of 20,240[1] students who enrolled in kindergarten in 2010 (ibid.). To identify this sample, NCES first defined a set of geographically contiguous primary sampling units. Then, they randomly selected 90 of these units for inclusion in data collection. Within the selected units, NCES randomly selected schools for participation. At each selected school, NCES randomly selected 20 students for participation in ECLS. Of the 20,240 students selected for participation, 18,170 responded to data collection in the fall or spring of the first year. NCES continued to follow first-year respondents from kindergarten through fifth grade (Tourangeau et al., 2016). Components of data collection for each of these students comprised teacher questionnaires, school administrator questionnaires, parent interviews, and, beginning in third grade, student questionnaires. After the first year, data collection occurred in the spring of each school year for all students, and a subsample of students participated in data collection

[1] All reports of sample size are rounded to the tens digit according to directions from NCES for using restricted data.

during the fall of first and second grade.

For my analysis of ECLS data, I use a weighted sample of students to account for non response. Given these weights, I use the subsample of students for whom valid data on the child questionnaire, parent interview, and teacher questionnaire exists in the fall of kindergarten, first grade, second grade, and third grade. Attrition from the sample of students who participated in kindergarten occurred for a variety of reasons including death and moving outside the United States (Tourangeau et al., 2016). Also, NCES selected a portion of students who moved within the United States to follow, so data collection discontinued for unfollowed students. With these conditions, my analytic sample had 6740 students. Using the sample weights provided by NCES, these students are representative of children who enrolled in kindergarten for the 2010-2011 school year and lived in the United States through the 2013-2014 school year.

Because it is nationally representative of students, ECLS data has benefits relative to other available sources of information that have been or could be used to study class size. To date, most research on class size that has drawn on a representative sample of a population has been limited in scope to, for example, one state in the United States (e.g., Statz & Stecher, Finn & Achilles (1990), and Hoxby (2000)). These limits in scope restrict the potential to understand class size in the United States as a whole. Some large U.S. datasets with measures of instruction and class size (e.g., NCTE, SII, and MET) use non random sampling strategies, which introduces barriers to drawing inference about a population beyond the sample. Thus, potential for generalization to kindergartners throughout the United States is a major benefit of using ECLS data to study the relationship between class size and instruction.

Two features of the ECLS dataset make it particularly useful to investigate ideas relevant to research on class size. First, ECLS is focused on early elementary school, which is the time in school when some have suspected the largest effects of class size exist (see, e.g., Lazear (2001) and Finn & Achilles (1990)). Second, NCES over sampled some segments of the population in order to facilitate comparisons among subgroups. Specifically, NCES over sampled students who are Asian, Native Hawaiian, or Other Pacific Islanders to meet goals for sample size (Tourangeau et al., 2016). This sampling strategy makes possible investigation of potential heterogeneous effects of class size on instruction in response to my third research question. Sample weights are available in the dataset in order to conduct analyses that are corrected for the oversampling of some subgroups.

Table 3.1: Measures of instruction available in the ECLS dataset organized by relevant aspect of the instructional triangle

Aspect of the instructional triangle	Source	Measure
Relationship between teacher and student	Teacher report on STRS	Closeness in relationship with child Conflict in relationship with child
Relationships among students	Teacher report on SSRS	Interpersonal skills Externalizing behaviors Internalizing behaviors
	ECLS Teacher Questionnaire	Participation in ELA achievement groups Participation in mathematics achievement groups
	ECLS Student Questionnaire	Prosocial behavior

Note: STRS refers to the Student Teacher Relationship Scale. SSRS refers to the Social Skills Rating Scale.

Measures

Thus far, I have given an overview of the dataset I will use to respond to my research questions. Now, I provide details about the measures I use for this study. I also explain how this measures connect with the theory and background previously established. I organize this subsection into measures of instruction, class size, and other covariates.

Measures of Instruction

In this subsection, I introduce all measures of instruction I use in this study and situate them within the framework I have established. I also provide relevant descriptive statistics for these measures.

Maintaining a connection with my framework, I categorize measures of instruction available in ECLS by relevant aspects of the instructional triangle (Cohen, Raudenbush, and Ball, 2003). Specifically, these aspects are the relationship between teacher and student and relationships among students. Table 3.1 contains an overview of my measures of instruction by aspect of the instructional triangle. As a note, some aspects of the instructional triangle are not included in this study because measures for these aspects are not available in the ECLS dataset. These relationships have received little attention in prior research, which has mostly focused on the vertices of the instructional triangle as represented in in Figure 2.1 such as teacher practices. My focus on relationships that occur during instruction offers a contribution to existing literature by giving direct attention to relevant aspects of the instructional triangle that can slip into the background of research on instruction.

Some may question my conceptualization of relationships that occur during instruction as measures of instruction. These people may describe relationships as measures of non-cognitive skills or social outcomes rather than measures of instruction. In addition to situating my measures of instruction within the instructional triangle, I also use the framing of the questions in the ECLS surveys to argue for their relevance to instruction. Specifically, ECLS questions are framed as a reflection on what has transpired over a preceding period of time (e.g., the previous one or two months) rather than an assessment at a particular point in time. As teachers and students respond to these questions about their relationships with each other, I posit they draw on their experiences during instruction to formulate responses because classroom instruction is the setting where teachers and students spend a majority of their time together. For this reason, I consider the measures of relationships in the ECLS data to be measures of instruction. That is, they measure the nature of interactions teachers and students have with each other as they work on relevant material (Cohen et al., 2003).

I now provide a description of each measure and explain its connection to the background and theory for this study. As a measure of the teacher-student relationship from the instructional triangle, I use the Student Teacher Relationship Scale (STRS) (Pianta & Steinberg, 1992) from the teacher survey in kindergarten through 3rd grade. Pianta & Steinberg developed this scale to measure teachers' feelings and beliefs about their relationships with individual students, and it reflects the teacher's perception of his or her closeness and conflict with the child. Researchers have conducted extensive analysis of the convergent and divergent validity of constructs measured with the Student Teacher Relationship Scale with related constructs such as student behavior, at-home behavior, decisions about grade retention, student feelings about and adjustment to school, and peer nominations (Birch & Ladd (1997); Doumen et al. (2009); Pianta & Steinberg (1992); others).

On the ECLS questionnaire, teachers responded to a set of 15 questions that combine into two scales to measure perceived closeness and conflict in their relationship with a particular student (see, e.g., Westat (n.d.)). As an introduction to these items, teachers had the prompt, "Below is a series of statements about your relationship with this child. For each statement, please code the category that most applies to your relationship with him or her" (ibid., p. 9). Teachers rated each of the items on a five-point scale, which had descriptors of 'Definitely does not apply,' 'Not really,' 'Neutral, not sure,' 'Applies sometimes,' and 'Definitely applies'. Given that the experiences on which teachers draw to respond to these items primarily occur in the setting of instruction, I consider these scales to be a measure of the relationship between teacher and student in the instructional triangle (Cohen et al., 2003). In other words, when teachers determine, for example, that an item 'applies sometimes' to their relationship with a student, the primary setting in which the item sometimes applies is instruction.

Examples of items on the STRS scale illustrate their plausible relevance to instruction. First, the closeness scale includes prompts such as "When I praise this child, he/she beams with pride" and "This child spontaneously shares information about himself/herself." Next, the conflict scale includes "This child and I always seem to be struggling with each other" and "This child remains angry or is resistant after being disciplined." These examples provide insight into instructional moments that could inform teachers' responses. Specifically, one common occurrence during instruction is for a teacher to tell a student he or she did a good job providing an answer to a mathematics problem or offering an insight into a book the class reads together. Another common occurrence during instruction is for a teacher to ask a child to refrain from a potentially distracting activity such as encroaching on another student's physical space or speaking loudly during independent work time. Because the situations posed in the STRS items are so common during instruction, these items have plausible relevance to instruction.

Data available in ECLS includes composite scores for the closeness and conflict scales. These composite scores are the average of of all responses to the 7 items on the closeness scale and eight items on the conflict scale (Tourangeau et al., 2016). Using this composite score rather than the individual variables provides a single measure of the underlying construct the entire scale is validated to measure. To generate a score, teachers needed to respond to five items on the scale. Both of these scales had a reliability between .86 and .90 as measures by Cronbach's alpha for kindergarten through third grade (ibid.).

Regarding students' relationship with each other, ECLS data contains survey responses from the perspective of both the teacher and student. For all years of the study, teachers responded to survey items about their observations of students' interactions with each other. These items came from a subset of the Social Skills Rating System (SSRS) (Gresham & Elliott, 1990), which included scales to measure students' internalizing and externalizing behavior. Although SSRS forms for parents, students, and teachers are available, ECLS surveys only contained questions for the teacher. Researchers have translated SSRS scales into many languages and found consistent results with a variety of populations (Pedersen, Worrell, & French, 2001; Van der Oord et al., 2005), used it as a screening tool and assessment for social skills (Gresham, Elliott, Vance, & Cook, 2011; Gresham, MacMillan, Bocian, Ward, & Forness, 1998), and demonstrated convergent validity with other measures of behaviors (Flanagan, Alfonso, Primavera, Povall, & Higgins, 1996). Teachers' responses on this questionnaire rely on their observations of students' interactions with peers. Although this observation provides incomplete information about students' interactions with each other, it provides information about features of students' relationships that teachers notice.

Because SSRS scales are copyright protected, I do not have access to prompts and example items to develop an argument for using SSRS scales within the context of my framework for instruction (Tourangeau et al., 2016). Nonetheless, available research using SSRS scales states that the teacher questionnaire is designed to measure the extent to which a particular student threatens and argues with other students as well as experiences anxiety or becomes embarrassed around other students (see, e.g., Gresham et al. (1998)). Additionally, researchers have shown differences in children's behavior across situations (Achenbach, McConaughy, & Howell, 1987), and SSRS scales use multiple informants to reflect these differences (Gresham & Elliott, 1990). Indeed, Flanagan et al. (1996) found a somewhat weak correlation between the SSRS social skills scale and another measure of social skills, and they suggested that the difference in the scales is associated with the specificity of the SSRS scale to the classroom environment. Because teachers primarily interact with children in an instructional setting, I use teachers' responses to SSRS items as a measure of their perception of a particular student's activity and relationship with other students during instruction.

ECLS data includes scores for interpersonal skills, externalizing behavior, and internalizing behavior based on the arithmetic mean of responses to relevant items (Tourangeau et al., 2016). To create a score for the scales, teachers needed to respond to four of five interpersonal skills items, four of six externalizing behavior items, and three of four internalizing behavior items. This number of items is fewer than the number on the full SSRS scales, and the wording of items in ECLS data collection also varies from wording of original SSRS items in some instances (ibid.). Reliability for each of these scales across all years of ECLS data collection was between .73 and .89, and internalizing behavior having the lowest reliability of all the measures (ibid.).

I use teachers' reports about how they organize students for work on academic content as another measure of students' relationships with each other during instruction. Specifically, I use teachers' reports about their use of pull out and achievement groups in their classes. I use these reports because this organization has consequences for which student-to-student interactions the teacher condones and how the students are positioned with respect to each other. That is, these grouping strategies provide information about constraints teachers place on students' interactions with their peers because these groups dictate, in part, with whom children work on academic content. Moreover, previous research has shown a relationship between class size and teachers' choices with respect to grouping students within the class (Blatchford et al., 2001). Because assignment to specific achievement groups within the classroom can position students as more or less capable than their peers, these assignments have potential to influence students' relationships with each other and is thus relevant to the study of instruction with my framework.

Each year of data collection, teachers had the opportunity to state the number of achievement groups they use in their class and the particular achievement group a particular student was in (see, e.g., Westat, n.d.). Specifically, the surveys prompted, "How many instructional groups based on achievement or ability levels in READING (MATHEMATICS) do you currently have in this child's class?" If teachers responded that they use two, three, four, or five or more instructional groups, the survey prompted, "In which reading (mathematics) instructional group is this child currently placed?" For this question teachers teachers wrote a number, where "1" was associated with the highest achieving group. With these prompts, ECLS surveys make it possible to examine the relationship between class size and achievement grouping within classes at the child level.

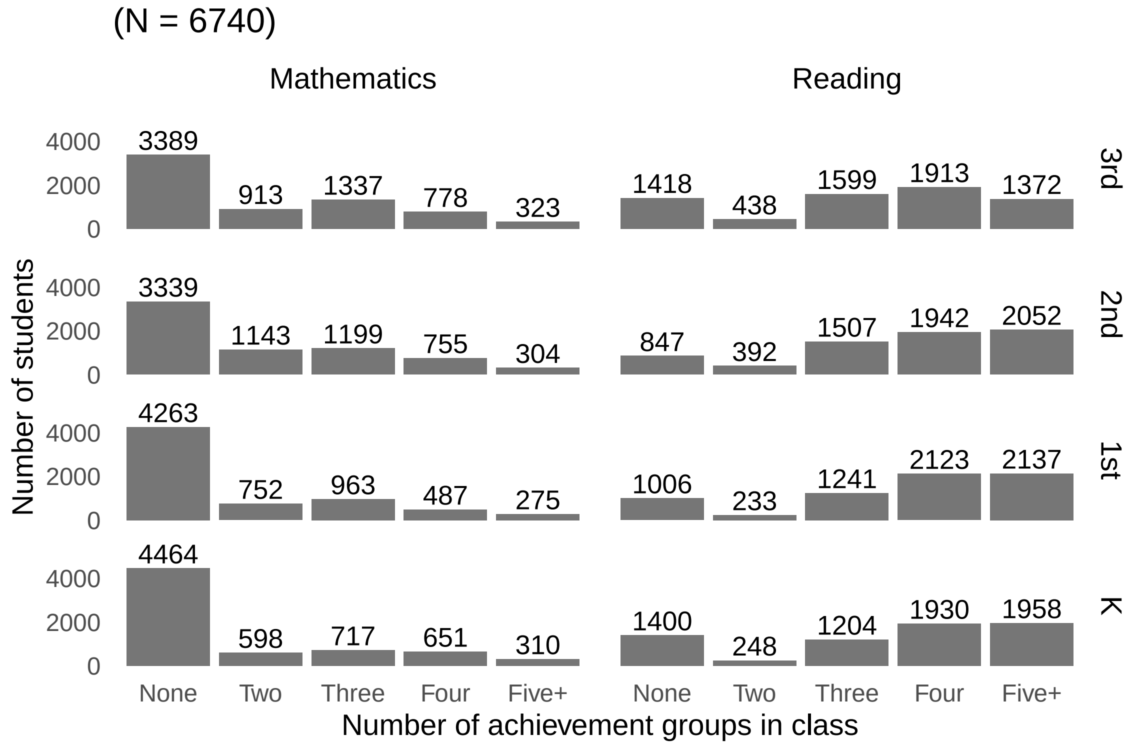
In latter years of data collection, children completed a survey in which they answered questions about their relationships and interactions with peers during instruction. Specifically, children responded to items from the Children's Social Behavior Scale-Self Report (Crick & Grotpeter, 1995), which included prompts such as, "I try to cheer up other classmates who are upset or sad about something." Children responded to these prompts by claiming how often the statement has been true for them. ECLS surveys included three of these items, which is a subset of the items that made up Crick and Grotpeter's scale of prosocial behavior. I conducted a factor analysis with these variables using a geominQ rotation to compute a score for students' reported prosocial behavior during instruction. I use the geominQ rotation because researchers have found it performs well even without knowledge of the true loading structure (Asparouhov & Muthén, 2009). In this factor analysis, all three questions loaded onto one underlying factor.

Descriptive statistics for measures of instruction are in Table 3.2 and Figure 3.1. Two features of these measures are important to note First, they appear robust across grades. Each measure of instruction has approximately the same mean, standard deviation, minimum, maximum, and percentiles from kindergarten through 3rd grade. Thus, throughout early elementary school, students performed nearly the same on these measures as a group. Second, each measure of instruction is skewed in one direction or the other. Measures of closeness and interpersonal skills are each skewed left with at least ten percent of students receiving the highest score. Measures of conflict, externalizing behavior, and internalizing behavior are skewed right with at least ten percent of students receiving the lowest score. In the case of conflict for all grades and externalizing behavior for kindergarten, at least 25 percent of all students received the lowest score on the scale.

Table 3.2: Descriptive statistics for measures of instruction in the ECLS dataset separated by grade (N = 6740)

Measure	Mean	sd	Min	25th percentile	Median	75th percentile	Max
Closeness							
Kindergarten	0.85	0.15	0.07	0.78	0.89	0.96	1
1st Grade	0.84	0.16	0.04	0.75	0.86	0.96	1
2nd Grade	0.82	0.17	0.04	0.71	0.86	0.96	1
3rd Grade	0.8	0.18	0	0.71	0.82	0.93	1
Conflict							
Kindergarten	0.14	0.19	0	0	0.06	0.22	0.94
1st Grade	0.14	0.19	0	0	0.06	0.22	1
2nd Grade	0.14	0.19	0	0	0.06	0.19	0.97
3rd Grade	0.14	0.19	0	0	0.06	0.19	1
Interpersonal Skills							
Kindergarten	0.73	0.21	0	0.6	0.73	0.93	1
1st Grade	0.73	0.21	0	0.6	0.73	0.93	1
2nd Grade	0.72	0.22	0	0.58	0.73	0.93	1
3rd Grade	0.72	0.22	0	0.58	0.73	0.93	1
Externalizing Behavior							
Kindergarten	0.19	0.2	0	0	0.13	0.33	1
1st Grade	0.23	0.2	0	0.06	0.17	0.33	1
2nd Grade	0.23	0.2	0	0.06	0.17	0.33	1
3rd Grade	0.22	0.2	0	0.06	0.17	0.33	1
Internalizing Behavior							
Kindergarten	0.16	0.16	0	0.08	0.08	0.25	1
1st Grade	0.17	0.16	0	0.08	0.11	0.25	1
2nd Grade	0.19	0.17	0	0.08	0.17	0.33	1
3rd Grade	0.19	0.17	0	0.08	0.17	0.33	1
Prosocial Behavior							
3rd Grade	0.75	0.16	0	0.67	0.75	0.83	1

Source: ECLS-K: 2010 - 2013.



Source: ECLS−K: 2010 − 2013.

Figure 3.1: Number of achievement groups teacher uses in the child's classroom for mathematics and reading

I scale all quantitative measures of instruction to be on a scale of zero to one. This scaling has the advantage of consistent interpretation across all scales rather than requiring readers to put each estimate in the results in the context of the measure from which it came. When I interpret estimates, I use the language of points, so I refer to a change of, say, 0.07 as 0.07 points.

While Table 3.2 and Figure 3.1 show the variation that exists in the population for each measure of instruction, one remaining question is the extent to which these measures vary within students during early elementary school. That is, individual students could receive the same scores every year to produce variation in the population that has zero variation within students. Table 3.3 shows the intraclass correlation (ICC) for each measure of instruction. Because students receive a score from one teacher each year and because the group of teachers changes every year, I use two-way random effects, single score ICC (McGraw & Wong, 1996) to calculate a measure of the variation within student for each measure of instruction. Low ICC ($< .6$) for each measure is evidence that a substantial amount of variation exists within students across years for

Table 3.3: Intraclass correlation for measures of instruction

		95% confidence interval	
	ICC	Lower bound	Upper bound
Closeness	0.286	0.272	0.299
Conflict	0.480	0.470	0.491
Interpersonal skills	0.415	0.404	0.426
Externalizing behavior	0.565	0.554	0.575
Internalizing behavior	0.286	0.275	0.298

Source: ECLS-K 2010-2013.

the measures of instruction available in ECLS data (LeBreton & Senter, 2008).[2]

Class Size

Teacher surveys from each year of ECLS contain information about class size. Teachers provided multiple measures of class size by reporting counts for multiple subgroups of students. Specifically, teachers reported the total number of children in the class, the number of boys and girls in the class, the number of students who identify as a particular race or ethnicity, and the number of students of a particular age.[3] These multiple measures of class size provide a reliable way to determine the class size every year for each student in the study. Further, these measures of class size are quantitative, so I can use them as such or create categories of large and small classes according the thresholds that are relevant for policy. Using relevant thresholds in this way can facilitate comparison to prior research related to class size. I represent the distribution of class size from kindergarten through third grade in Figure 3.2. These bar charts show the mean and variation in class size for each grade and the evolution of this distribution during early elementary school. Notably, mean class increased each year from kindergarten though third grade from 20 to 21.99 students per class, and the difference in mean class size among years is statistically significant ($p < 0.001$).

[2] NB: Low ICC could also be an indication of low variability among subjects. This is potentially the case for closeness and internalizing behavior, which both have low ICC and there is little variation in the measures

[3] While these measures of class size most often agreed, there was some deviation. This mean deviation, however, was less than 0.07 in magnitude in all cases. Further, this mean was skewed by large values such a maximum difference of 25 students when comparing the number of students reported by race and number reported by sex.

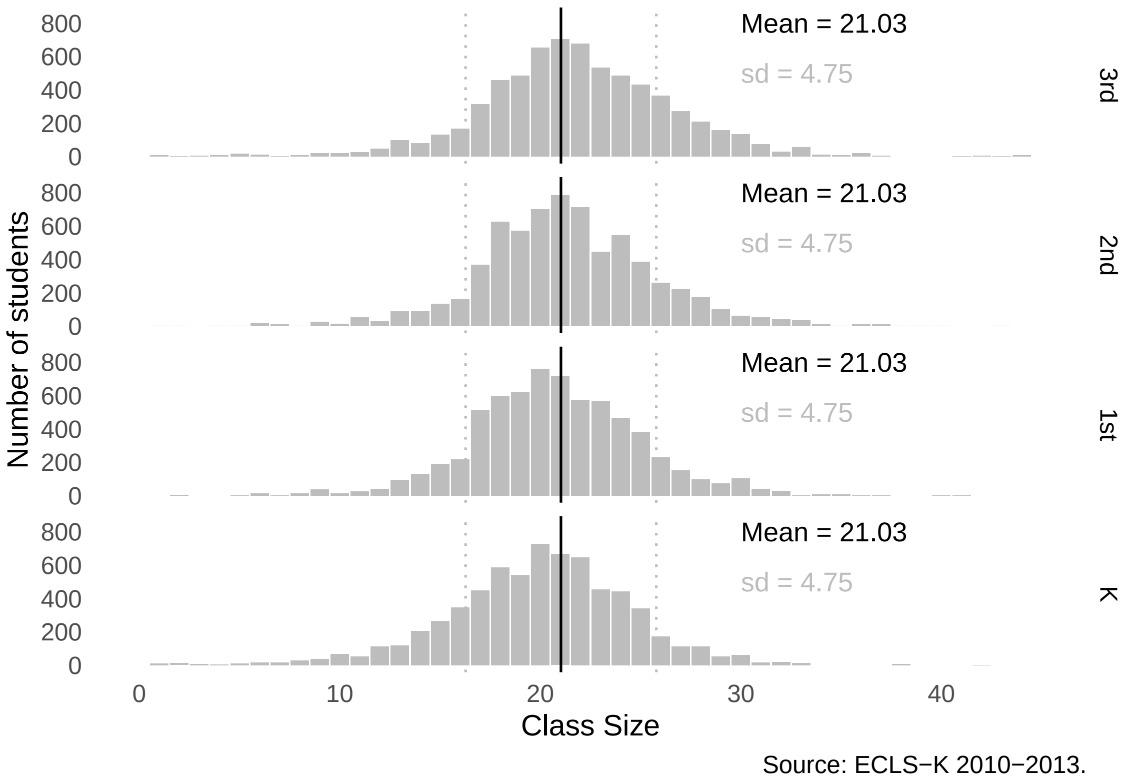
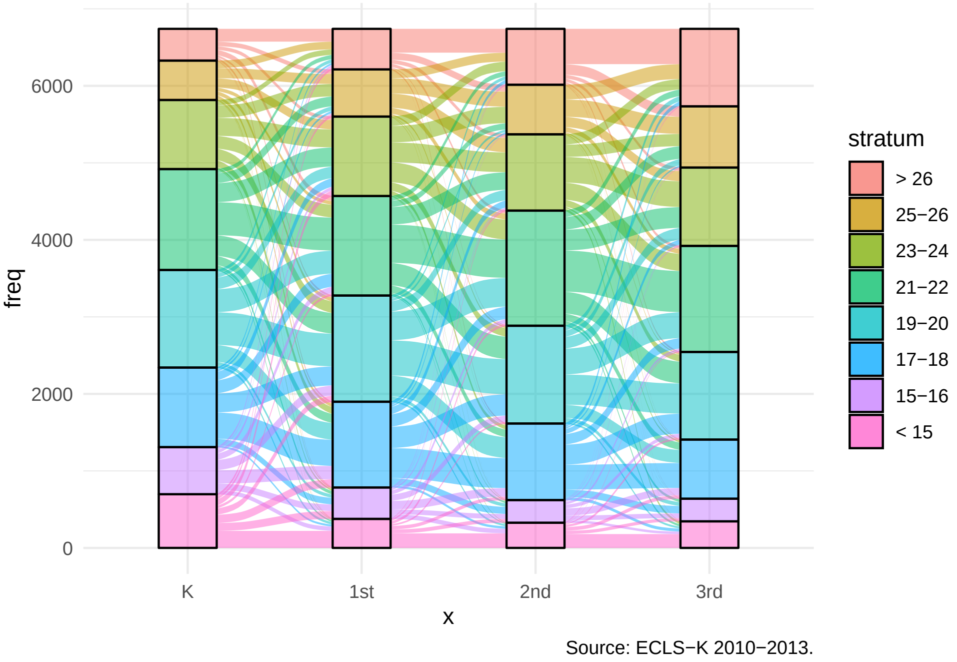


Figure 3.2: Distribution of class size in the ECLS data by grade (N = 6740)

For my analysis, one concern about the data is the possibility of limited variation in class size within children. That is, children who, for example, are in a small class in kindergarten might tend to remain in a small class through third grade because children proceed through school as a cohort. This situation would limit my ability to attribute variation within children in outcomes to class size. To measure class size variation within children, I calculated the intraclass correlation (ICC) using two-way mixed effect, single score ICC (McGraw & Wong, 1996). Using this measure of ICC accounts both for random variation among students and variation due to a fixed effect for each grade. This calculation resulted in a value of .448, which suggests a substantial amount of variation in class size within students. Figure 3.3 illustrates this variation within students by showing the movement of children to classes of different size in successive years of elementary school. In the figure, the thickness of lines exiting columns to the left and entering columns to the right represents the portion of students who shared a common class size in consecutive years. Although the thickest lines reflect students having roughly the same class size from year to year, there is nonetheless sufficient shuffling of students among classes of varying size to produce the low ICC reported.



Source: ECLS–K 2010–2013.

Figure 3.3: Grade to grade class size transitions in the ECLS data (N = 6740)

Other Covariates

A set of covariates and control variables I will consider for use in my analysis appears in Table 3.4. These variables are organized into groups according to how they are relevant in the instructional triangle and whether they are constant or varying over time. Each group of characteristics maps onto a component of the instructional triangle in 2.1, with school as well as family and household characteristics as part of the environment.

This set of covariates and control variables serves two purposes in my work. First, I will use this information to respond to my first research question and describe the predictors of enrollment in a small class in terms of socio-demographic student characteristics. Given the breadth of information available in ECLS, I will be able to construct a detailed description of students who are enrolled in smaller rather than larger classes. Second, I will use socio-demographic information as control variables in my analysis of the relationship between class size and instruction. This set of control variables will enable me to consider possible bias in my estimates that could arise from observable student characteristics and will also make it possible to estimate the

Table 3.4: Covariates and control variables available in the ECLS dataset

	Constant	Time varying
Student		
	Race and ethnicity	BMI
	Sex	Has diagnosed disability or received therapy services
	Age upon kindergarten entry	Received special education services
		Moved
Family/household		
	Language at home prior to kindergarten	Total number of people living in household
	Household SES prior to kindergarten	Parents/guardians living in household
School		
		Locale
		State
		Public or private
		Number of students
Teacher		
		Views of students and teaching
		Highest degree earned
		Years of experience
Class		
		Behavior rating
		Percent boys

Note: Variables are classified as either constant or time varying at the student level. That is, for example, students' age uppon entry into kindergarten never changes, but the number of students enrolled at the school the student attends changes from year to year.

relationship between class size and instruction with greater precision. Descriptive statistics for time constant categorical covariates are in Table 3.5, time constant quantitative covariates are in Table 3.6, time varying categorical covariates are in Table 3.7, and time varying quantitative covariates are in Table 3.8.

Bi-variate tests showed some relationships between the independent variables of interest in the study–sex, race, ethnicity, and family SES– and other covariates. For instance, one-way ANOVA showed differences in mean family SES based on student race or ethnicity ($p < 0.001$). Also, a chi-squared test showed differences in the school locale students attend based on race ($p < 0.001$). Another chi-squared test shows differences in in the size of school students attend based on race ($p < 0.001$). A final chi-squared test shows differences in the size of school based on locale ($p < 0.001$). Differences in these groups of students among my independent variables of interest provides some necessary context to interpret the results of this study.

Analytic Strategy

In this subsection, I describe the methods of analysis I use to respond to my research questions. I organize this subsection around my approaches to answer the research questions I have articulated. That is, I first specify the models I use to explain the

Table 3.5: Descriptive statistics for time-constant binary and categorical student variables

	Proportion
Female	0.49
Asian/Pacific Islander	0.07
Black/African American	0.09
Hispanic	0.21
White	0.57
Other	0.05
English only	0.83
Mom married at first birth	0.74
Mom received WIC	0.37
Student received pre-k care	0.81
Student received WIC	0.41

Source: ECLS-K 2010-2013.

Table 3.6: Descriptive statistics for time-constant quantitative variables

	Mean	sd
Mom age at first birth	25.25	5.81
Family pre-K SES	0.09	0.81
Kindergarten entry age	5.54	0.36

Source: ECLS-K 2010-2013.

Table 3.7: Descriptive statistics for time-varying binary and categorical variables

	Proportion			
	Kindergarten	1st Grade	2nd Grade	3rd Grade
Teacher Male	0.02	0.03	0.05	0.08
Advanced degree	0.45	0.49	0.48	0.51
Teacher Native American	0.02	0.01	0.01	0.01
Teacher Asian	0.02	0.02	0.02	0.02
Teacher Black/African American	0.05	0.05	0.05	0.06
Teacher Hawaiian/Pacific Islander	0.01	0.00	0.00	0.00
Teacher Hispanic	0.09	0.09	0.09	0.09
Teacher White	0.93	0.92	0.92	0.92
Class Behaves Exceptionally well	0.11	0.09	0.10	0.11
Class Behaves Well	0.39	0.38	0.41	0.41
Class Occasionally misbehaves	0.40	0.39	0.37	0.36
Class is Often difficult	0.08	0.12	0.10	0.10
Class is Almost always difficult	0.02	0.01	0.01	0.02
Less than six people in household	0.81	0.80	0.80	0.79
Two parents	0.77	0.76	0.75	0.73
One parent	0.16	0.17	0.17	0.18
One parent plus another adult	0.05	0.06	0.06	0.07
Other guardian	0.02	0.02	0.02	0.02
500+	0.50	0.53	0.52	0.52
300-499	0.32	0.31	0.32	0.32
<300	0.18	0.17	0.16	0.17
Public	0.87	0.88	0.88	0.88
City	0.27	0.27	0.27	0.27
Suburban	0.37	0.37	0.37	0.37
Town/rural	0.36	0.35	0.36	0.36
Diagnosed disability	0.20	0.15	0.16	0.15
Received special ed	0.05	0.06	0.08	0.09

Source: ECLS-K 2010-2013.

Table 3.8: Descriptive statistics for time-varying quantitative variables

	Kindergarten		1st Grade		2nd Grade		3rd Grade	
	Mean	sd	Mean	sd	Mean	sd	Mean	sd
Teacher experience	14.77	9.80	15.05	9.98	15.50	9.84	14.52	9.45
Teacher views of learning	2.35	0.70	2.41	0.73	2.46	0.72	2.65	0.77
Teacher views of teaching	4.38	0.72	4.29	0.75	4.19	0.76	4.11	0.80

Source: ECLS-K 2010-2013.

relationship between socio-demographic characteristics and class size during early elementary school in the United States. Then I specify my models to estimate the relationship between class size and instruction. I use different strategies to analyze this relationship based on the data available in ECLS. Specifically, for measures of teachers' decisions (i.e., their choices about how to organize students for academic work) and students' reports of pro-social behavior, cross-sectional data are available. For measures of teachers' perceptions of students (e.g., closeness in the teacher-student relationship), longitudinal data are available. Therefore, I specify different models

Table 3.9: Analytic strategy for each outcome of interest

Analysis	Outcome
Regression and survival analysis	Class size
Fixed effects regression	Closeness in relationship with child Conflict in relationship with child Interpersonal skills Externalizing behaviors Internalizing behaviors
Bayesian additive regression trees	Participation in ELA achievement groups Participation in mathematics achievement groups Prosocial behavior

for estimating the relationship between class size and these measures of instruction. My presentation of each model follows the same structure. First I specify the model mathematically and explain relevant assumptions. Then I provide details about calculating standard errors and measuring uncertainty in the model estimates. Finally, I describe how I explore the sensitivity of my models to the analytic choices I made.

Class Size and Socio-Demographic Characteristics

As a preliminary for my analysis of the relationship between class size and instruction, I first examine the socio-demographic characteristics that explain class size in the United States. I conduct this analysis for reasons related to equality and equity. Regarding equality, I wonder if students enroll in smaller and larger classes at the same rates along lines of race, ethnicity and family SES. Like other resources, class size has potential to be unequally distributed among people who are more and less privileged. Regarding equity, I wonder if class size might be used to interrupt patterns of injustice. Prior research suggests students who are members of historically marginalized groups might receive greater benefits than their counterparts from smaller classes. If class size happens to be unequally distributed, it might be distributed in such a way as to target benefits towards those who need them the most. In total, this analysis of the relationship between class size and socio-demographics shows who is positioned to benefit from any relationship between class size and instruction I take up in subsequent analysis.

Model Specification

I use two approaches to analyze the relationship between socio-demographic characteristics and class size in early elementary school. First, I conduct a cross-sectional analysis of class size to gain insight into the predictors of class size in each grade. A benefit of this approach is the ability to observe which variables have a significant relationship with class size in a particular grade, which may not be the same variables from grade to grade. Because I use a separate regression for each grade, these estimates are specific to each grade rather than the average effect across all grades. Second, I conduct an event history analysis to explain the relationship between socio-demographic characteristics and enrollment in a large or small class at some point in early elementary school. This analysis takes advantage of the longitudinal structure of the ECLS data to consider the occurrence and timing of enrollment as part of the relationship between socio-demographic characteristics and enrollment in large or small classes.

Cross-sectional Analysis For the cross-sectional analysis, I specify a linear regression model for each grade as

$$y_{is} = \beta_0 + \boldsymbol{\beta}_1 \boldsymbol{X}_i + \boldsymbol{\beta}_2 \boldsymbol{X}_s + \epsilon_{is}, \tag{3.1}$$

where y_{is} is the class size of student i attending school s. Further, $\boldsymbol{X}_i$ and $\boldsymbol{X}_s$ are vectors containing information about students and schools respectively. In light of research that suggests students who are members of historically marginalized groups may receive the greatest benefit from class size (e.g., Finn and Achilles (1990)), I include variables for race, ethnicity, sex, and family SES in every model specification. Also, because class size is related to state-level legislation and policies such as school funding, I include state level fixed effects as a component of $\boldsymbol{X}_s$ in every model specification. I fit the model in equation (3.1) by the method of least squares.

In addition to sex, race, ethnicity, and family SES, I include a variety of other variables that are plausibly related to class size in the model. Including these variables accounts for potential omitted variable bias to. These other variables fall into the categories of school, student, teacher, and household characteristics in Table 3.4.

Linear regression models such as the one I have specified rely on multiple theoretical assumptions for valid statistical inference (Rencher & Schaalje, 2008), and I address two salient assumptions here. First, linear regression assumes homoscedasticity, or, in other words, constant variance in the outcome. In this study, homoscedasticity means the variance of class size is the same for all values of the independent variables. To expound, the linear model model assumes variation in class size is the same for White and African American students; for students who speak English at home and those who do not; for students who attend a public school and those who attend private school; and for all combinations of all values of all covariates. Because of potential heteroscedasticity in my data, I calculate robust standard errors using the HC1 method, which researchers have shown effectively reduces bias in estimates with unbalanced data (Hinkley, 1977).

Second, linear regression assumes independence of observations. If observations are independent, knowledge about one observation will provide zero information about another observation. If, on the other hand, observations are dependent, knowledge about one observation provides information about other observations. ECLS data violates the linear regression assumption of independence because many students in the sample attend the same school. Moreover, ECLS data contains repeated observations of the same student across grades in elementary school. Observations in the data are thus dependent because information about one student's class size in a particular school provides information about the class size of other students who attend the school and the student's class size in other grades.

Violation of the independence assumption has two consequences. First, calculation of ordinary standard errors will yield incorrect inference in my analysis. Given a sample of a constant size, an ordinary calculation of standard errors appropriate for independent observations are less than standard errors adjusted for clustering. These smaller standard errors could lead to erroneous statistically significant results. For this reason, I account for clustering in the data at the school level in the calculation of standard errors.

To estimate the standard error of coefficient estimates in the linear models, I cluster at the school level and use the CR2 adjustment. Researchers have shown this adjustment has better properties for statistical inference than other common options (Pustejovsky & Tipton, 2018). In this study, clustering at the school level is necessary because students' assignment to class size is correlated within schools, and clustering standard errors models this dependence (Abadie, Athey, Imbens, & Wooldridge, 2017). Other plausible levels of aggregation for this analysis include the teacher, district, and state because each of these levels have connections to class size. Given that NCES did not systematically sample schools from districts and that I use state-level fixed effects in the model specifications, the school is the highest level of aggregation for clustering standard errors.

I check for sensitivity to outliers, influential observations, and collinearity in my models using studentized residuals, Cook's distance, and variance inflation factor. After specifying the preferred model, I use these tests to identify potentially problematic observations or covariates. For studentized residuals, Cook's distance, and variance inflation factor, I consider values, respectively, greater than two standard deviations from the mean, greater than $4/n$ and, greater than 10 to be potentially problematic. These cutoffs, while imperfect, are consistent with advice in existing research (Fox, 1997). With potentially problematic components of the model identified, I then delete these components from the model and compare results across model specifications. In this comparison, I look for substantive changes to the inference I draw from the preferred model and the models excluding potentially problematic components.

Event History Analysis As a second analytic approach, I use an event history analysis to explain student enrollment in a large or small class at some point in early elementary school. Whereas the regression analysis I specified previously describes the characteristics associated with class size within each grade, this approach adds insight into the occurrence and timing of enrollment in a large or small class during early elementary school.

For event history analysis of class size, I use a discrete time hazard model. I choose this model for two reasons. First, students in the ECLS data experience a potential change in class size once per year, which amounts to four time points during the period of observation. Second, many students experience the event of enrolling in a large or small class at the same time. In event history analysis, multiple students experiencing the event of interest at the same point in time is called a tie. Ties in the ECLS data are due to the structure of the school experience–that students generally experience a change in class size once per year. For these reasons, a discrete time model is preferable to other options such as a cox proportional hazard model.

To conduct the analysis, I define large and small classes according to the thresholds used in project STAR (Finn & Achilles, 1990). Specifically, I use class size of 17 or less as small and 22 or greater as large. I estimate models for first enrollment in a large or small class separately. For both of the models I use a binary dependent variable that indicates enrollment in, for example, a large class (22 or more students) or not a large class (less than 22 students). I construct an analogous binary outcome to represent enrollment in a small class or not a small class. Conducting analyses separately for enrollment in a large and small class makes it possible to observe different sets of variables that have a significant relationship with different outcomes.

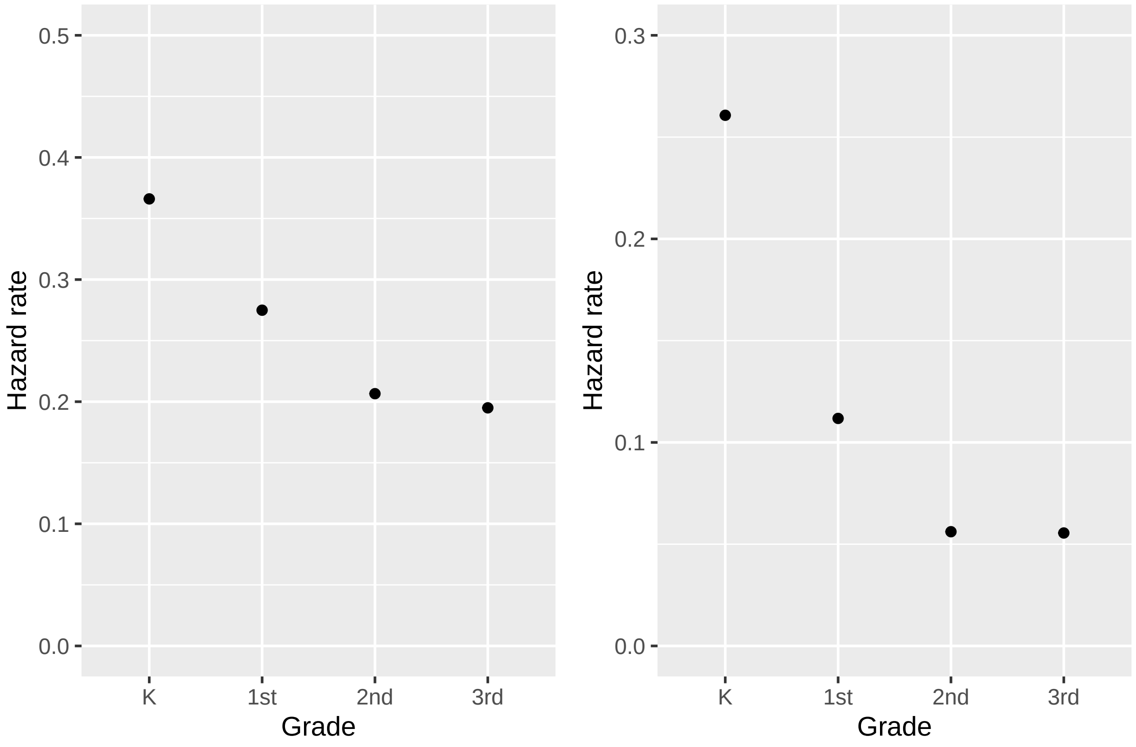
I compile data for event history analysis with one observation at the student-year level. For data to analyze enrollment in a large class, once a student enrolled in a class of 22 or more students for the first time, subsequent student-year observations are removed from the data set. A student who, for example, enrolled in a class of 20 students for kindergarten and a class of 28 students for first grade had observations in the data set for kindergarten and first grade only. Students who always enroll in a class of at most 21 students remained in the data set for all four years and are considered 'censored' after third grade. In other words, these students never enrolled in a 'large' class during the period of observation for this study. I conduct this analysis for small classes using the same data structure and a class size threshold of 17.

To fit the discrete time hazard model, I estimated a logistic regression with the binary outcome of enrollment in a large (or small) class. Each model I specify included a variable to account for grade, which captures potential differences in the probability of enrolling in a large (or small) class in each year. These models also included the control variables corresponding to the instructional triangle

I estimated models that included time constant and time varying covariates. Conducting these separate analyses makes it possible to observe the extent to which the occurrence and timing of enrollment in a large or small class is responsive to changing conditions for students or is more or less fixed by a set of initial conditions. For the models with all time constant covariates, I included the first observed value of time varying covariates as a time constant covariate. As an example, school population was a variable that could change from year to year for a particular student. For the time constant analysis, I included the population of the school each child attended for kindergarten as a time constant covariate. For the time varying analysis, I allowed these covariates to change over time within child.

Like the linear models I specified previously, an assumption of logistic regression is independent and identically distributed observations. Because observations in my data are statistically dependent due to students attending the same school and by state policies regarding class size, I again cluster standard errors at the school level and include fixed effects for state in my model. Including these terms in the model specifies the relationship among observations in my data and adjusts the standard errors to account for dependence among observations.

Figure 3.4 shows the portion of students who enrolled in a large and small class each year during early elementary school. Inspection of the figure raises some doubts about the students who enroll in a large or small class early in the observation period. Especially for enrollment in a small class, the portion falls by what seems to be larger magnitude after kindergarten than it does after subsequent grades. This pattern suggests possible heterogeneity among students in my sample with respect to enrollment in a small class.



Source: ECLS−K 2010−2013.

Source: ECLS−K 2010−2013.

Figure 3.4: Hazard rates for enrollment in a large and small class

Because of the patterns in Figure 3.4, I allowed for the possibility in my models of unobserved heterogeneity in the distribution of risk to enroll in a large or small class. Because the rate of enrollment in a large or small class drops after kindergarten such that children who enrolled in a large or small class in kindergarten could plausibly be fundamentally different than students who remain in the sample for subsequent years due to unobserved characteristics. This difference could bias my estimate of the relationship between enrollment and my independent variables of interest. This varying risk groups of students face based on state and school is known as a shared frailty in event history analysis. A member of a group with greater shared frailty is more likely to experience the event of interest than a member a group with lesser frailty. Including random effects at the child level accounts for potential unobserved heterogeneity within my sample.

For the survival analysis models, I conducted sensitivity checks for the choice of threshold for large and small classes. In the models described above, I chose thresholds for large and small classes that aligned with thresholds found in project STAR (Finn & Achilles, 1990). Although these thresholds have special interest for research, they can also obscure patterns in the data. For this reason, I examine the extent to which results of the survival analysis are sensitive to the choice of threshold. To do so, I estimate survival analysis models with a thresholds in the range of 20 to 24 students for large classes and 15 to 19 students for small classes. These ranges are supported by the data given that they are no more than approximately one standard deviation away from the mean class size. By conducting the survival analysis with a range of thresholds, I gain insight into how the choice of threshold does or does not affect the results of my analysis.

Differences in Teachers' Classroom Organization Choices and Students' Reports of Prosocial Behavior

Of central interest in this study is the effect of class size on instruction. Endogeneity is a major challenge for studying this relationship. In the context of this study, endogeneity means that class size might influence instruction, but instruction also plausibly influences class size. To animate this statement, imagine a school that is committed to a clear vision of high quality instruction. Details about this vision are unimportant for now. What is important to know is that administrators at the school have reason to believe that enactment of this vision requires class sizes between 20 and 25 students. Supposing the administrators can successfully advocate for their perspective, they will manipulate class size to support their vision of high quality instruction. In this situation, class size and instruction are certainly related, but manipulation of class size to support a particular vision of instruction obscures the effect of class size on instruction in the absence of such a vision. That is, one cannot reasonably conclude from the preceding example that a class of 20 to 25 students led to the observed instruction and that similar instruction would be observed in other classes of 20 to 25 students.

Researchers have devised multiple methods to solve the endogeneity problem. Most powerful among these methods is randomization. Project STAR (Finn & Achilles, 1990), for example, endeavored to measure differences in students' educational outcomes after random assignment to a small or regular class size. Because ECLS data contains observations of rather than random assignments to class size, I need to find another strategy to address endogeneity.

Identifying an instrumental variable is one such strategy. To conduct instrumental variable analysis, one uses a variable that is, in the context of this study, unrelated to instruction but related to class size (Angrist & Pischke, 2009). I attempted to use cross-county mobility within the United States as an instrumental variable in this study. My logic for exploring this instrument was that schools can predict some mobility into and out of their boundaries, but these predictions contain imprecise information for which schools cannot fully plan. I reasoned that this imprecision might lead to slightly more or slightly fewer students than planned for in each school. Further, I reasoned that these unplanned for students would lead to variation in class size for which administrators could not account. That is, my instrument would isolate some variation in class size that is entirely unrelated to instruction. Alas, this instrument suffered the same fate as so many others. Specifically, cross-county mobility is related to class size in the ECLS data, but this relationship is too weak to yield useful inference about the relationship between class size and instruction.

To learn about the relationship between class size and instruction, Bayesian Additive Regression Trees (BART) (Chipman, George, & McCulloch, 2010) is one method I used. BART is a machine learning technique to predict outcomes. Researchers have suggested BART is a useful technique for analysis of observational data with many covariates such as ECLS (Hill, 2011). BART is useful in this study because I can use it to specify a class size and formulate predictions of measures relevant to instruction. By comparing predictions across a range of class sizes, I can draw inference about the relationship between class size and instruction. Although this method does not eliminate endogeneity that exists in the relationship between class size and instruction, it uses information in the data to predict plausible changes in measures of instruction that occur due to changes in class size. In the remainder of this subsection, I describe my implementation of the BART model to analyze the relationship between class size and teachers' choice about grouping students as well as students' reports of prosocial behavior.

Model Specification

To create a prediction, BART uses the sum of predictions from a predetermined number, m, of trees. Each tree in the model consists of multiple nodes. Each node partitions the data by a combination of values of covariates. At the end of each node is a mean value for the outcome. This value is called a leaf on the tree. Predicted values for observations in the data are then the sum of all leaves corresponding to the covariates for that observation.

To provide a simple illustration of BART, I adapt an example from existing literature (Tan, 2019) to my topic for this research. To begin, all the values in Figure 3.5 are fabricated, but I use them to show trees BART could possibly produce. For this brief example, I constructed two trees. A full implementation of BART has more trees.

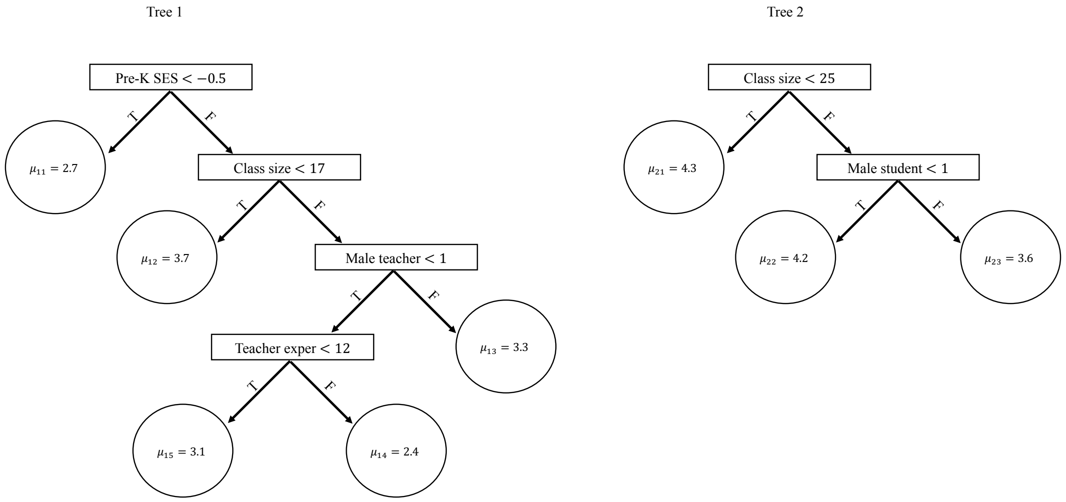


Figure 3.5: Illustration of a possible BART tree

In Figure 3.5, Tree 1 has five nodes represented by circles. Each of these nodes corresponds to a distinct subset of the data. One node, for example, corresponds to all observations with SES prior to kindergarten less than −0.5. For these observations, the mean value of the outcome is 2.7. This value is one leaf on the tree. Another node on Tree 1 corresponds to all observations with SES prior to kindergarten greater than −0.5 and class size less than 17. For this partition of the data, the mean outcome is

3.7. One can similarly interpret other nodes on every tree in the model to specify all leaves.

As mentioned previously, BART uses the sum of regression trees to produce predicted values for observations. Specifically, this sum uses all the leaves relevant to an observation to predict the outcome for that observation. Using the trees in Figure 3.5, an observation with Pre-K SES of 1.2 and class size of 15 would have a predicted value of $\frac{\mu_{12}}{2} + \frac{\mu_{21}}{2} = 4$. Note that, in the calculation of the sum, the leaf of each tree is divided by the number of trees, m.

In formal terms, the BART model is written

$$Y = \Sigma_{j=1}^{m} g(x; T_j, M_j) + \epsilon, \epsilon \sim N(0, \sigma^2). \tag{3.2}$$

In this equation, T_j represents a tree, and M_j represents the set of leaves corresponding to tree j. Further, g is a function that selects the appropriate leaf in tree j for the value of the covariate, x. Finally, the outcome, Y, is the sum of these m leaves plus a normally distributed error, ϵ (Chipman et al., 2010).

One risk of using a machine learning algorithm is over fitting the data. That is, machine learning techniques have the potential to use too much information in the data to generate predictions for an outcome. Excessive use of information happens because, along with systematic variation, data also contains random variation. Over fitting leads to mistaking random, idiosyncratic variation in the data as evidence of systematic relationships between predictors and an outcome. Such over fitting undermines the trustworthiness of results.

Selection of prior distributions for parameters in the BART model protects against over fitting the data (Chipman et al., 2010; Rocková & Saha, 2018). Within the Bayesian framework for statistical analysis that governs BART, priors make it possible to incorporate knowledge about a phenomenon of interest that exists before analysis of the current data. People are likely to have at least some knowledge about parameters related to a phenomenon before they formally collect data. A frequentist framework for statistical analysis requires strict reliance on formally collected data to learn about the relationship between parameters in the model and the outcome. A Bayesian framework, however, uses priors to take advantage of previously existing knowledge. Researchers then update their knowledge about parameters in the model by analyzing information available in the data. Priors can vary in strength according to the confidence researchers have in their previously existing knowledge. Weak priors allow collected data the ability to freely update previously existing knowledge, and strong priors constrain this ability. Thus, researchers using a Bayesian framework can alleviate concerns about over fitting data by adjusting the strength of prior distributions.

One prior for the BART model influences how trees, each T_j, grow and renders each tree a "weak learner". Forcing each tree to be a "weak learner" means individual trees have less correlation with the outcome of interest than a single "strong learner." Using the sum of many "weak learners" has the advantage of reducing error variance in the overall estimate. One component of the prior for trees in BART is the probability a tree at depth d grows deeper, which is given by the following equation:

$$\frac{\alpha}{(1+d)^{\beta}}. \tag{3.3}$$

In this expression, $\alpha \in (0, 1)$ determines the baseline probability a node splits. Next, $\beta > 0$ imposes a penalty for growth at depth, d. By construction, a great value for β corresponds to a large penalty (Chipman et al., 2010; Tan, 2019). For purposes of prediction, recommended default values for these parameters are $\alpha = 0.95$ and $\beta = 2$ (Chipman et al., 2010). I use these default values in my model specification.

Additional components of the priors for trees in the BART model are distributions for the selection of covariates and cut points with the trees. Following default recommendations, I use a uniform distribution for both of these priors (Chipman et al., 2010). Use of uniform distributions makes all variables and all cut points equally likely candidates for selection in the model.

Remaining parameters in the BART model for which prior distributions are needed are every $\mu_{ij} \in M_j$ and σ. For each μ_{ij}, the prior is a normal distribution with mean and variance such that they are tuned by a parameter, k, to match the range of the observed outcome, Y (Chipman et al., 2010). Chipman, George, and McCullogh propose $k = 2$ as a default that yields good results, which is the value I use in my models.

For σ, the prior is an inverse chi-square distribution with degrees of freedom, ν, and a scaling parameter, λ, such that the distribution is aligned with the observed standard deviation, $\hat{\sigma}$, of the outcome. To determine λ researchers specify a quantile, q, of the distribution to match $\hat{\sigma}$. I follow Chipman, George, and McCullogh's suggestion to use $\nu = 3$ and $q = 0.90$ as a default that avoids over fitting the data.

Choosing the number, m, of trees to include in the model requires balancing competing needs for prediction of the outcome and selection of relevant covariates. Chipman, George and McCullogh (2010) have shown that, on one hand, greater m improves prediction. On the other hand, m too great can impose unnecessary computational cost because marginal improvement in prediction decreases as m increases. In other words, the quality of prediction more or less reaches a plateau rather than continually increasing with greater and greater m. Further, m too great can compromise the ability of the algorithm to identify relevant predictors of the outcome. With many trees, more predictors can enter the model, but fewer trees forces predictors to compete for entry. This competition leads to inclusion of the most relevant covariates in the model. With these competing needs in mind, common advice is to choose $m = 200$ (Chipman et al., 2010; Tan, 2019), although some researchers have argued for the merits of fewer (e.g., $m = 20$) trees (Bleich, Kapelner, George, & Jensen, 2014). I set the number of trees in my model to 200.

Having describe the choice of prior distributions and the number of trees in the model, I turn now to a brief description of the process whereby trees in the BART model are created (Chipman et al., 2010). While I avoid technical details, this description establishes a foundation for understanding the results of the analysis. BART begins with m single node trees. At the beginning of an iteration, the algorithm uses the uniform prior distributions to propose a new tree, T_j. Using information in the data and constrained by α and β in expression (3.3), the algorithm then determines whether to accept or reject the proposal. These two steps repeat for all m trees.

Next, the BART algorithm uses the data and the most recent set of trees to obtain M_j. This task is accomplished by taking a draw from the normal prior distribution for each leaf (i.e., each μ_{ij}) in the existing trees. After these draws, every tree in the model has an established set of nodes and a corresponding leaf for each node.

With the complete tree structure in place, the final step of an iteration is to draw a value for the variance, σ^2, of the error term. Given the current tree structure and the data, the BART algorithm draws from the inverse chi-square prior distribution for this value. Having updated all the parameters (T_j, M_j, and σ) in the BART model, the iteration is complete.

For inference about the outcome, Y, BART completes many, N, iterations of the algorithm. Because the BART process begins in an arbitrary state with m single node trees, iterations of the BART algorithm near the beginning of the process can be erratic before settling into a relatively stable state. For this reason, these beginning iterations are called the burn-in period. Hence, the total number of iterations in the process can be expressed as $N = b + k$, where b is the number of burn-in iterations. Because the erratic burn-in iterations of the BART algorithm are discarded, k iterations are available for drawing inference about Y.

These k iterations of the BART algorithm constitute a sample of draws from the posterior distribution of T_j, M_j, and σ. This sample is called a posterior distribution because it is the knowledge one has about the parameters after updating the priors with information contained in the data. Because this posterior distribution is a sample from the distribution of parameters affecting the outcome, Y, the set of predicted Y values from the k iterations reflects the distribution of Y.

Main results from BART, then, are (1) the set of k predicted outcomes and (2) the k sets of trees used to make those predictions. With the predicted values, I calculate statistics to learn about the relationship between class size and the outcomes related to instruction. With the sets of trees, I examine the frequency with which variables appear in the model. This frequency provides information about the usefulness of particular variables for predicting outcomes related to instruction.

To examine how class size is related to students' reports of prosocial behavior and teachers' use of achievement groups, I calculate partial dependencies of the outcomes on class size. That is, I calculate the predicted value of measures of instruction over the range of class size. By doing so, I can explore how predictions of instruction relevant outcomes vary over the entire distribution of class size and test for significant differences using ANOVA. Also, I test for differences in instruction by the thresholds for class size used in project STAR of 13 to 17 and 22 to 25 students (Finn & Achilles, 1990). Because STAR is an influential experiment in class size literature, I use the thresholds in this study to understand possible differences in instruction that occur in classes of these sizes.

One strength of BART is the ability to detect interactions among covariates (Chipman et al., 2010). This strength is important for my study because my third research question focuses on the possibility that class size has heterogeneous effects on students. Some existing literature suggests that students who have historically been disenfranchised in schools reap greater benefits from class size reduction policy (Krueger, 2002). BART offers the ability to examine the extent to which such benefits might operate through measures of instruction available in ECLS data. Specifically, I use the k sets of trees from the BART model to calculate a prediction of the relationship between class size and outcomes relevant to instruction conditional on student race and ethnicity as well as family SES. Researchers have demonstrated results from BART are well suited to conduct these calculations (Green & Kern, 2012; Hill, 2011). Within each level of the variable on which I condition, I calculate the relationship between class size and the outcomes relevant to instruction.

Estimating uncertainty for results from BART is straightforward (Chipman et al., 2010). Because the results provide a distribution of predicted outcomes, the variance in this distribution measures uncertainty in the results. To find the endpoints of, for example, a 95% confidence interval, I use the distribution of predicted outcomes to calculate the 2.5th and 97.5th quantile.

Class Size and Changes in Teachers' Perceptions of their Students

Model Specification

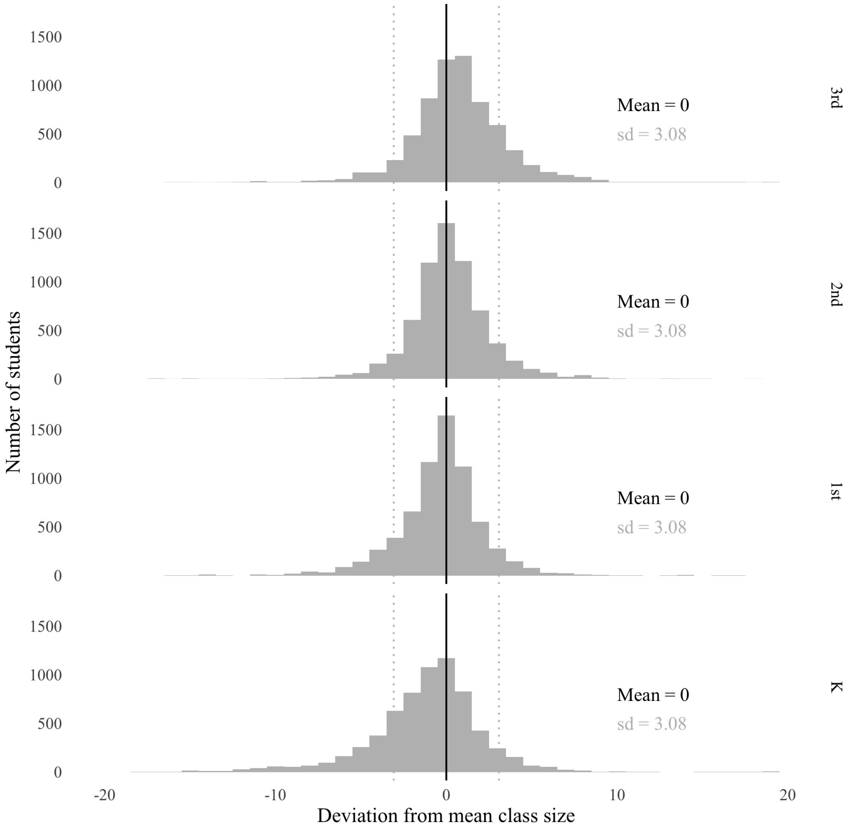
I use a fixed effects regression model to investigate changes in the ways in which students are perceived by their teachers that occur in association with class size. Teachers completed surveys about ECLS students every year, so I observe changes in how students are perceived by teachers in their behavior and relationships over time and in connection with variation in class size. An advantage of a fixed effects model for this analysis is the ability to control for all unobserved student characteristics that do not change over time (Allison, 2009). Some characteristics that may have a tremendous influence on relationships, such as personality, are arguably constant over time. To the extent that these characteristics are constant, they will not affect my estimates of the effect of class size on students' relationships. Of course, many factors that could affect students' relationships in school vary over time, and I will draw on the set of covariates available in ECLS to account for factors such as changes in family composition, moving to a new area, and teacher characteristics.

To implement the fixed effects regression model, I use the hybrid method Allison (2009) proposed. This method makes it possible to estimate the effect of variables that are constant within student over time. Examples of such variables are sex, race, ethnicity, and family SES prior to kindergarten. Estimating the effects of these variables is necessary for responding to my third research question about potentially heterogeneous relationships between class size and instruction along lines of race, ethnicity, sex, and family SES.

When implementing the hybrid method, researchers first calculate the child-specific mean and year-to-year deviations from the mean for all time varying covariates. Then, both the mean and deviations are included in the regression model along with a child-specific random effect. Because mean class size in the population increases between kindergarten and third grade, I include a state-grade fixed effect in the model to account for natural increases in class size at the state level. :

$$y_{igs} = \mu + \beta_1(\overline{C_i} - C_{ig}) + \boldsymbol{\beta}_2\boldsymbol{X}_{ig} + \boldsymbol{\beta}_3\boldsymbol{X}_{sg} + \alpha_i + \epsilon_{igs}, \tag{3.4}$$

where y_{igs} is the outcome of the measure relevant to instruction for student i attending grade g at school s. In the model, $\overline{C_i}$ is the mean class size for student i in early elementary school and C_{ig} is the class size class size for student i in grade g. Distributions by grade for this variable are represented in Figure 3.6. Inspection of this figure shows the bulk of the distribution falls in the range of -10 to 10 for each grade.

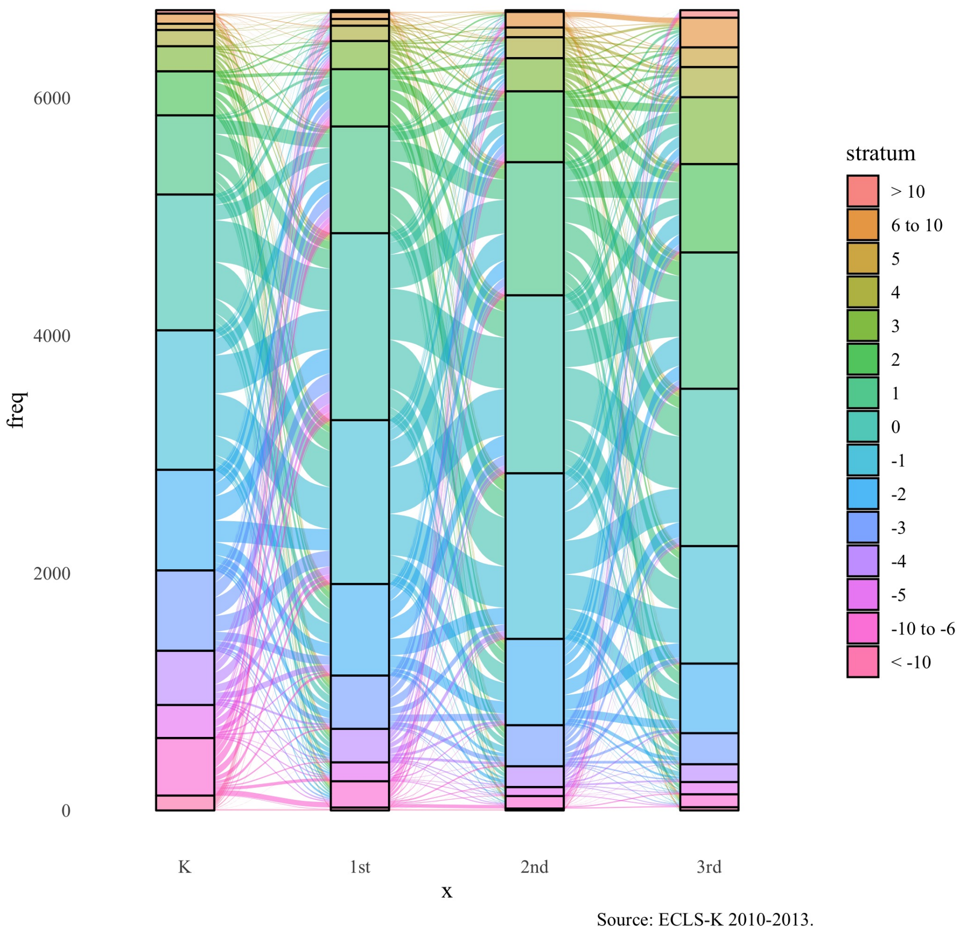


Source: ECLS-K 2010-2013.

Figure 3.6: Distribution of devaiation from child-centered class size by grade

Completing the model, $\boldsymbol{X}_{ig}$ is a vector of covariates specific to child i in grade g and includes the child-year predicted class size. Next, $\boldsymbol{X}_{sg}$ is a vector of school-grade level covariates, which includes a fixed-effect for state-grade. Finally, α_i is the random effect for each child. I fit this random effect model using the method of restricted maximum likelihood (REML).

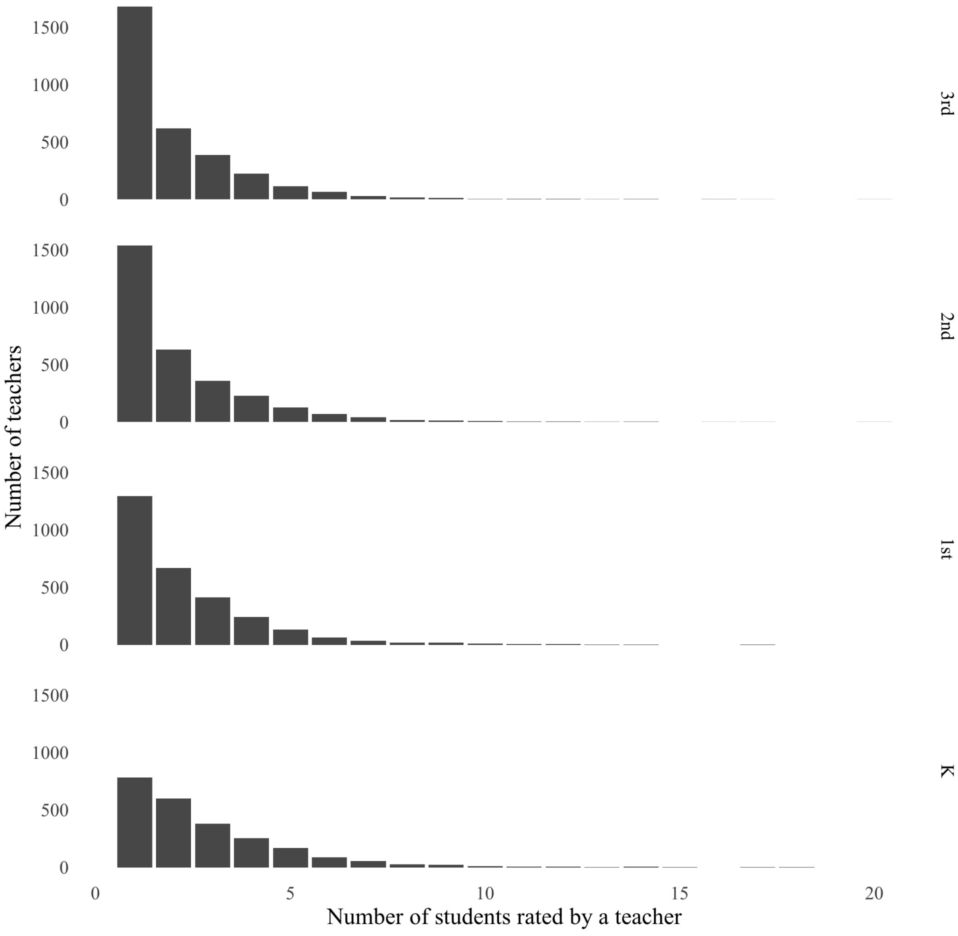
A central consideration for using the model specified in (3.4) is whether there is sufficient variation within students to draw meaningful inference about the effect of a change in class size on measures relevant to instruction. Sufficient variance within student needs to exist both in the outcomes, y_{igs}, and the deviations from predicted class size, C^*_{ig}. Low intraclass correlations reported in Table 3.3 suggest sufficient variation within child for the outcomes. Intraclass correlation for deviations from class size are low, with a value of 0.33. A representation of year to year deviations is found in Figure 3.7. This figure shows students with a given deviation in one year often have a different deviation the following year. Low ICC for both outcomes relevant to instruction and deviations from class size indicates a promising amount of variation exists to run the model specified in (3.4).



Source: ECLS-K 2010-2013.

Figure 3.7: Year to year deviations in class size

One concern about using the measures relevant to instruction available in ECLS teacher surveys is the ‘rater’ effect that comes from different teachers rating the child each year. In other words, a reasonable question about year to year changes in the measure for, say, externalizing behavior is the extent to which these changes arise because of the the teacher rather than changes that arise because of another factor such as class size. Histograms in Figure 3.8 represent the number of children in the sample each teacher rated by year. This figure shows most teachers only rated one student, and relatively few teachers rated more than 5 students. Moreover, the number of teachers who completed a survey for one student grew from less than 1000 in kindergarten to more than 1500 in third grade. Because individual teachers rated so few students in the sample, estimation of a rater effect is not feasible. Variation in scores based on teacher perception are thus limited by the unknown effect of idiosyncratic teacher variation in the scores.



Source: ECLS-K 2010-2013.

Figure 3.8: Number of students rated by a teacher for each grade

A last assumption about the fixed effects model I address pertains to the effect of time invariant variables. These effects are assumed to be constant across every measurement (Allison, 2009). If the effect of time invariant variables is not constant, the interaction of these variables and time needs to be in the model. Examples of time invariant variables in my data include sex, race, ethnicity, and family SES prior to kindergarten. I consider these and other interactions as candidates for the model both to check this assumption and to respond to my third research question about the potentially heterogeneous effect of class size on socio-demographic characteristics of interest.

For the fixed effect regression models I estimate, I consider including multiple interactions in order to respond to my research questions. These interactions can severely complicate the interpretation of output from a regression model. To facilitate interpretation, I calculate the marginal effect of class size on the outcome at specific values of interest such as the contrast between male and female students.

Issues Related to Standard Errors and ECLS Data

Accurate estimation of standard errors is vital for statistical inference. Standard errors measure the uncertainty within a sample, and this measurement is necessary to determine knowledge about the population one can infer from the sample. In this subsection, I describe challenges associated with accurate estimation of standard errors in this study and how these challenges shape my analysis.

Two specific challenges relevant to estimating standard errors arise with using ECLS data. First, the sampling design NCES used and the process of collecting data for ECLS affects estimation of standard errors. Specifically, NCES identified a stratified random sample of students by first identifying primary sampling units and then school within these units for participation in data collection. Moreover, NCES over sampled Asian and Pacific Islander students and encountered typical non-response and attrition over time in the sample of ECLS students (Tourangeau et al., 2012). To account for oversampling and non-response in the data, NCES weighted observations to reflect representation of the entire population in the ECLS data. Ordinary calculation of standard errors does not account for this weighting and can yield incorrect inference.

Second, ECLS data has a hierarchical structure that is common in education research. This structure arises from the fact that children attend a classrooms within a school, and multiple schools make up a district, which is one of many districts within a state that has specific laws and policies governing education. Ordinary calculation of standard errors does not account for this structure and tends to underestimate the true standard error in the sample.

I address challenges associated with calculating standard errors by using fixed effects, design effects, and clustering standard errors. State level fixed effects in my models provide a measure of state level policies affecting class size and instruction and they serve as a proxy for stratified random sampling NCES utilized in ECLS data collection. Design effects provide a strategy for understanding the relationship between standard errors calculated using ordinary methods and standard errors adjusted for sampling strategy and attrition. Given that I chose to use the sample of students who participated in all four years of data collection, the weights associated with my sample had a design effect of 1.47 (Tourangeau et al., 2016). This value means standard errors that account for the ECLS sampling strategy and attrition are, in general, 1.47

times standard errors calculated using ordinary methods. I use this design effect to calculate an adjusted weight according to the recommendations from NCES (ibid.). With this weight, estimated standard errors in my model account for the sampling strategy and attrition in my sample.

I also cluster standard errors at the school level in order to account for the hierarchical structure of schools. I select the school level for clustering because NCES randomly selected schools for participation in the study. After selecting schools, NCES randomly selected 20 students within the school for participation, so clustering at the school level reflects the relationship among students in the data who attended the same school.

Although up to 20 students in the sample could attend the same school, Figure 3.9 shows that most schools only have one student in the sample after kindergarten. This phenomenon occurs because a subsample of children selected in each school end up participating in every year of data collection and because children move to different schools between grades. When children move to a different school, they often end up as the only ECLS participant in their school. Because a limited number of children attend each school and there are many schools in the ECLS data, the impact of clustering at the school level in the estimation of standard errors may be limited. Nonetheless, I cluster standard errors at the school level in order to obtain conservative standard errors.

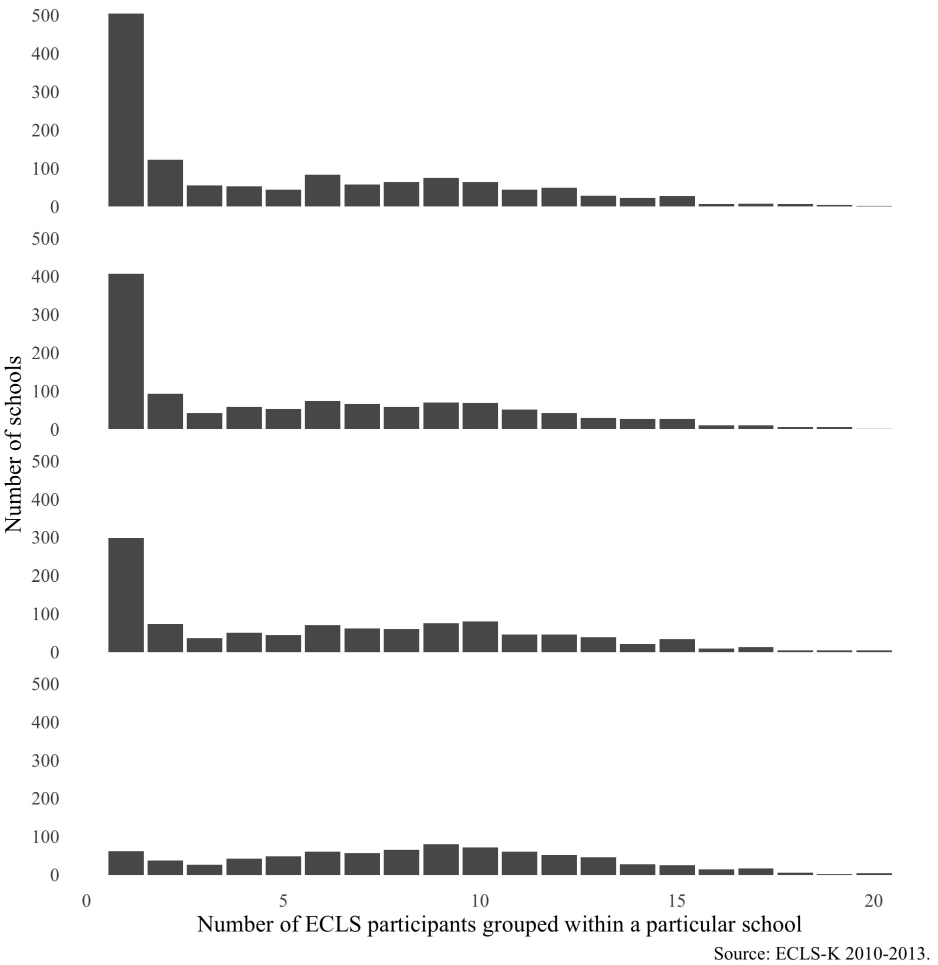


Figure 3.9: Concentration of students within schools by grade for the ECLS data

Handling Missing Data

Although the weighting scheme I apply in my analysis limits the volume of missing data, some missingness still exists in the ECLS dataset. That is, the NCES weighting scheme eliminates missingness due to non-participation, but item-level missingness (due to, e.g., a participant skipping a survey question) remains in the data. I use multiple imputation to account for these missing values (Little & Rubin, 2002). Other options for handling missing values include using only observations for which complete data is available or plugging in a value (e.g., the mean) for missing data and adding an indicator that the value was originally missing. Multiple imputation has at least two advantages over other methods for handling missingness in the data set.

First, analyzing all available cases with multiple imputation rather than elimination cases with incomplete data preserves an argument for external validity. One strength of using ECLS data to analyze the relationship between class size and instruction is that the data comprise a random sample of students from the United States. Because of this sampling design, the results are arguably generalizable to all students in the United States. Eliminating observations for missingness, however, undermines this argument for generalizabiliity. Using multiple imputation to analyze all cases maintains the argument that the sample of students in the ECLS data are representative of the population of United States students.

Second, multiple imputation leads to more robust results than others methods for handling missing values. Specifically, other methods are only valid under extremely restrictive assumptions, yield untrustworthy standard errors, or both (Van Buuren, 2018). By generating multiple sets of the data, multiple imputation accounts for uncertainty in the imputed values to address issues associated with other methods.

I implement multiple imputation of the ECLS data using a decision tree method and accounting for the structure of the data (Van Buuren, 2018). Decision trees are a non-parametric method that can accommodate any type of data. By using a non-parametric method, I avoid imposing any specific functional form on the variables in my data in the process of imputing missing values. For each variable with missing values, I use the other variables to create a decision tree and identify a plausible value to use in place of the missing data. I also account for the longitudinal and hierarchical structure of the data by grouping the data by child and using fixed effects at the school level.

My data consists of five imputed sets constructed using 20 iterations for each set. Having five imputed data sets available enables me to observe the extent to which the uncertainty in the imputed data affects the results of my analysis. Repeated iterations within each data set refine the imputed values. Researchers have suggested that 10 to 20 iterations normally provides adequate values (Van Buuren, 2018).

Limitations

Prior to presenting any results, I acknowledge some limitations of this study. These limitations arise due to shortcomings of data, scope, and analytic design.

Although the ECLS dataset has potential to provide useful information about the relationship between class size and instruction, it is less than ideal in some ways.

One limitation of all the measures of instruction contained in ECLS is that NCES collected them by administering an annual survey. Research suggests that surveys can yield poor quality information about how much time teachers spend on particular topics or using particular practices as well as the quality of instruction (Mayer, 1999). Despite the limitations of surveys, some researchers have used them effectively to study instruction (e.g., Cohen & Hill (2001)). Further, my measures of instruction are somewhat atypical in that they focus less on the work of teaching and more on relationships to occur during instruction. That is, some problems with surveys for studying teaching include that teachers may have trouble reporting particular practices they used and the language researchers use to describe practice may not align with how teachers think about practices. I limit concern about such problems by using scales that require teachers to recall their relationship and experiences with particular students.

Concerning generalizability of this study, I have made an effort to conduct analysis on a sample from the 2010 cohort of kindergartners who are representative of the national cohort. Nonetheless, the generalizability of my results is compromised by issues such as attrition and the age of ECLS data. Over time, students dropped out of ECLS data collection for a variety of reasons. Although NCES employed strategies to identify a representative sample of students among those who persisted throughout data collection, there is not way to know for certain whether attrition occurred in a systematic rather than random fashion. Also, given the dynamic nature of school policy and practices, findings pertaining to the 2010 cohort of kindergartners may not provide reliable information about future cohorts.

Another limitation of the ECLS dataset is the lack of measures of classroom level characteristics and other relevant components of the instructional triangle. That is, there are no measures from, for example, administrative data for characteristics such as prior student achievement in the class, the percent of students in the class who are eligible for subsidized lunch, or the percent of students who have an IEP. Other information that could be valuable in my study such as school policies regarding curriculum, student discipline, and teacher professional development are also not present in ECLS data. Inasmuch as such information would provide insight into the choices teachers make in the nature of relationships in the classroom, this missing information potentially introduces bias into my results.

A final limitation I note pertains to the scope of this study. Specifically, most of my analysis is limited observed changes in the instruction students experience that occurred in connection with changes in class size. Within child, the standard deviation for class size in early elementary school was approximately four. With observed changes, this small, my analysis cannot do a good job of capturing the potential impact of large changes in class size. That is, natural changes in class size like the ones I observe in the ECLS data cannot provide good information about the effect of large changes (say a change of 10 students) on the instruction students experience in schools. As a consequence, extrapolating any of these results to what might happen if schools were to hire more teachers to reduce average class size would introduce labor market considerations that are outside the scope of my work. Further, teachers' practice in small classes could change with opportunities to learn and develop professionally in ways that I do not detect in this study. That is, teachers may be capable of changing their instruction in productive ways when they have the support to do so, but my analysis will not capture these possibilities.

A limitation of my BART analysis is endogenous variation in class size. Although I can predict outcomes for all observations under the condition that one variable, such as class size, changes, these predictions are based on information from observational data. As such, these predictions are subject to variation in class size that occurs purposefully in pursuit of some goal. More to the point, school administrators could decide to manipulate class size in order to promote prosocial behavior among students. Because I have not isolated variation in class size that occurs independent of such decision making, my estimates of the relationship between class size and measures relevant to instruction lack a plausibly claim to causality.

CHAPTER 4

Results

I now present results from the regression analyses I conducted. In this section, I limit the presentation to direct interpretation of the regression results. Following this section, I interpret results across models and form a response to my research questions in the discussion section.

This section is organized in two subsections. First, I consider socio-demographic characteristics associated with class size in kindergarten through third grade from two perspectives–first considering each grade in isolation followed by considering the experience students have over time throughout early elementary school. These results are relevant to my first research question about the characteristics of students who enroll in large and small classes. Second, I present results pertaining to the relationship between class size and instruction, which is relevant to my second and third research questions. In particular, my results provide information about the relationship between class size and teachers use of achievement groups, student reports of pro-social behavior, student participation in classroom activities, students' relationship with their peers, and other measures relevant to the instructional triangle (Cohen et al., 2003) as described in the framework for this study.

The Relationship Between Socio-Demographic Characteristics and Class Size

Cross-Sectional Analysis

I begin my presentation of results with with a cross-sectional analysis of the relationship between class size and socio-demographic characteristics. Results of the linear regression are available in Table 4.1. None of the central variables of interest for this study were significant explanatory variables for class size. That is, results indicate that sex, race, ethnicity, and family SES do not explain differences in class size in which students were enrolled. Moreover, interactions involving these variables were not statistically significant.

Table 4.1: Results from linear regression for the relationship between socio-demographic characteristics and class size in kindergarten.

	Kindergarten	1st Grade	2nd Grade	3rd Grade
Female	0.018	0.009	-0.1	0.039
	(0.112)	(0.097)	(0.097)	(0.116)
Black/African American	-0.161	-0.269	-0.157	-0.288
	(0.289)	(0.261)	(0.249)	(0.307)
Hispanic	0.318	0.377*	0.173	-0.133
	(0.213)	(0.186)	(0.194)	(0.231)
Asian or Pacific Islander	-0.074	-0.043	-0.303	-0.533
	(0.656)	(0.307)	(0.324)	(0.373)
Other race	0.053	0.15	0.163	-0.033
	(0.283)	(0.258)	(0.216)	(0.23)
Family SES	0.033	0.142	0.069	0.133
	(0.114)	(0.093)	(0.096)	(0.101)
Public school	2.126**	1.455	1.162	2.116**
	(0.685)	(0.777)	(0.658)	(0.726)
Suburb	-0.985*	-0.407	-0.418	0.162
	(0.436)	(0.336)	(0.3)	(0.354)
Town/Rural	-1.054**	-1.063**	-0.839**	-0.542
	(0.393)	(0.325)	(0.315)	(0.316)
300 - 499 students	1.479**	1.628**	2.418***	1.946***
	(0.526)	(0.521)	(0.459)	(0.426)
500+ students	2.384***	2.616***	3.424***	3.128***
	(0.491)	(0.473)	(0.439)	(0.447)
Received SPED	-0.773*	-0.632*	-1.197***	-0.73**
	(0.354)	(0.306)	(0.276)	(0.263)
K entry age (months)	0.002	-0.002	0.014	-0.006
	(0.015)	(0.014)	(0.014)	(0.015)
Diagnosed disability	-0.152	-0.303	-0.039	-0.338
	(0.123)	(0.159)	(0.165)	(0.176)
Student moved		0.064	-0.106	1.2*
		(0.126)	(0.125)	(0.501)
N	6740	6740	6740	6740
R2	0.245	0.301	0.311	0.236
Adj R2	0.238	0.295	0.305	0.229

Note: State fixed effects, teacher characteristics, student household characteristics, and an interaction between race and school size are included in every model. Standard errors are clustered at the school level, adjusted for heteroskedasticity, and weighted to account for design effects in the ECLS surveys. Source: ECLS-K 2010-2013. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

School characteristics had a statistically significant relationship with class size. School size was a significant predictor class size. For example, children who attended a school of 300 to 499 students, on average, were enrolled in classes with 1.48 to 2.42 more students than children who attend a school with fewer than 300 students. School locale was also a significant predictor of class size for each grade level. For kindergarten through second grade, students who attended school in a town or rural area were enrolled in smaller classes, on average, than their counterparts in urban schools. This results was also true for kindergarten students in suburban schools. Finally, school type had a significant relationship with class size, though this relationship was statistically significant only in kindergarten and third grade. In both of these grades, students in public schools were enrolled in classes and average of more that two students bigger than their peers in private schools.

In every grade, receipt of special education services had a significant relationship with class size regardless of model specification. Specifically, students who receive special education services, on average, were enrolled in a smaller class than students who do not receive these services. The magnitude of this relationship across early elementary school had a range from -0.632 to -1.197.

Event History Analysis

I now report the results of an event history analysis regarding student enrollment in at least one large or small class at some point in early elementary school. These results provide evidence about school and student characteristics associated with greater or lesser odds of enrollment in a either a large or small class. As a reminder, I use the threshold of greater than 25 students to define a large class and less than 17 students to define a small class. I first present results related to enrollment in a large class followed by results related to enrollment in a small class.

Event history analysis results for enrollment in a large class are found in Table 4.2. This table contains results from six regression models. Three of these models have only time constant covariates, and three of the models allow time-varying covariates. Within each group of three models, Model 1 omits clustering of standard errors at the school level. Models 2 and 3 include clustering of standard errors at the school level to account for the relationship in class enrollment among students who attend the same school. In addition to clustering standard errors at the school level, model 3 also includes a random effect for each student. This random effect accounts for the possibility that students have unobserved characteristics that affect the probability they enroll in a large class at some point in early elementary school.

Table 4.2: Event history analysis of enrollment in a large class at least once during early elementary school

	Time constant			Time varying		
	Mod 1	Mod 2	Mod 3	Mod 1	Mod 2	Mod 3
1st grade	0.86	0.86	0.86	0.82*	0.82	0.82
	(0.07)	(0.11)	(0.11)	(0.07)	(0.1)	(0.11)
2nd grade	0.69***	0.69**	0.69**	0.65***	0.65**	0.65**
	(0.07)	(0.1)	(0.1)	(0.06)	(0.09)	(0.09)
3rd grade	0.78*	0.78	0.78	0.75**	0.75*	0.75*
	(0.08)	(0.12)	(0.12)	(0.08)	(0.11)	(0.11)
Female	0.97	0.97	0.97	0.98	0.98	0.98
	(0.06)	(0.04)	(0.04)	(0.06)	(0.04)	(0.04)
API	1.2	1.2	1.2	1.19	1.19	1.19
	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)
Black/African American	1.02	1.02	1.02	1.06	1.06	1.06
	(0.12)	(0.14)	(0.14)	(0.12)	(0.14)	(0.14)
Hispanic	1.1	1.1	1.1	1.13	1.13	1.13
	(0.12)	(0.1)	(0.1)	(0.12)	(0.11)	(0.1)
Other race	1.09	1.09	1.09	1.13	1.13	1.13
	(0.16)	(0.12)	(0.12)	(0.17)	(0.13)	(0.12)
Pre-K family SES	1.02	1.02	1.02	1.04	1.04	1.04
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Public school	1.76***	1.76*	1.76*	2.21***	2.21**	2.21**
	(0.22)	(0.45)	(0.45)	(0.3)	(0.62)	(0.6)
City	1.2*	1.2	1.2	1.25*	1.25	1.25
	(0.11)	(0.19)	(0.19)	(0.12)	(0.2)	(0.2)
Town/rural	0.73**	0.73*	0.73*	0.79*	0.79	0.79
	(0.07)	(0.11)	(0.11)	(0.08)	(0.12)	(0.12)
300-499	1.14	1.14	1.14	1.3*	1.3	1.3
	(0.13)	(0.24)	(0.24)	(0.15)	(0.26)	(0.26)
500+	2.08***	2.08***	2.08***	2.28***	2.28***	2.28***
	(0.23)	(0.42)	(0.42)	(0.26)	(0.45)	(0.44)
Observations	16570	16570	16570	16570	16570	16570
AIC	5933.19	5933.19	5933.19	5921.19	5921.19	5921.19

Note: Point estimates are odds ratios. Standard errors are in parentheses. All models include state fixed effects. For time constant analysis, all variables are held constant at the first observed value. For the time varying analysis, variables that can vary over time are allowed to do so. For example, school population is held at the value of the first school each student attended for the time constant analysis. School population takes on the value corresponding to the school each student attended in a particular year for the time varying analysis. State fixed effects, teacher characteristics, and student household characteristics are included in every model. Source: ECLS-K 2010-2013. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

To examine the relationship between student and school characteristics and enrollment in at least one large class between kindergarten and third grade, I focus on the results from the time varying model 3. Although I prefer one model above the others, I include results from all models in Table 4.2 to show the sensitivity of statistical inferences to modeling choices. Inference from all six models has many similarities, but I prefer time varying model 3 for a few reasons. First, Table 4.2 provides reason to prefer the time varying models over the time constant models. Specifically, AIC for the time varying models is less than that for the time constant models. This lower value indicates the models allowing time varying covariates explain more variation in enrollment in a large class.

Second, looking across the models contained in Table 4.2, there is a noticeable difference between difference between models 2 and 3 compared to model 1. Specifically, clustering standard errors at the school level results in changes in the estimate of the standard error. For example, the standard error for 1st grade is estimated to be 0.07 in model 1, and this estimate is 0.1 for models 2 and 3. These observed differences in estimated standard errors are empirical evidence of the intuitive idea that adjusting for the relationship among children in the same school is important in a study of class size.

Third, accounting for unobserved heterogeneity among students is valuable based on observed patterns in large class enrollment. Specifically, the portion of students who enroll in a large class for the first time is less in first through third grade than in kindergarten as seen in Figure 3.4. This pattern suggests that students who enroll in a large class in kindergarten could have unobserved characteristics associated with greater risk of enrolling in a large class than their counterparts. Including a random effect for each child accounts for such heterogeneous risk if it exists.

Interpreting model 3 from Table 4.2, no student characteristics of central interest in this study had a statistically significant relationship with enrollment in a large class. That is, race, ethnicity, sex, and family SES all had statistically insignificant relationships with enrollment in a large class at least one time in early elementary school. As in the cross-sectional analysis, students who attended a public school were more likely to enroll in a large class than students who attended a private school. Also, students who attended schools with 500 or more students had greater odds of enrolling in a large class at some point in early elementary school compared to students who attended a school with less than 300 students.

Results from event history analysis of enrollment in a small class at some point in early elementary school are in Table 4.3. I built models for this analysis in the same way as for the preceding analysis, and Table 4.3 is organized in the same way as Table 4.2. I prefer time varying Model 3 for the same reasons given in the previous analysis, though this model has no substantive differences compared to time varying Model 2.

Table 4.3: Event history analysis of enrollment in a small class at least once in early elementary school

	Time constant			Time varying		
	Mod 1	Mod 2	Mod 3	Mod 1	Mod 2	Mod 3
1st grade	0.46***	0.46***	0.46***	0.49***	0.49***	0.49***
	(0.04)	(0.07)	(0.07)	(0.05)	(0.07)	(0.07)
2nd grade	0.22***	0.22***	0.22***	0.23***	0.23***	0.23***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)
3rd grade	0.21***	0.21***	0.21***	0.22***	0.22***	0.22***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Female	1.05	1.05	1.05	1.05	1.05	1.05
	(0.08)	(0.06)	(0.06)	(0.08)	(0.06)	(0.06)
API	1.22	1.22	1.22	1.21	1.21	1.21
	(0.25)	(0.28)	(0.28)	(0.25)	(0.27)	(0.27)
Black/African American	1.7***	1.7***	1.7***	1.68***	1.68***	1.68***
	(0.22)	(0.23)	(0.23)	(0.21)	(0.24)	(0.24)
Hispanic	1.04	1.04	1.04	1.01	1.01	1.01
	(0.13)	(0.11)	(0.11)	(0.13)	(0.11)	(0.11)
Other race	0.96	0.96	0.96	0.92	0.92	0.92
	(0.17)	(0.16)	(0.16)	(0.17)	(0.16)	(0.16)
Pre-K family SES	0.81**	0.81***	0.81***	0.81***	0.81***	0.81***
	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)
Public school	0.45***	0.45***	0.45***	0.45***	0.45***	0.45***
	(0.06)	(0.1)	(0.1)	(0.06)	(0.1)	(0.1)
City	0.89	0.89	0.89	0.91	0.91	0.91
	(0.1)	(0.15)	(0.15)	(0.1)	(0.15)	(0.15)
Town/rural	1.15	1.15	1.15	1.15	1.15	1.15
	(0.14)	(0.2)	(0.2)	(0.14)	(0.19)	(0.19)
300-499	0.86	0.86	0.86	0.65***	0.65*	0.65*
	(0.11)	(0.18)	(0.18)	(0.08)	(0.13)	(0.13)
500+	0.45***	0.45***	0.45***	0.33***	0.33***	0.33***
	(0.06)	(0.09)	(0.09)	(0.04)	(0.07)	(0.07)
Observations	20330	20330	20330	20330	20330	20330
AIC	4654.14	4654.14	4654.14	4608.95	4608.95	4608.95

Note: Point estimates are odds ratios. Standard errors are in parentheses. All models include state fixed effects. For time constant analysis, all variables are held constant at the first observed value. For the time varying analysis, variables that can vary over time are allowed to do so. For example, school population is held at the value of the first school each student attended for the time constant analysis. School population takes on the value corresponding to the school each student attended in a particular year for the time varying analysis. State fixed effects, teacher characteristics, and student household characteristics are included in every model. Source: ECLS-K 2010-2013. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

For enrollment in a small class at some point in early elementary school, one student characteristic of central interest in this study was statistically significant. Family SES prior to kindergarten, however, had a statistically significant relationship with the odds students enroll in a small class during early elementary school. When controlling for school characteristics and other relevant factors a one standard deviation increase in family SES was associated with an average decrease of 19 percentage points in the odds of enrolling in a small class. As a group, race and ethnicity did not have a significant relationship with class size ($p > 0.05$), so I do not interpret the point estimate for Black/African American students in Table 4.3.

To summarize results from the analysis of socio-demographic characteristics associated with class size in the United States, school characteristics appear to be more consistently have a significant relationship with class size than student characteristics of central interest in this study.

The Relationship Between Class Size and Instruction

I now turn to a presentation of results related to my second and third research questions. These results fall into two broad categories based on the data available for analysis. First, I present results from analysis of cross sectional data about differences in teachers' choices and students' behavior. These results provide insight into the relationship between class size and instruction for a population of students at a particular point in time. More precisely, the measures relevant to instruction available for this analysis pertain to teachers' choices regarding use of achievement groups in reading and mathematics. Second, I present results from an analysis of longitudinal data changes in the ways in which students are perceived by their teachers. These results provide insight into the relationship between class size and instruction for a student over time. More precisely, the measures relevant to instruction available for this analysis pertain to teacher observations of students' relationships, habits, and demonstrations of emotion in the classroom.

Teachers' Use of Achievement Groups and Students' Prosocial Behavior

One observation researchers have made about the relationship between class size and instruction is related to the use of groups. Specifically, some researchers have claimed larger classes are associated with teachers sorting students into more groups within the class (see, e.g., Blatchford et al. (2001)). ECLS data provides information about the number of achievement groups for reading and mathematics teachers use in their classroom. I begin this subsection by reporting analysis results related to these groupings.

For each school year from kindergarten to third grade, I used a Bayesian Additive Regression Trees model to predict the probability teachers use a number of reading achievement groups between zero and five or more. A representation of results from this analysis is available in 4.1. Inspection of this figure reveals a couple noteworthy patterns. First, for kindergarten through third grade, the mean probability teachers use zero achievement groups was roughly constant with respect to class size. In other words, as class size increases, the probability teachers work with the entire class rather than with some number of achievement groups was approximately the same. Further, for all grades and all class sizes, this probability was greater than 10% and less than 25%. Also, zero reading achievement groups was never the most probable teacher choice for classroom organization.

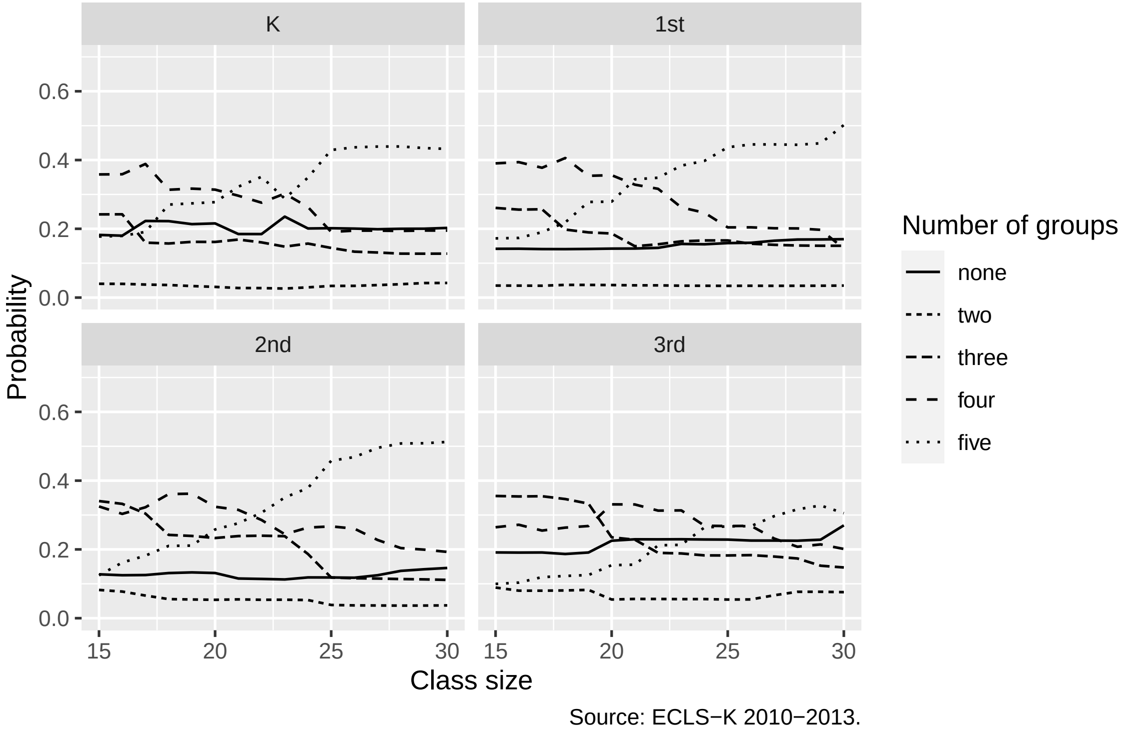
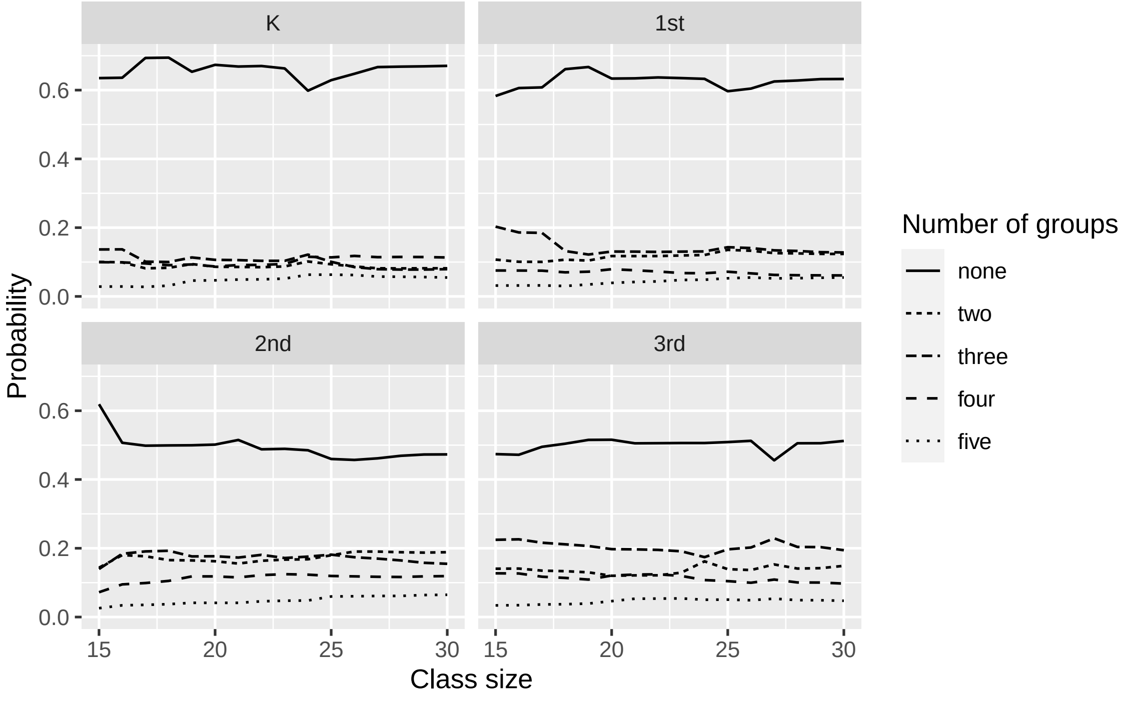


Figure 4.1: Probability teachers use zero to five or more achievement groups in reading by class size.

Second, for kindergarten through second grade, an increase in class size was associated with an increase in mean probability that teachers use five or more reading achievement groups. For a class size of 15 students, this probability was between 10% and 20%. For a class size of 30, the mean probability of teachers using five or more achievement groups was greater than 40%. Also, for each of these grades, five or more reading achievement groups was the most probable teacher choice for classroom organization in classes with at least 23 students. Corresponding to the increasing probability that teachers use five or more achievement groups as class size increases is a decreasing probability that teachers use three or four achievement groups in the classroom.

Third grade shows a somewhat different pattern than kindergarten through second grade for use of achievement groups. In this grade, the probability teachers use five or more achievement groups in a class of 30 students was always less than 40. Further, for classes of 15 to 19 students, the probability teachers use three achievement groups was greater than 30 percent.

Teacher use of achievement groups in mathematics during early elementary school, represented in Figure 4.2, had fairly consistent patterns for all grades. For kindergarten through third grade, the most probable number of achievement groups was zero regardless of class size. Moreover, this mean probability was greater than 40% for all grades and all class sizes.



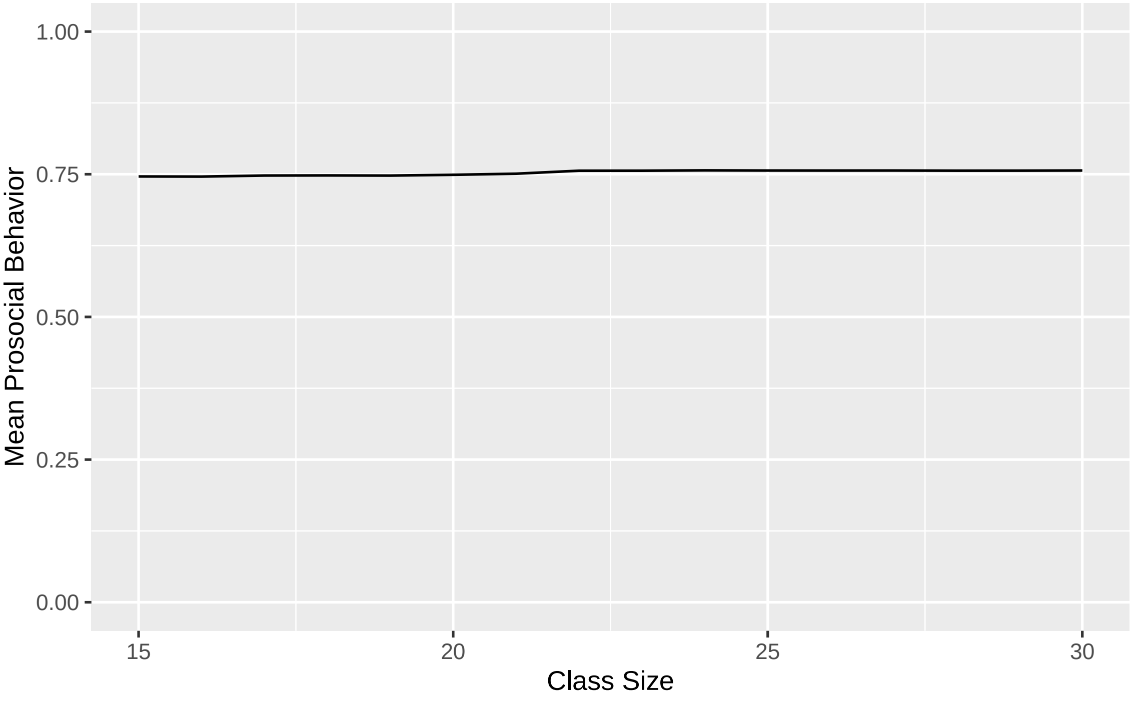
Source: ECLS−K 2010−2013.

Figure 4.2: Probability teachers use zero to five or more achievement groups in mathematics by class size.

Not only was the most probable number of achievement groups in mathematics zero, the least probable number was five or more. Once again, this result was true for all grades and regardless of class size. Together, these patterns suggest a rather limited use of achievement groups in mathematics throughout early elementary school.

Transitioning away from teachers choices regarding use of achievement groups, I now present results related to students' reports of prosocial behavior. As a reminder, this scale was based on student responses to three items such as "I try to cheer up other classmates who are upset or sad about something." These results are useful for gathering evidence relevant to the hypothesis that student relationships with each other are associated with class size.

Student reports of prosocial behavior had no meaningful relationship with class size. As shown in Figure 4.3, which is based on analysis using Bayesian Additive Regression Trees, mean predicted prosocial behavior was almost constant with respect to class size. Specifically, using the BART model to predict prosocial behavior if all students enrolled in a class of size 15 yielded an average of 0.746 on the prosocial behavior scale. Replacing class size enrollment with 30 yields a mean prosocial behavior score of 0.757.



Source: ECLS−K 2010−2013.

Figure 4.3: Predications of prosocial behavior by class size based on analysis using Bayesian Additive Regression Trees

Cross-sectional analysis using BART found some evidence of differences in teachers' practice associated with class size. Specifically, analysis showed some evidence that teachers' use of achievement groups was associated with class size. On the other hand, analysis showed similarities in use of achievement groups in mathematics regardless of class size. Also, students reported similar levels of prosocial behavior in third grade regardless of class size.

Changes in the Ways Students are Perceived by their Teachers

In this subsection I present results from fixed effects regression analysis of the relationship between teachers' perceptions of students and class size. A focus of this analysis is exploration of heterogeneous effects of changes in class size on teachers' perceptions. Theory and existing empirical evidence suggest changes in class size might have a different effect on different groups (e.g., Lazear, 2001). For this study, I am interested in whether these effects differ by student sex, race, ethnicity, and family SES.

To facilitate interpretation of the results, I scale the class size variable to units of four students. I choose to scale changes in class size to four students because four is the whole number nearest to a one standard deviation change in class size as seen in Figure 3.6. This rescaling is useful to understand changes in teachers' perceptions of students associated with more noticeable changes in class size (e.g., 23 students compared to 19) than a small change of one student. Also, using a larger unit of change in class size makes interpretation of small effects on instructional scales more comprehensible than interpreting values near zero.

Results from regressions for each outcome are contained in Table 4.4. Class size deviation is the main independent variable of interest in these models and it represents the differences between a child's predicted class size and observed class size in a particular year. By capturing this difference, the variable provides an estimate of the relationship between year-to-year changes in class size during early elementary school and relationships that occur during instruction.

Table 4.4: Results of fixed effects regression analysis for five outcomes pertainting to relationships that occur during instruction.

	Outcome				
	Closeness	Conflict	Interpersonal skills	Externalizing behavior	Internalizing behavior
Class size deviation	-0.007	0.001	-0.006	0.001	-0.001
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Female	0.06***	-0.075***	0.085***	-0.085***	-0.007
	(0.004)	(0.006)	(0.006)	(0.006)	(0.005)
Asian/Pacific Islander	-0.026*	-0.009	-0.008	-0.023	-0.023
	(0.011)	(0.015)	(0.016)	(0.016)	(0.012)
Black/African American	-0.013	0.049***	-0.039***	0.048***	-0.011
	(0.007)	(0.01)	(0.01)	(0.011)	(0.008)
Hispanic	-0.012	-0.012	0.012	-0.019	-0.012
	(0.007)	(0.009)	(0.01)	(0.01)	(0.007)
Other	-0.021*	0.009	0.005	-0.001	0.003
	(0.01)	(0.013)	(0.014)	(0.014)	(0.01)
Pre-K SES	0.015***	-0.019***	0.027***	-0.023***	-0.015***
	(0.003)	(0.004)	(0.005)	(0.005)	(0.004)
Asian/Pacific Islander X Class size deviation	-0.002	0.001	0.006	0.001	0.001
	(0.009)	(0.009)	(0.01)	(0.008)	(0.009)
Black/African American X Class size deviation	0.005	-0.007	0.006	-0.006	-0.009
	(0.006)	(0.006)	(0.007)	(0.006)	(0.006)
Hispanic X Class size deviation	0.001	0.003	0.004	0.002	0.001
	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)
Other X Class size deviation	0.001	-0.004	0.013	-0.007	-0.017
	(0.01)	(0.01)	(0.012)	(0.01)	(0.01)
Pre-K SES X Class size deviation	0.002	-0.001	0.002	0.001	0
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Female X Class size deviation	0.01**	-0.007	0.009*	-0.007	-0.001
	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)
Intercept	0.815***	0.28***	0.558***	0.405***	0.243***
	(0.046)	(0.059)	(0.064)	(0.064)	(0.049)

Note: Point estimates come from a linear regression model. Standard errors are in parentheses. All models include controls for student, teacher, school, classroom, and household characteristics as well as state-grade fixed effects. Source: ECLS-K 2010-2013. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Each of these models includes an interaction between change in class size and family SES, sex, race, and ethnicity. These interactions are in the models regardless of significance in order respond to my third research question about the potentially heterogeneous relationship between class size and relationships that occur during instruction. Including interaction terms in each model makes direct interpretation of results from the model challenging. To illustrate, results in Table 4.4 show a significant relationship between class size and teacher rating of closeness with the student, but this estimate only pertains to white, male students whose family had average SES prior to kindergarten because of the interactions terms included in the model. For this reason, I interpret results from the model by calculating average marginal changes in the outcomes for a one unit change in class size.

Calculation of marginal changes in relationships occurring during instruction based on results from fixed effects regression analysis showed class size did not have a statistically significant relationship with any outcomes. Further, not only was the point estimate for the relationship between class size and teacher ratings of students near zero, the plausible range of values for this relationship was also near zero. For all the scales in Table 4.5, the standard error of of the estimate was less than 0.003. Using this standard error yields a 95% confidence interval for the estimate of each relationship that contains only points less than 0.01 in magnitude. For the internalizing behavior scale, as an example, a 95% confidence interval for the estimate of the relationship with class size is the interval $(-0.006, -0.001)$ Thus, results of the fixed-effect regression analysis indicate with high confidence that class size is associated with less than an average change of 0.01 points on scales ranging from zero to one for these five measures relevant to classroom instruction.

Table 4.5: Average marginal effect of changes in class size on five outcomes pertaining to relationships that occur during instruction

	Closeness	Conflict	Interpersonal skills	Externalizing behavior	Internalizing behavior
Class size deviation	-0.001	-0.003	0.001	-0.003	-0.003
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)

Note: Standard errors are in parentheses. Marginal effects of class size are based on results of fixed effects regression analysis. Source: ECLS-K 2010-2013. *** p < 0.001; ** p < 0.01; * p < 0.05

Moving beyond the average relationship, fixed effects regression analysis provided evidence of a different relationship between class size and closeness for male and female students. Table 4.6 contains p-values for tests of joint significance for interactions of race, ethnicity, family SES, and sex with class size on each of the five measures relevant to instruction. This table indicates a statistically significant interaction was present for teacher reports of closeness and interpersonal skills based on student sex. All other p-values were greater than 0.05, which indicates that class size had the same relationship with measures relevant to instruction regardless of race and family SES.

Table 4.6: P-values for the interaction of class size with variables of central interest in this study.

	Closeness	Conflict	Interpersonal skills	Externalizing behavior	Internalizing behavior
Family SES	0.408	0.777	0.538	0.693	0.904
Sex	0.009	0.087	0.037	0.054	0.885
Race/Ethnicity	0.950	0.627	0.770	0.670	0.266

Note: Values less than 0.05 provide evidence of a heterogeneous relationship between class size and the corresponding measure of instruction. Source: ECLS-K 2010-2013.

Accounting for other factors such as household and other student characteristics, the marginal effect of a 4 student change in class size was associated with a positive change in teacher rated closeness for female students and a negative change for male students. Specifically, an increase of four students in class size was associated with an average increase of 0.003 points on the closeness scale for females students. Male students, on the other hand, had a 0.005 point average decrease on the closeness scale for the same change in class size. Thus, an increase of four students in class size was associated with a statistically significant increase in the difference between scores for female and male students by approximately 0.008 points.

Although somewhat afield of the focus of this study, some findings from Table 4.4 are noteworthy separate from class size. Specifically, I feel compelled to bring attention to some significant differences in measures relevant to instruction along the lines of sex and race when controlling for class size, teacher views of teaching and students, household characteristics, and other relevant factors. While setting aside for a moment the relationship between class size and measures of instruction, results displayed in Table 4.4 are relevant to the varying experiences students have in the classroom along lines of race and sex, which is central to this study. These results therefore provide some useful context for further discussion in the subsequent section.

Across the board, male students received from their teachers less desirable ratings than their female counterparts. Compared to female students, male students had lower closeness scores, higher conflict scores, lower interpersonal skill scores, higher externalizing behavior scores, and higher internalizing behavior scores. For all but the internalizing behavior scale, the magnitude of these differences was greater that 0.05 on a scale ranging from 0 to 1.

In some cases, students of color received from their teachers less desirable ratings than their White counterparts. In particular, Black/African American students received higher conflict and externalizing behavior scores and lower interpersonal skills scores than White students. The magnitude of this difference ranged from 0.039 points to 0.049 points on a scale from 0 to 1.

Combining statistically significant differences from Table 4.4 across race and sex provides an estimate of the different experiences White female and Black/African American male students had during instruction. For example, White female students, on average, received a closeness score of 0.861 on a scale ranging from 0 to 1. Black/African American studnets had an average score of 0.787 on the same scale. As another example, White female students had an average externalizing behavior score of 0.273, and Black/African American students had an average score of 0.411 on a scale from 0 to 1.

In this section, I have reviewed the results from the analyses I conducted to respond to my research questions. Thus far, I have made an effort to provide only a direct interpretation of output from the models. I now turn to interpreting and discussing the results in the contexts of my research questions and the broader literature.

CHAPTER 5

Discussion, Limitations, and Future Directions

In this section, I situate the results of this study within the context of theory and hypotheses I laid out previously. While interpreting these results in their broader context, I urge caution according to the limitations of the methods I used to conduct the research. In alignment with these limitations, I suggest potentially insightful areas of future research on the relationship between class size and instruction. After highlighting what I think are the most salient results from my research in this manner, I offer some parting remarks regarding my current thinking on the topic of class size and instruction.

With respect to my first research question, results from my study indicate there are limited direct relationships between class size and race, ethnicity, or family SES. I limit this statement to direct effects because there is a possibility that other factors associated with class size disproportionately affect students along lines of race and family SES. Specifically, school characteristics including school size, urbanicity, and public status were significant predictors of class size. Referring back to the bi-variate relationships reviewed in the methods section, chi-squared and ANOVA analysis showed these school characteristics are associated with SES as well as race and ethnicity.

Therefore, on a localized level controlling for relevant characteristics, there are limited differences in the size of class in which students enroll based on race and family SES. Nonetheless, inasmuch as, for example, students of color attend large, urban public schools at higher rates than White students, they may very well enroll in classes of greater size than White students.

Beyond any indirect relationships between class size and family SES or race, there is one direct effect of note. Specifically, event history analysis showed Black/African

American students and students whose family has below average SES are more likely than their White and higher SES counterparts to enroll in a small class at some point in early elementary school. One reason this phenomenon could occur is receipt of targeted services that takes place in small classes. Although my analysis accounted for students who receive special education services, there is a possibility these services also pull in non-special education students of color and students from economically disadvantaged backgrounds. Further investigation of the phenomenon observed in my study requires more detailed data than is available from ECLS.

One thing the event history analysis adds that the cross sectional analysis cannot is the finding that kindergarten is a risk for both entering into a small class and a large class. This means that, although the mean class size is fairly close together from kindergarten through 3rd grade (Figure 3.2) and students' class size varies from year to year (Figure 3.3), those who enroll in a small or large class are likely to do it for the first time in kindergarten.

With respect to my second and third research questions, I found some statistically significant relationships between class size and measures relevant to instruction, but these relationships were small enough to be of little practical interest. Moreover, because of the number of relationships I tested, statistically significant findings could be due to chance rather than actual differences in the population. These small relationships include interactions of change in class size with sex. To be more precise, all of the relationships I found between class size and instruction had a point estimate less than 0.01 on a scale from 0 to 1 for a change in class size of four students. I argue that such a small change in scores on measures of instruction associated with change in observed class size is evidence that, in terms of the variables available in the ECLS dataset, class size did not have a meaningful relationship with the measures of instruction available in ECLS.

Despite small magnitudes in the relationship between class size and measures relevant to instruction, some patterns in (the lack of) relationships are noteworthy. These patterns are noteworthy for both theoretical and practical reasons. That is, the results of this study push against some common thinking about class size and provide detail and texture beyond the findings of previous studies.

First, my results showed a heterogeneous relationship between class size and closeness for male and female students. Although the magnitude of this differential relationship was small, the direction is intriguing. Namely, the relationship between change in class size and closeness went in opposite directions for male and female students. A finding in which both groups received lower (or higher) scores–but one group to a greater extent than the other–in connection to observed increases in class size would draw less interest for me. A differential relationship in opposite directions, however, leads me to wonder about the mechanisms that drive such a finding. What roles do male and female students take on in larger and smaller classes that might explain teachers responding to them in opposite ways in terms of closeness? To what extent might male and female students take on societal gender roles when the teacher's work load might increase with more students in the classroom? What affect do these roles have on the academic work male and female students are available to do in the classroom? These sorts of questions flow naturally from the instructional triangle as represented in Figure 2.1 (Ball, 2018; Cohen et al., 2003) and could lead to understanding about ways the broader environment might influence instruction and potentially reproduce undesirable aspects of our society.

Second, in the relationships between class size and instruction I studied, I found no significant interaction with race, ethnicity and family SES. This result stands somewhat in contrast with suggestions from prior research. Previous researchers who have studied class size have more or less agreed that class size reduction policy is likely to most benefit, in terms of educational outcomes, students who are members of historically disenfranchised groups (e.g., Hanushek (1998)). Supposing such differential benefits are real, it stands to reason that instruction is the mechanism whereby students obtain these benefits. In any event, the measures relevant to instruction available in ECLS data did not provide evidence about the details of instruction that might benefit students who identify with historically disenfranchised groups. Granted, I am using ECLS data for secondary analysis, and researchers who design data collection for a careful study of class size and the mechanisms whereby it improves educational outcomes will be able to make decisions about aspects of instruction most likely to have a relationship with class size.

Third, I found evidence of limited change in teacher practice. Hoxby (2000) suggested teachers are unlikely to change their practice in light of transient changes in class size. My analysis of teachers use of achievement groups provides some support for this suggestion. Specifically, class size had no relationship within a grade with the probability that teachers choose to use achievement groups or not in mathematics and reading. Viewed from that binary perspective, my study provides some empirical support for Hoxby's thought. Moreover, the probability with which teachers used a particular number of achievement groups in mathematics was fairly constant within each grade with respect to class size. This observation provides more support for the notion that teachers do not change their practice in repose to class size

From another perspective, however, teachers may change their practice. In kindergarten through third grade, the probability with which teachers chose to use, say, five achievement groups increased in connection with an increase in class size. That is, teachers may not have, for example, a curriculum for low, medium, and high ability readers that they use regardless of class size. Instead teachers might refine the groupings based on the students in their class, which is affected by class size. Still, teachers might use the exact same curriculum and just create two of each group. For now, the point is that teachers may very well change their practice in some ways and not in others in response to changes in class size. These decisions are likely based on content area, student needs, and other factors relevant to instruction. Again referring to the instructional triangle (Cohen et al., 2003), future research might explore how teachers make decisions about (not) changing their practice in response to environmental factors such as transient changes in class size due to natural variation in the population of students.

To note, a limitation of the ECLS data for studying teacher response to transient changes in class size is the design of following a cohort of students through early elementary school. One byproduct of following a student cohort is obtaining repeated cross-sections of teachers. Further, these cross-sections of teachers are not representative of the population of teachers, for they are only the teachers who happen to teach a representative sample of students. A data set focused on a representative sample of teachers is both necessary for answering questions about teacher practice in the United States and, to my knowledge, rare or non-existent.

Fourth, findings in this study did not detect a relationship between class size and classroom community. hooks (1994) proposed that small classes are necessary to enact pedagogy that prioritizes well being. Ladson-Billings (2009) wrote about the strong sense of community in the classroom of successful teachers of students of color, and I hypothesized that, in light of documented effects of class size on educational outcomes for students belonging to historically marginalized groups, class size might correspond to the sense of community in a classroom. Within my framework for this study, I situated ideas about community in the relationship between students. ECLS data contained several measures of this relationship, but none of the available measures provided good evidence for differences in relationships among students based on class size. To be clear, findings from this study can do little to refute the ideas of the above authors. Instead, my work can only serve to document that empirical evidence for these ideas may not be so easy to come by that it flows from secondary data analysis of the sort I conducted here. In fact, I suspect empirical evidence for the relationship between class size and complex ideas like classroom community may only come from work carefully and directly designed to find it.

One way future researchers might study relationships among students and understand its relationship to class size is by incorporating social network analysis methods. These methods are well suited to studying groups of a size commonly found in classrooms, and have potential to offer profound insight into the relationships students form in the classroom. Social network analysis can, for example, help researchers understand the principles–such as reciprocity–that seem to dictate the formation of social ties among a group of people. Using Social network analysis methods, researchers might study the relationship between class size and the likelihood of students reciprocating social ties within the classroom. Differences in the likelihood of reciprocation based on class size could provide empirical evidence for the sense of community that may or may not be present in smaller and larger classes.

As an aside, the analysis I conducted with ECLS data is not relevant to existing theory about class size and student behavior. Lazear (2001) suggested an advantage of smaller rather than larger classes could be that the probability the behavior of a particular student at a particular time distracts from academic content is less in a small class than in a large class. Externalizing behavior in the ECLS data, however, cannot test Lazear's theory because ECLS measures occur at the student level, and Lazear's theory pertains the the entire class. From my findings, an individual students' externalizing behavior may decrease, on average, in connection with an increase in class size, but the probability of at least one student in the class engaging in disruptive behavior at a particular time could still be greater in the larger class than in the smaller class.

To this point, the limitations I have mentioned have focused on particular measures available in ECLS data. I now mention a more general limitation of the analysis I conducted that is important to bear in mind in any interpretation of the results. First, all the measures I used, with the exception of student reports of prosocial behavior, were based on teacher perceptions of students. That is, my measure of, say, interpersonal skills was based entirely on teacher perception. While researchers have validated these scales for use, they are nonetheless limited in the information they can offer. I have mentioned that the statistically significant relationships I found between class size and measures of instruction were small in practical terms. Given that these measures were limited to teacher perception, however, valid questions about students actual experiences during instruction remain open. At the same time, teachers' perceptions of students are relevant to the experience students have in schools, and my analysis provides evidence that these perceptions change little, if at all, in association with class size.

Reflecting across all the work done in this study, I remain curious about the relationship between class size and the experience children have during instruction. Admittedly, a reasonable policy take away from this work is that, if class size matters for the instruction students experience in school, it matters little enough that there are likely other factors that deserve higher priority. Nonetheless, I suspect that researchers are able to design future work on the relationship between class size and instruction that is targeted at likely areas of difference more so than the secondary data analysis I have conducted here. I have suggested a few possible avenues for this research to pursue, and I am sure there are others. Such research could lead to findings that provide insight into how policy might leverage class size to improve instruction for students.

Ultimately, my curiosity stems from the persistent attention education stakeholders–including teachers, parents, and others–continue to give to class size. I am confident my experience as a parent that I used to introduce my research problem resonates with many other parents. I have observed class size take a prominent position in labor disputes leading to teacher strikes. Even though reasearchers' efforts, including my own, have not yielded conclusive ideas about how class size affects instruction, I suggest the issue seems to be enough on the mind of people close to instruction that researchers have a responsibility to figure out what is going on and provide answers about how to use class size to benefit students.

APPENDIX: Regression Results with Original Data

Table 5.1: Results from linear regression for the relationship between socio-demographic characteristics and class size in kindergarten.

	Kindergarten	1st Grade	2nd Grade	3rd Grade
Female	0.105	-0.011	-0.187	-0.072
	(0.13)	(0.102)	(0.1)	(0.125)
Black/African American	-0.219	-0.308	-0.418	-0.45
	(0.314)	(0.285)	(0.299)	(0.3)
Hispanic	0.167	0.326	0.104	-0.172
	(0.259)	(0.194)	(0.208)	(0.235)
Asian or Pacific Islander	-0.449	-0.086	-0.502	-0.68
	(0.73)	(0.324)	(0.356)	(0.4)
Other race	0.227	0.389	-0.05	-0.208
	(0.337)	(0.271)	(0.229)	(0.242)
Family SES	0.06	0.172	0.118	0.173
	(0.125)	(0.095)	(0.102)	(0.11)
Public school	1.858*	1.063	0.966	1.874*
	(0.743)	(0.796)	(0.66)	(0.732)
Suburb	-0.533	-0.281	-0.573	0.155
	(0.484)	(0.347)	(0.316)	(0.358)
Town/Rural	-1.438***	-1.214***	-1.382***	-1.116***
	(0.426)	(0.323)	(0.32)	(0.306)
300 - 499 students	1.339*	1.44**	2.31***	1.932***
	(0.585)	(0.532)	(0.434)	(0.437)
500+ students	2.309***	2.461***	3.246***	2.996***
	(0.565)	(0.48)	(0.419)	(0.459)
Received SPED	-0.92*	-0.563	-1.29***	-0.744**
	(0.404)	(0.32)	(0.292)	(0.268)
K entry age (months)	0.005	-0.004	0.007	-0.008
	(0.017)	(0.015)	(0.015)	(0.016)
Diagnosed disability	-0.098	-0.279	0.112	-0.274
	(0.146)	(0.165)	(0.163)	(0.186)
Student moved		0.028	-0.085	1.387**
		(0.129)	(0.135)	(0.514)
N	5630	6290	6220	6200
R2	0.215	0.284	0.299	0.226
Adj R2	0.209	0.278	0.294	0.22

Note: State fixed effects, teacher characteristics, student household characteristics, and an interaction between race and school size are included in every model. Standard errors are clustered at the school level, adjusted for heteroskedasticity, and weighted to account for design effects in the ECLS surveys. *** p < 0.001; ** p < 0.01; * p < 0.05

Table 5.2: Event history analysis of enrollment in a large class at least once during early elementary school

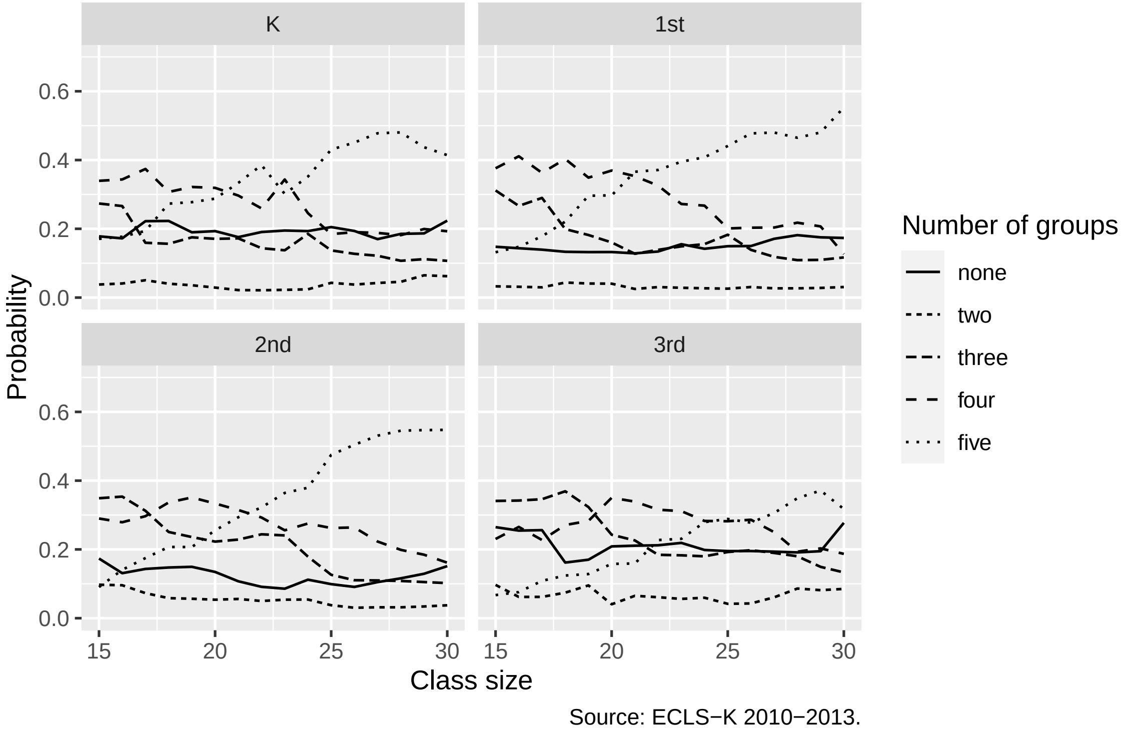
	Time constant			Time varying		
	Mod 1	Mod 2	Mod 3	Mod 1	Mod 2	Mod 3
1st grade	0.86	0.86	0.86	0.79**	0.79	0.79
	(0.08)	(0.12)	(0.12)	(0.07)	(0.1)	(0.11)
2nd grade	0.68***	0.68*	0.68*	0.62***	0.62**	0.62**
	(0.07)	(0.1)	(0.1)	(0.06)	(0.09)	(0.09)
3rd grade	0.76*	0.76	0.76	0.72**	0.72*	0.72*
	(0.09)	(0.12)	(0.12)	(0.08)	(0.11)	(0.11)
Female	0.97	0.97	0.97	0.97	0.97	0.97
	(0.07)	(0.05)	(0.05)	(0.07)	(0.05)	(0.04)
API	2.53	2.53**	2.53**	1.9	1.9	1.9
	(1.22)	(0.86)	(0.86)	(1.04)	(0.71)	(0.71)
Black/African American	1	1	1	1.33	1.33	1.33
	(0.3)	(0.36)	(0.36)	(0.4)	(0.45)	(0.46)
Hispanic	1.96*	1.96*	1.96*	2.22**	2.22**	2.22**
	(0.56)	(0.58)	(0.58)	(0.62)	(0.67)	(0.67)
Other race	1.25	1.25	1.25	1.44	1.44	1.44
	(0.5)	(0.41)	(0.41)	(0.6)	(0.48)	(0.5)
Pre-K family SES	1	1	1	1.02	1.02	1.02
	(0.06)	(0.05)	(0.05)	(0.06)	(0.05)	(0.05)
Public school	1.82***	1.82*	1.82*	2.35***	2.35**	2.35**
	(0.26)	(0.48)	(0.48)	(0.35)	(0.68)	(0.67)
City	1.37	1.37	1.37	1.11	1.11	1.11
	(0.38)	(0.48)	(0.48)	(0.25)	(0.4)	(0.4)
Town/rural	1.24	1.24	1.24	0.97	0.97	0.97
	(0.34)	(0.42)	(0.42)	(0.22)	(0.35)	(0.35)
300-499	1.34	1.34	1.34	1.73***	1.73*	1.73*
	(0.21)	(0.34)	(0.34)	(0.28)	(0.42)	(0.43)
500+	2.41***	2.41***	2.41***	2.81***	2.81***	2.81***
	(0.37)	(0.6)	(0.6)	(0.44)	(0.69)	(0.69)
Observations	13940	13940	13940	15050	15050	15050
AIC	4909.39	4909.39	4909.39	5323.88	5323.88	5323.88

Note: Standard errors are in parentheses. All models include state fixed effects.

Table 5.3: Event history analysis of enrollment in a small class at least once in early elementary school

	Time constant			Time varying		
	Mod 1	Mod 2	Mod 3	Mod 1	Mod 2	Mod 3
1st grade	0.45***	0.45***	0.45***	0.48***	0.48***	0.48***
	(0.05)	(0.06)	(0.06)	(0.05)	(0.07)	(0.07)
2nd grade	0.22***	0.22***	0.22***	0.24***	0.24***	0.24***
	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)
3rd grade	0.2***	0.2***	0.2***	0.22***	0.22***	0.22***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)
Female	1.06	1.06	1.06	1.02	1.02	1.02
	(0.09)	(0.06)	(0.06)	(0.09)	(0.06)	(0.06)
API	1.37	1.37	1.37	1.31	1.31	1.31
	(0.31)	(0.33)	(0.33)	(0.29)	(0.31)	(0.31)
Black/African American	1.77***	1.77***	1.77***	1.89***	1.89***	1.89***
	(0.25)	(0.26)	(0.26)	(0.26)	(0.29)	(0.29)
Hispanic	1.07	1.07	1.07	1.08	1.08	1.08
	(0.15)	(0.12)	(0.12)	(0.15)	(0.12)	(0.12)
Other race	0.98	0.98	0.98	0.95	0.95	0.95
	(0.19)	(0.17)	(0.17)	(0.18)	(0.17)	(0.17)
Pre-K family SES	0.82**	0.82***	0.82***	0.81**	0.81***	0.81***
	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Public school	0.46***	0.46**	0.46**	0.46***	0.46***	0.46***
	(0.07)	(0.11)	(0.11)	(0.07)	(0.11)	(0.11)
City	1.41	1.41	1.41	1.85	1.85	1.85
	(0.52)	(0.8)	(0.8)	(0.64)	(0.78)	(0.78)
Town/rural	1.39	1.39	1.39	1.84	1.84	1.84
	(0.51)	(0.79)	(0.79)	(0.63)	(0.77)	(0.77)
300-499	0.83	0.83	0.83	0.62***	0.62*	0.62*
	(0.11)	(0.18)	(0.18)	(0.08)	(0.12)	(0.12)
500+	0.45***	0.45***	0.45***	0.33***	0.33***	0.33***
	(0.06)	(0.1)	(0.1)	(0.04)	(0.07)	(0.07)
Observations	17080	17080	17080	18470	18470	18470
AIC	3856.75	3856.75	3856.75	4095.02	4095.02	4095.02

Note: Standard errors are in parentheses. All models include state fixed effects. For time constant analysis, all variables are held constant at the first observed value. For the time varying analysis, variables that can vary over time are allowed to do so. For example, school population is held at the value of the first school each student attended for the time constant analysis. School population takes on the value corresponding to the school each student attended in a particular year for the time varying analysis.



Source: ECLS−K 2010−2013.

Figure 5.1: Probability teachers use zero to five or more achievement groups in reading by class size.

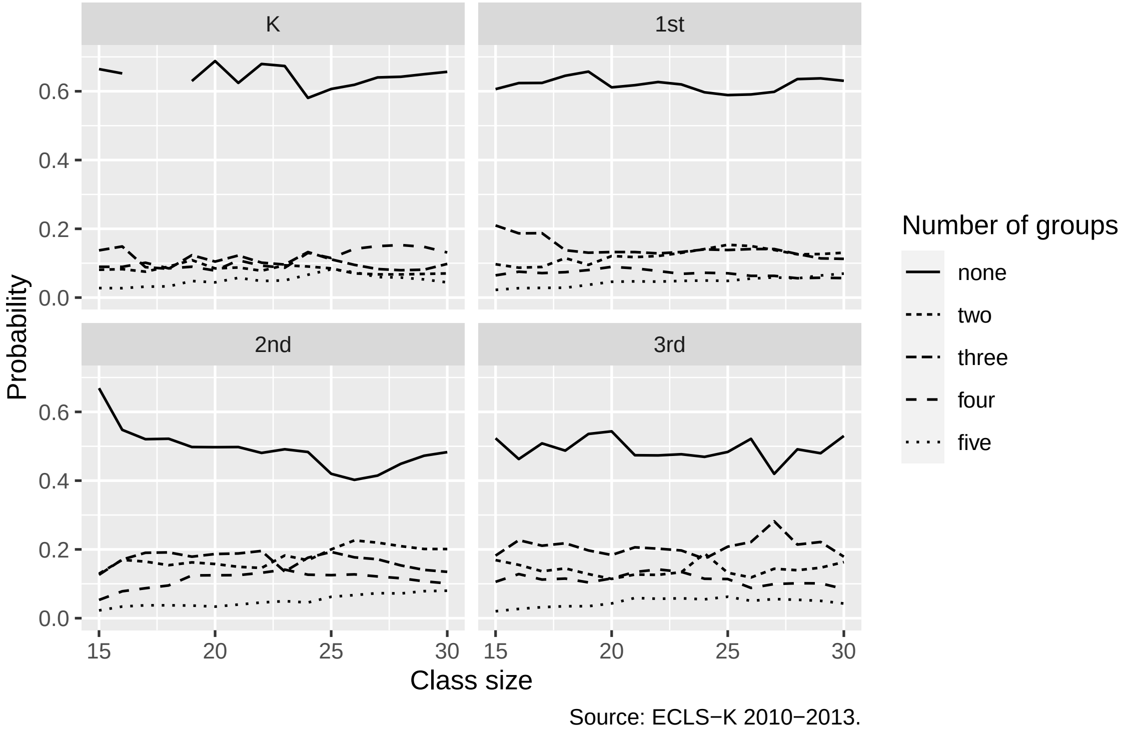


Figure 5.2: Probability teachers use zero to five or more achievement groups in mathematics by class size.

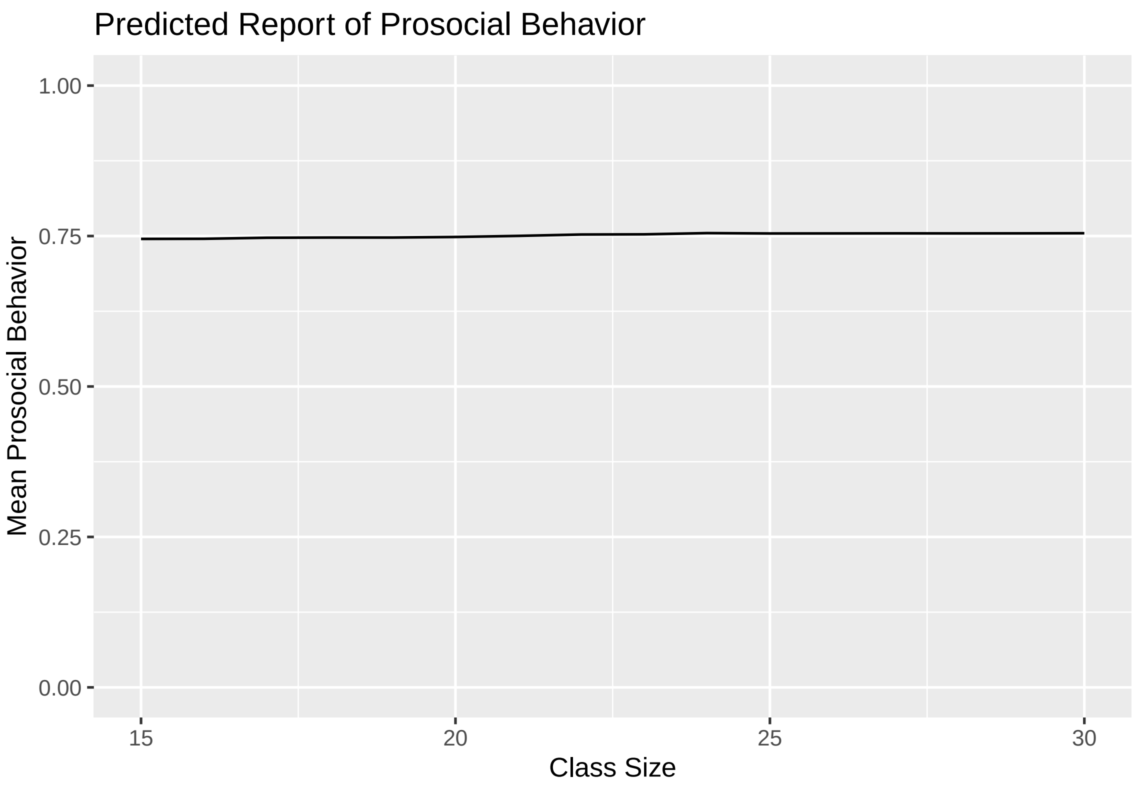


Figure 5.3: Predications of prosocial behavior by class size based on analysis using Bayesian Additive Regression Trees

Table 5.4: Results of fixed effects regression analysis for five outcomes pertainting to relationships that occur during instruction.

	Outcome				
	Closeness	Conflict	Interpersonal skills	Externalizing behavior	Internalizing behavior
Class size deviation	-0.007**	-0.004	-0.005*	0.001	-0.001
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
Female	0.062***	-0.071***	0.087***	-0.081***	-0.006*
	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)
Asian/Pacific Islander	-0.029***	-0.011	-0.001	-0.027**	-0.023***
	(0.006)	(0.008)	(0.008)	(0.009)	(0.006)
Black/African American	-0.013*	0.047***	-0.029***	0.038***	-0.015**
	(0.005)	(0.007)	(0.007)	(0.007)	(0.005)
Hispanic	-0.01*	-0.01	0.014*	-0.016**	-0.011*
	(0.004)	(0.006)	(0.006)	(0.006)	(0.005)
Other	-0.023***	0.001	0.007	-0.013	-0.002
	(0.006)	(0.008)	(0.008)	(0.008)	(0.006)
Pre-K SES	0.016***	-0.018***	0.026***	-0.021***	-0.014***
	(0.002)	(0.003)	(0.003)	(0.003)	(0.002)
Asian/Pacific Islander X Class size deviation	0.016***	-0.034***	0.047***	-0.039***	-0.041***
	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)
Black/African American X Class size deviation	0.018***	0.021***	-0.003	0.026***	0.017***
	(0.005)	(0.006)	(0.007)	(0.007)	(0.005)
Hispanic X Class size deviation	0.008**	0.013***	-0.007	0.012**	-0.002
	(0.003)	(0.004)	(0.004)	(0.004)	(0.003)
Other X Class size deviation	-0.009	0.002	-0.001	0.005	0.001
	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)
Pre-K SES X Class size deviation	0.003	-0.01*	0.007	-0.009*	-0.007
	(0.005)	(0.005)	(0.006)	(0.004)	(0.005)
Female X Class size deviation	-0.002	0.006	0.003	-0.001	0.002
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)
Intercept	-0.044*	0.045*	-0.055*	0.03	0.01
	(0.019)	(0.022)	(0.025)	(0.024)	(0.02)

Note: Point estimates come from a linear regression model. Standard errors are in parentheses. All models include controls for student, teacher, school, classroom, and household characteristics as well as state-grade fixed effects. *** p < 0.001; ** p < 0.01; * p < 0.05

REFERENCES

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). *When should you adjust standard errors for clustering?* https://doi.org/10.3386/w24003

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213–232. https://doi.org/10.1037/0033-2909.101.2.213

Allison, P. D. (2009). *Fixed effects regression models*. Thousand Oaks, CA: SAGE publications.

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*, 533–575. https://doi.org/10.1162/003355399556061

Angrist, J. D., lavy, V., Leder-Luis, J., & Shany, A. (2017). *Maimonides rule redux*. National Bureau of Economic Research.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438. https://doi.org/10.1080/10705510903008204

Ball, D. L. (2018). Just dreams and imperatives: The power of teaching in the struggle for public education. *Annual meeting of the American Educational Research Association*. New York City, NY.

Ball, D. L., & Forzani, F. M. (2007). What makes education research 'educational'? *Educational Researcher*, *36*, 529–540. https://doi.org/10.3102/0013189X07312896

Ball, D. L., Thames, M. H., Phelps, G., & others. (2008). Content knowledge for teaching: What makes it special. *Journal of Teacher Education*, *59*, 389–407. https://doi.org/10.1177/0022487108324554

Betts, J. R., & Shkolnik, J. L. (1999). The behavioral effects of variations in class size: The case of math teachers. *Educational Evaluation and Policy Analysis*, *21*, 193–213. https://doi.org/10.3102/01623737021002193

Birch, S. H., & Ladd, G. W. (1997). The teacher-child relationship and children's early school adjustment. *Journal of School Psychology*, *35*, 61–79.

Blatchford, P., Baines, E., Kutnick, P., & Martin, C. (2001). Classroom contexts: Connection between class size and within class grouping. *The British Psychological Society*, *71*, 283–302. https://doi.org/10.1348/000709901158523

Blatchford, P., Bassett, P., & Brown, P. (2011). Examining the effect of class size on classroom engagement and teacher–pupil interaction: Differences in relation to pupil prior attainment and primary vs. Secondary schools. *Learning and Instruction*, *21*, 715–730. https://doi.org/10.1016/j.learninstruc.2011.04.001

Bleich, J., Kapelner, A., George, E. I., & Jensen, S. T. (2014). Variable selection for BART: An application to gene regulation. *The Annals of Applied Statistics*, *8*, 1750–1781. https://doi.org/10.1214/14-AOAS755

Bowne, J. B., Magnuson, K. A., Schindler, H. S., Duncan, G. J., & Yoshikawa, H. (2017). A meta-analysis of class sizes and ratios in early childhood education programs: Are thresholds of quality associated with greater impacts of cognitive, achievement, and socioemotional outcomes? *Educational Evaluation and Policy Analysis*, *39*, 407–428. https://doi.org/10.3102/0162373716689489

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, *4*, 266–298. https://doi.org/10.1214/09-AOAS285

Cohen, D. K. (2011). *Teaching and its predicaments*. Cambridge, MA: Harvard University Press.

Cohen, D. K., & Hill, H. C. (2001). *Learning policy: When state education reform works*. Hillsborough, NC: Yale University Press.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*, 119–142. https://doi.org/10.3102/01623737025002119

Dee, T. S., & West, M. R. (2011). The non-cognitive returns to class size. *Educational Evaluation and Policy Analysis*, *33*, 23–46. https://doi.org/10.3102/0162373710392370

Dewey, J. (1923). *Democracy and education: An introduction to the philosophy of education.* Macmillan.

Doumen, S., Verschueren, K., Buyse, E., De Munter, S., Max, K., & Moens, L. (2009). Further examination of the convergent and discriminant validity of the Student-Teacher Relationship Scale. *Infant and Child Development*, *18*, 502–520. https://doi.org/10.1002/icd.635

Dunn, E. W., Gilbert, D. T., & Wilson, T. D. (2011). If money doesn't make you happy, then you probably aren't spending it right. *Journal of Consumer Psychology*, *21*, 115–125. https://doi.org/10.1016/j.jcps.2011.02.002

Dynarski, S., Hyman, J., & Schanzenbach, D. W. (2013). Experimental evidence on the effect of childhood investmetns on postsecondary attainment and degree completion. *Journal of Policy Analysis and Management*, *32*, 692–717. https://doi.org/10.1002/pam.21715

Education Next/PEPG Surveys. (2020). https://www.educationnext.org/wp-content/uploads/2020/07/EN-PEPG_Complete_Polling_Results-2007-19.pdf.

Englehart, J. M. (2007). Discourse in a small class: The 'diverge-converge patter' and 'relaxed freedom'. *Education*, *35*, 83–97. https://doi.org/10.1080/03004270601103467

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. https://doi.org/10.1080/01619569209538671

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal*, *27*, 557–577. https://doi.org/0.3102/00028312027003557

Flanagan, D. P., Alfonso, V. C., Primavera, L. H., Povall, L., & Higgins, D. (1996). Convergent validity of the BASC and SSRS: Implications for social skills assessment. *Psychology in the Schools*, *33*, 13–23.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods.* Sage Publications, Inc.

Graue, E., Hatch, K., Rao, K., & Oen, D. (2007). The wisdom of class-size reduction. *American Educational Research Journal*, *44*, 670–700. https://doi.org/10.3102/0002831207306755

Green, D. P., & Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly*, *76*, 491–511. https://doi.org/10.1093/poq/nfs036

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research*, *66*, 361–396. https://doi.org/10.3102/00346543066003361

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

Gresham, F. M., Elliott, S. N., Vance, M. J., & Cook, C. R. (2011). Comparability of the Social Skills Rating System to the Social Skills Improvement System: Content and psychometric comparisons across elementary and secondary age levels. *School Psychology Quarterly*, *26*, 27–44. https://doi.org/10.1037/a0022662

Gresham, F. M., MacMillan, D. L., Bocian, K. M., Ward, S. L., & Forness, S. R. (1998). Comorbidity of hyperactivity-impulsivity-inattention and conduct problems: Risk factors in social, affective, and academic domains. *Journal of Abnormal Child Psychology*, *26*, 393–406.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis*, *19*, 141–164. https://doi.org/10.3102/01623737019002141

Hanushek, E. A. (1998). *The evidence on class size.* Rochester, NY: W. Allen Wallis Institute of Political Economy.

Hanushek, E. A. (1999). Some findings from an independent investigation of the tennessee star experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis*, *21*, 143–163. https://doi.org/10.3102/01623737021002143

Hanushek, E. A. (2003). The failure of input-based schooling policies. *The Economic Journal*, *113*, F64–F98. https://doi.org/10.1111/1468-0297.00099

Herbst, P. G. (2003). Using novel tasks in teaching mathematics: Three tensions affecting the work of the teacher. *American Educational Research Journal*, *40*, 197–238. https://doi.org/10.3102/00028312040001197

Herbst, P., Nachlieli, T., & Chazan, D. (2011). Studying the practical rationality of mathematics teaching: What goes into "installing" a theorem in geometry? *Cognition and Instruction*, *29*, 218–255. https://doi.org/10.1080/07370008.2011.556833

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*, 430–511. https://doi.org/10.1080/07370000802177235

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*, 217–240. https://doi.org/10.1198/jcgs.2010.08162

Hinkley, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics*, *19*, 285–292. https://doi.org/10.1080/00401706.1977.10489550

hooks, b. (1994). *Teaching to transgress: Education as the practice of freedom.* New York City, NY: Routledge.

Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics*, *115*, 1239–1285. https://doi.org/10.1162/003355300555060

Jackson, P. W. (1990). *Life in classrooms.* New York, NY: Teachers College Press.

Krueger, A. B. (2002). Understanding the magnitude and effect of class size on student achievement. In L. Mishel & R. Rothstein (Eds.), *The class size debate* (pp. 7–35). Washington, D.C.: Economic Policy Institute.

Ladson-Billings, G. (2009). *The dreamkeepers: Successful teachers of African American children* (2nd ed.). San Francisco, CA: Josey-Bass.

Lazear, E. P. (2001). Educational production. *The Quarterly Journal of Economics*, *116*, 777–803. https://doi.org/10.1162/00335530152466232

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815–852. https://doi.org/10.1177/1094428106296642

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Lortie, D. C. (1975). *Schoolteacher: A sociological study.* University of Chicago Press.

McDonald, M., Kazemi, E., Kelley-Petersen, M., Mikolasy, K., Thompson, J., Valencia, S. W., & Windschitl, M. (2014). Practice makes practice: Learning to teach in teacher education. *Peabody Journal of Education*, *89*, 500–515. https://doi.org/10.1080/0161

956X.2014.938997

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Morton, K., & Riegle-Crumb, C. (2019). Who gets in? Examining inequality in eighth-grade algebra. *Journal for Research in Mathematics Education*, *50*, 529–554. https://doi.org/10.5951/jresematheduc.50.5.0529

Pedersen, J. A., Worrell, F. C., & French, J. L. (2001). Reliabiilty of the Social Skills Rating System with rural Appalachian children from families with low incomes. *Journal Fo Psychoeducational Assessment*, *19*, 45–53. https://doi.org/10.1177/073428290101900103

Pianta, R. C., & Steinberg, M. (1992). Teacher-child relationships and the process of adjusting to school. *New Directions for Child Development*, *57*, 61–80. https://doi.org/10.1002/cd.23219925706

Pong, S.-l., & Pallas, A. (2001). Class size and eighth-grade math achievement in the United States and abroad. *Educational Evaluation and Policy Analysis*, *23*, 251–273. https://doi.org/10.3102/01623737023003251

Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, *36*, 672–683. https://doi.org/10.1080/07350015.2016.1247004

Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics*. Hoboken, NJ: John Wiley & Sons.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*, 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Rocková, V., & Saha, E. (2018). On theory for BART. *arXiv Preprint arXiv:1810.00787*.

Shalaby, C. (2017). *Troublemakers: Lessons in freedom from young children at school*. New York City, NY: The New Press.

Silverman, J., & Thompson, P. W. (2008). Toward a framework for the development of mathematical knowledge for teaching. *Journal of Mathematics Teacher Education*, *11*, 499–511. https://doi.org/10.1007/s10857-008-9089-5

Stasz, C., & Stecher, B. M. (2000). Teaching mathematics and language arts in reduced size and non-reduced size classrooms. *Educational Evaluation and Policy Analysis*, *22*, 313–329. https://doi.org/10.3102/01623737022004313

Tan, J., Y. V. And Roy. (2019). Bayesian additive regression trees and the general BART model. *Statistics in Medicine*, *38*, 1–22. https://doi.org/10.1002/sim.8347

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., Hagedorn, M. C., Daly, P., & Najarian, M. (2012). *Early childhood longitudinal study, kindergarten class of 2010-11 (ECLS-K: 2011): User's manual for the ECLS-K:2011 kindergarten data file and electronic codebook (NCES 2013-061).* Washington, D. C.: U.S. Department of Education, National Center for Education Statistics.

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Vaden-Kiernan, N., Blaker, L., & Najarian, M. (2016). *Early childhood longitudinal study, kindergarten class of 2010-11 (ECLS-K: 2011): User's manual for the ECLS-K:2011 kindergarten-third grade data file and electronic codebook, restricted version (NCES 2016-092).* Washington, D. C.: U.S. Department of Education, National Center for Education Statistics.

Van Buuren, S. (2018). *Flexible imputation of missing data.* Boca Raton, FL: CRC Press.

Van der Oord, S., Van der Meulen, E. M., Prins, P. J. M., Oosterlaan, J., Buitelaar, J. K., & Emmelkamp, P. M. (2005). A psychometric evaluation of the Social Skills Rating System in children with Attention Deficit Hyperactivity Disorder. *Behaviour Research and Therapy*, *43*, 733–746. https://doi.org/10.1016/j.brat.2004.06.004

Westat. (n.d.). *Spring 2014 teacher questionnaire child level.* Washington, D. C.: U.S. Department of Education, National Center for Education Statistics.