2018

Classroom controversies: the academic impact of charter schools, suspension bans, and ability groups

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

CLASSROOM CONTROVERSIES: THE ACADEMIC IMPACT OF CHARTER SCHOOLS, SUSPENSION BANS, AND ABILITY GROUPS

by

DOMINIC MASTERMAN ZARECKI

B.A., Yale University, 2009
M.A., Boston University, 2013

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2018

Approved by

First Reader _____

Douglas Kriner, Ph.D.
Professor of Political Science
Pennsylvania State University, College of the Liberal Arts

Second Reader _____

Katherine Levine Einstein, Ph.D.
Assistant Professor of Political Science

Third Reader _____

Dino Christenson, Ph.D.
Associate Professor of Political Science

“An intelligent heart acquires knowledge,

and the ear of the wise seeks knowledge.”

Proverbs 18:15 (English Standard Version)

ACKNOWLEDGMENTS

This dissertation would have been impossible without the aid and support of many people. The students and faculty of the Political Science department at Boston University provided a very encouraging environment to learn about social science research. I would like to specifically thank my readers - Dino Christenson, Katie Einstein, David Glick, and Doug Kriner – for their feedback and guidance. As my first reader, Doug Kriner provided thoughtful guidance that profoundly improved this dissertation. Special thanks also go to Taylor Boas and John Gerring for their assistance early in my dissertation journey, as well as to Greg Winger for his insightful feedback.

A few other institutions provided crucial support. As an undergraduate at Yale University, Ian Shapiro's "Moral Foundations of Politics" course convinced me to pursue a major in Ethics, Politics, and Economics. Yale's Religion and Politics Colloquium – in particular Philip Gorski, Sigrun Kahl, and Vivek Sharma – made me excited about pursuing graduate school. My time at the California Charter Schools Association – where I worked for three years after earning my Master's Degree – planted the seeds of the first two chapters of this dissertation. Elizabeth Robitaille and Allison Kenda provided excellent leadership, and it was a pleasure to "learn the ropes" of education data work alongside Leilani Cannon, Elyce Martinez, Jonathan Slakey, and Sheila Xiao. My current employer, Fortune School of Education, provides an inspiring example of an education agency that uses research to improve outcomes for the students we serve. Thank you to the CEO, Margaret Fortune, as well as my boss, Matt Taylor, for their vision, passion, and hard work that formed the basis of chapter three.

Two groups motivated me to make education policy the topic of my dissertation. One is the Technical Design Group, a seven-member committee that provides statistical advice to the Academic Accountability Unit of the California Department of Education – an amazing team led by the indefatigable Jenny Singh. A special thank you to members Ed Haertel and Christine Hikido for their very helpful feedback. The other group is the Strategic Data Project, a two-year fellowship connected to Harvard University's Center for Education Policy Research. The staff, speakers, and "fellow fellows" in Cohort 8 give me reason to be optimistic about the future of education research.

Finally, I thank my family. My kids, Eliora and J.P., are a source of joy, and my parents and parents-in-law have provided selfless support that enabled me to finish. Most importantly, my wife has been a pillar of support and keen advisor. When I had nearly made up my mind to quit graduate school, my wife convinced me that I needed to finish and that we would make it work. She encouraged me to follow my gut and switch dissertation topics, and she wisely guided me toward what is now the third chapter of this dissertation. I thank God for bringing her and everyone mentioned above into my life.

CLASSROOM CONTROVERSIES: THE ACADEMIC IMPACT OF CHARTER SCHOOLS, SUSPENSION BANS, AND ABILITY GROUPS

DOMINIC MASTERMAN ZARECKI

Boston University Graduate School of Arts and Sciences, 2018

Major Professor: Douglas Kriner, Professor of Political Science, *Pennsylvania State University*

ABSTRACT

Education policy is frequently in the crosshairs of ideological disagreement. This dissertation analyzes three controversial policies over which elected school boards often have control: charter schools, suspension bans, and ability groups.

How do charter schools impact district academic growth? Researchers typically focus on large districts with many charter schools, but the most common experience is an average-sized district shifting from no charters to one. A difference-in-differences design analyzing a decade of charter expansion in California reveals that impact is contingent on charter type: locally funded charters (i.e. affiliated with the district) lead to either static or decreased growth while directly funded charters (i.e. independent of the district) lead to higher academic growth.

Many policymakers have banned or limited suspensions for all but the most serious offenses. The 2013 suspension ban in Los Angeles Unified School District provides a natural experiment; it led to a substantial, 0.2 standard deviation decrease in academic growth among middle schools that had previously issued the banned suspensions. Four subsequent suspension bans – in San Francisco, Pasadena, Oakland,

and (grades K-3) all of California – also appear to have harmed academic growth. Simultaneously, suspension bans have an uncertain relationship with dropout rates, the primary mechanism by which bans are meant to impact the school-to-prison pipeline. Instead of banning suspensions, policymakers should carefully test other efforts that decrease suspension and dropout rates without harming academic growth.

Finally, educators have utilized between-class ability grouping – sorting students in one grade into different classes by prior ability – for over a century. Proponents rely on a previously untested mechanism: decreased classroom dispersion in prior academic ability allows teachers to target their instruction more narrowly. This final paper measures classroom dispersion directly for the same students over four trimesters. Multivariate regressions and multilevel models evaluate the relationship between classroom dispersion and academic growth while controlling for other classroom characteristics as well as student, teacher, and school effects. Analyses reveal that English classrooms with less dispersion in prior ability experience slightly less growth. However, there is a trade-off: between-class ability grouping improves equity at the expense of overall academic growth.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CADRE .....Community Asset Development Re-defining Education

LAUSD ..... Los Angeles Unified School District

LCSC.....Labor Community Strategy Center

TCE .....The California Endowment

INTRODUCTION

Overview

This dissertation analyses the academic impact of three controversial education policies. Increased academic growth would greatly enhance economic development (Hanushek et al. 2015) and reduce the disparity in achievement that exists between racial groups (e.g. Harris and Herington 2006) and income brackets (e.g. Reardon 2013). While many people – such as parents (Galindo and Sheldon 2012), teachers (Kane et al. 2013), and superintendents (Whitehurst et al. 2014) – influence academic growth, school board members have unique levers of control over relatively large numbers of students. Of all institutional variance in student scores (i.e. differences explained by teachers, schools, and districts), school districts account for 10% of academic performance (Whitehurst et al. 2013 and 2015).

Each chapter highlights a specific way school boards can impact academic growth:

1. Districts experience static or decreased academic growth after the first locally funded charter opens, but increased growth after the first directly funded charter school opens.
2. An effort to reduce suspensions by Los Angeles Unified School District (LAUSD) caused a substantial decrease in academic growth for many of its middle schools.

3. A charter school organization's decision to sort students into English classes by their ability slightly depressed overall academic growth, but it increased growth for the lowest-achieving students.

There are two overall messages. The first is that specific decisions that school boards make can have enormous impact – positive or negative – on academic growth. The second message, unfortunately, is that school boards operate in an environment much more influenced by ideology than by evidence. Even when extant research can provide guidance, school boards make limited and selective use of this information (Asen et al. 2013; Penuel et al. 2017). More importantly, school boards rarely oversee research necessary to evaluate the impact of their specific efforts. In the struggle between ideology and information, this dissertation attempts to push the balance slightly more toward rigorous evidence.

The Research Question

How do school boards impact academic growth? While it is beyond the scope of one dissertation to analyze every way school boards can make an impact, the three papers each analyze a distinct type of action school boards can take. It is important to explain the focus on school boards as well as the specific outcome of academic growth.[1]

Understanding how to improve academic growth is important for both economic growth and social justice. An analysis for the National Bureau of Economic Research analyzing the impact of student achievement on economic growth for U.S. states finds

[1] These papers will all use standardized test score gains to operationalize academic growth. Standardized tests are imperfect measures of student learning, and they are subject to distortions that threaten the validity of the findings in these papers. However, standardized tests are the only widely available measure of academic growth that we have, and education researchers routinely use them in their work.

"enormous scope for state economic development through improving the quality of schools" (Hanushek et al. 2015). Since the publication of the Coleman Report "Equality of Educational Opportunity" in 1966, scholars have noted large disparities in academic achievement between racial groups and income brackets. In additional to being intrinsic issues of social justice, these disparities fuel income inequality, which is known to harm economic growth (Dabla-Norris et al. 2015) and trust (d' Hombres et al. 2013).

Why focus on school boards? While many actors influence academic growth, the role of school boards is a uniquely powerful combination of scope and depth. Those who may have a stronger impact, such as parents and teachers, influence far fewer numbers of students. State and national policy-makers influence many more students, but in much less direct ways. Local school boards possess specific authority over a wide range of educational matters that can dramatically impact the students they serve. A small but growing literature on local school board members finds that their racial diversity (Hughes et al. 2017), quality of interaction (Grissom 2012), and attitudes (Lorentzen 2013) influence student outcomes. The implication is that particular actions school boards take can impact academic growth.

Nevertheless, positively influencing school board behavior will take creativity. Even when they are inclined to make decisions informed by evidence, school board members operate in environments where research tends to be used to support decisions that have already been made (Asen et al. 2013; Penuel et al. 2017). Disseminating information about district performance does not impact reelection prospects or even superintendent turnover (Kogan et al. 2016a). The only effect of a very clear signal of

district underperformance – "failing AYP" (Academic Yearly Progress) in former federal accountability parlance – was that voters became less likely to approve education-related tax via referendum (Kogan et al. 2016b). Unfortunately, such direct democracy tends to result in less instructional spending, more teacher turnover, and decreased academic growth (Kogan et al. 2017). Education researchers should think carefully about the linkages between important findings and improved outcomes.

Three Chapters

This dissertation takes the form of three chapters. One analyzes how academic growth changes for districts after the first charter school opens within its boundaries. Another chapter evaluates the impact of a policy that suddenly reduced the number of suspensions in many Los Angeles schools. The final chapter estimates the effectiveness of sorting students into English classes based on their prior achievement. I choose these particular topics for three reasons.

First, each of the topics concern potential school board actions. Charter schools can open in nearly every US state, and school boards can encourage these schools to open within their borders if not authorize them directly. As for suspensions, school boards can take actions that go beyond statewide regulations, often at the behest of interest groups. Lastly, school boards can at least recommend the practice (or not) of sorting students by ability even when they cannot require that schools do so.

Second, each chapter makes a theoretical contribution to its respective literature. Research on the impact of charter schools typically focuses on a few large school districts containing many charter schools. In contrast, the first chapter of this dissertation

analyzes the far more common experience of the first charter school opening in a (usually average-sized) district. This chapter also highlights a previously overlooked distinction between two types of charter schools: those that are affiliated with their local school district, and those that operate independently. Only the latter type of charter school appears to cause increased growth for local school districts, suggesting that districts only improve if they feel competitive pressure from charter schools.

The school discipline literature contains two competing hypotheses about the expected impact of a suspension ban on schoolwide academic growth. The second chapter of this dissertation tests those hypotheses for the first five large-scale suspension bans in the United States. The ban appears to harm growth in every case. The results suggest that at least some suspensions might be better conceptualized as a symptom of other problems instead of as an independent cause of negative outcomes. Such a conceptualization would explain how trying to directly reduce suspensions (a symptom) fails to lead to positive outcomes.

The literature on between-class ability groups spans nearly a century. However, it never directly tests the primary mechanism that causes this grouping practice to impact academic growth: students in classrooms have less dispersion in their prior academic ability. In theory, less dispersion would enable teachers to target their instruction more narrowly and accurately for all students. The final chapter of this dissertation tests this classroom dispersion mechanism directly and finds that less dispersion typically leads to less academic growth. This is the opposite result of recent scholarship, and it suggests

that the results of between-class ability grouping may be dependent on factors such as curricular design and teacher training.

Third, each chapter uses a research methodology that produces more rigorous evidence than most of the existing literature. The charter school chapter alleviates concerns of omitted variable bias through two innovative analyses. The first looks at school districts where the first directly funded charter school was denied by the local school board but then approved on appeal to the county or state. The fact that these districts also increased their academic growth proves that the finding is not caused by district willingness to open a charter school. The second analysis compares directly funded and locally funded charters and finds that only directly funded ones cause an increase in district academic growth. This means that having a community that can produce a successful charter petition is insufficient to cause district growth; the charter school also needs to have independence.

The literature on behavior management uses non-experimental methods to argue that suspensions depress academic performance, both for suspended students and their non-suspended peers. The second chapter exploits a policy shift in the second largest US school district to argue that this is probably not true; LAUSD middle schools experienced a huge drop in academic growth after the school board banned suspensions based on defiance. This is based on a natural experiment within LAUSD. LAUSD schools that gave no defiance suspensions prior to the suspension ban are very similar demographically to LAUSD schools that gave a few or many defiance suspensions. However, schools that gave no defiance suspensions prior to the ban saw no reduction in

their academic growth, while other LAUSD schools experienced a substantial reduction. This chapter also relies on difference-in-differences analyses that compare LAUSD to the rest of California. While other researchers have used difference-in-differences, they had to utilize natural fluctuations in suspension rates, which opens the possibility of omitted variable bias (Perry and Morris 2013). This chapter does not suffer from those same concerns because suspension bans are the result of a sudden, forced drop in suspension rates.

Finally, the ability grouping literature typically relies on standard OLS (ordinary least squares) regression of cross-sectional student data. In contrast, the third dissertation chapter utilizes multilevel analyses to more accurately model the nested structure of the data: students within grades within teachers within schools. Additionally, having four trimesters of data for the same students allows for a test that includes student controls. This test finds that as the same student moves from a less dispersed to a more dispersed classroom, that students tends to experience slightly higher academic growth. These methodological advances provide a rationale to trust the surprising findings.

All three chapters concern a school board policy, make theoretical contributions, and push the methodological rigor of their respective literatures. In addition, each chapter also demonstrates a distinct way that researchers can help improve outcomes for students. The chapter on charter competition demonstrates that school boards do not always make decisions in the best interest of student academic growth. This suggests a need to lobby state policymakers to ensure that school boards have appropriate checks on their power, such as the ability for charter schools to appeal to a county or state. The

second chapter highlights the need for ad hoc assistance in policy evaluation. California pioneered the use of suspension bans, but policymakers have not attempted to analyze their impact on academic growth – or even other outcome like dropout rates. Academics can either push to build evaluations into policymaking, or we can evaluate policies ourselves. The third chapter shows how researchers embedded within school districts are well positioned to engage in detailed policy analysis.[2] The impact of ability grouping may hinge on curricular design, teacher training, or other nuanced factors. If a researcher is working in a school district that is committed to making a policy work, they can engage in the meticulous work necessary to try and test many variations of implementation. These are not roles that academics are taught to play in most graduate schools, but such work may be a promising way to improve academic growth for students.

[2] Placing researchers in education agencies is an explicit goal of the Strategic Data Project fellowship (https://sdp.cepr.harvard.edu/), a program out of Harvard University's Center for Education Policy Research.

CONTINGENT ON TYPE: HOW THE FIRST CHARTER SCHOOL IMPACTS DISTRICT ACADEMIC GROWTH

Overview

How do charter schools impact the academic growth of the school districts in which they operate? We can estimate the causal effect of charter schools on their own students when they use a lottery to determine enrollment; we have no analogous way to evaluate the effect on nearby non-charter students. This causes recent literature reviews to reach a vague conclusion: charter schools have either a neutral or positive impact on the academic performance of the districts in which they operate (Buddin et al. 2015, Booker and Gill 2016). However, the most methodologically rigorous analysis in the literature – an instrumental variable analysis in a Southwestern school district – is also the one negative finding (Imberman 2011).

The extant literature suffers from two important issues. One is that its findings are driven by large school districts. Many studies focus exclusively on large districts: Chicago and Denver (Zimmer et al. 2009), Milwaukee (Nisar 2012), Ney York City (Cordes 2016; Winters 2012), Philadelphia and San Diego (Zimmer et al. 2009), and an anonymous Southwestern district (Imberman 2011). The problem is that most school districts are small; 93% of public school districts serve fewer than 10,000 students. Combined, these relatively small districts serve approximately half of the 50 million public school students in the United States. Even studies that cover entire states – or in the case of Davis (2013), the entire country – obtain most variation in charter competition from a small number of large urban districts. To take California as an example, in 2015

less than a third of all districts – 217 of 939 – contained any charter schools, and approximately half of those – 135 of 217 – only had one! Large districts with a growing number of charter schools are the exception, not the norm.

The second main issue for the extant literature is the reliance on relatively narrow observational data. This leaves findings vulnerable to issues such as omitted variable bias, selection bias, and sampling bias. Factors that encourage the opening of charter schools – such as parent engagement or community activism – may also impact the academic performance of nearby district schools. Such factors could bias an observed relationship between charter competition and district performance. This casts doubt on the internal validity of much of the literature. Imberman (2011) stands out by using the availability of facilities for charter schools as an instrument for charter competition. His finding of a negative relationship with district academic growth is at odds with the rest of the literature. However, his work also highlights sampling bias: it is unclear whether Imberman's finding is unique because of the particular district he studies or because the rest of the literature suffers from omitted variable bias.

This paper advances the literature in three ways. First, it shifts the focus away from large districts by analyzing the most common change in charter competition: a district-gradespan shifting from no charter schools to one. This occurs in a much more representative set of districts than other changes in charter competition, alleviating concerns of sampling bias. Second, this paper brings some certainty to the literature by dealing with omitted variable bias and selection bias with two innovative analyses detailed below. Finally, this paper highlights the important theoretical and practical

difference between two types of charter schools, referred to in California as locally funded and directly funded. We find that only directly funded charter schools boost the academic performance of their districts. While the boost is small, it undermines the common critique that charter schools harm the academic growth of other students by obtaining funding that used to go to their local district.

This paper proceeds in five parts. It begins by unpacking the literature on the impact of charter schools on their local school districts, leading to the hypotheses the literature would make concerning this paper's research question: how does the opening of a district's first charter school impact district academic growth? The second part explains why the charter sector in California is an ideal context for this study and describes two types of charter schools: locally funded and directly funded. The third part details the data and main analyses we use to answer that question. The fourth part explains threats to inference and efforts to overcome them. The final part reviews the findings and examines their implications.

Theory and Hypotheses

The literature views the impact of charter schools on their districts through the lens of economic competition (Clark 1961; Hirschman 1970). Charters are tuition-free and must accept all students; therefore, charter schools and their local districts typically are capable of serving overlapping sets of students. Theory hypothesizes that this competition for students compels school districts to change their behavior and provide better outcomes for students (Friedman, 1962; Hoxby, 2003; Hess, 2004).

While theory does not specify how districts would respond – e.g. increased learning time, teacher quality, or innovative practices – it does imply that the magnitude of change should be related to the degree of competition. This testable implication of theory dominates the literature. From this context, it makes sense that researchers have developed various ways to measure charter competition. One method is to focus on absolute presence: the distance to the closest charter school (e.g. Davis 2013) or the number of charter schools within a certain radius of a district school (e.g. Zimmer et al. 2009). A second approach focuses on relative presence: usually measured as the percent of students in a district who attend charter schools (e.g. Booker et al. 2008). The third approach can be called qualified presence because it combines one of the above approaches with some other factor seen as crucial to determining degree of competition. Cremata and Raymond (2014) argue that charter schools need to be relatively high performing in order to make districts feel competitive pressure. In their reviews of the literature, Booker and Gill (2015) and Epple et al. (2015) argue that charter sectors must be growing, not just large, in order to have positive impacts on districts. Researchers often conduct their analyses with more than one measure of charter competition (e.g. Booker et al. 2008).

One disagreement in the literature concerns where exactly to look for a reaction to competition from charter schools. At the most micro-level, it could be at the grade level. Jinnai (2013) finds evidence for this in North Carolina, with charters having a significant positive impact only on students in traditional schools' overlapping grades, not non-overlapping grades. Most of the literature assumes a school-wide impact, looking for an

impact at nearby schools serving the same general grade span (e.g. elementary, middle, or high schools). A more macro-level approach would look for a district-wide impact, looking for an impact at all schools district-wide serving that grade span.

Three reasons make it reasonable to suspect that charter competition has district-wide effects. One, district administrators play important roles that support academic growth, with staff often bearing responsibility for particular grade spans. Second, charter schools typically are not allowed to use geography as a criteria for enrolling students, meaning that charters often serve students from all over a district instead of just one attendance zone. Third, a potential mechanism of competitive pressure is that the opening of one charter school makes people across the district fear the opening of another charter. Importantly, district-wide impact is the most stringent test of any charter competition hypothesis because such a finding could vastly underestimate a larger impact that occurs only within certain schools or grades. For all these reasons, this paper uses this most stringent test: the impact on district-gradespans, i.e. all schools in a particular gradespan (elementary, middle or high) in the school district.

The literature has spent less time exploring differences between various charter school types.[3] Most analyses lump all charters into one category and assume they will have the same effect. While researchers always need to assume that some differences are not theoretically relevant, it is crucial to consider whether specific distinctions might be.

[3] Charters are extremely diverse: from fully online organizations to schools that operate in closed down traditional school buildings, from no-excuses programs that require uniforms to Montessori campuses that emphasize student-led learning, and from stand-alone schools that focus on dropout recovery to large networks of schools that focus on college going and completion.

The main exception is Buddin and Zimmer (2005), who explore the difference between startups – charters that are brand new organizations – and conversions – charters that used to be traditional district schools. The authors argued that conversions should exert less competitive pressure because they do not need to attract as many students from other district schools as startups do. Despite this theoretical expectation, they find no significant difference between startups and conversions.

This paper highlights a different charter type: locally funded and directly funded charter schools.[4] As explained in more detail in later sections, locally funded charters have budgetary and operational constraints imposed by their districts. At the extreme, districts even can revert locally funded charters back to traditional public schools; indeed, this occurred eight times between 2007 and 2013. If a district feels competitive pressure from a locally funded charter, then the district could respond by changing the charter instead of changing its own behavior. Since locally funded charters cannot maintain competitive pressure on districts, we hypothesize that this charter type does not cause districts to improve student outcomes:

- *<u>Novel Hypothesis 1</u>: Locally funded charters <u>do not</u> cause districts to increase academic growth*

[4] It is possible that some districts may strategically support the opening of locally funded charters in an effort to prevent the opening of a directly funded charter. To the extent that this occurs, districts with locally funded charters will be different from other districts in unobservable ways. However, California law states that districts are supposed to approve the opening of any charter school petition that meets particular conditions. As long as there is still community interest in a directly funded charter school, the existence of a locally funded charter does not prevent the approval of the directly funded one. Later sections will deal with the case where a district denial of a directly funded charter is overturned by the county or state board of education.

In contrast, directly funded charters operate independently from their districts. These charters have the authority to operate differently than the district in an effort to best serve students and families. The need to maintain enrollment gives directly funded charters an incentive to serve students and families well. While academic growth is not the only important outcome, it is certainly one important outcome. Districts feeling competitive pressure from a directly funded charter cannot simply make the charter change. Districts will have to change their own behavior to maintain their students. In some cases, charter schools may be able to offer services – such as longer school days or innovative curricula – that districts are unable to provide. However, all districts and charter schools can compete directly in terms of scores on statewide assessments. Therefore, we hypothesize that directly funded charters will tend to cause districts to improve student academic growth:

- *Novel Hypothesis 2: Directly funded charters do cause districts to increase academic growth*

These hypotheses are a novel contribution to the literature based on a previously overlooked difference in the extent to which charter schools are independent of their district.

In addition, the literature generates a set of competing hypothesis about what to expect about the impact of the first charter school on its district. In this study the absolute presence of charter schools is always small because there is only one charter within the district.

- *Absolute Hypothesis: The first charter will have no impact on district academic growth*

The relative presence of charter schools varies depending on the size of the district: a very small district may see a large percentage of its students enroll in the first charter school, while a large district may lose less than 1% of its students.

- *Relative Hypothesis: The first charter will impact district academic growth proportional to the percent of schools that are charters*

There are two distinct theories that fall in the category of qualified presence. In order to generate competitive pressure, either charter schools need to be relatively high performing (Cremata and Raymond 2014), or the charter sector must be large and growing (Booker and Gill 2015, Epple et al. 2015).

- *Qualified Hypothesis A: The first charter will impact district academic growth proportional to the relative academic performance of the charter school*

- *Qualified Hypothesis B: The first charter – followed by no others for several years – will have no impact on district academic growth*

The next section explains why California, the U.S. state with the largest charter sector in the country, is a good setting to test these hypotheses.

Background

The charter sector in California has three features that make it ideal for this analysis. First, the size and nature of charter sector growth provide a large enough number of observations to address our research question: how does the first charter impact its district? The number of California charter schools increased rapidly in a short

timeframe over a geographically dispersed area. The number of charter schools increased from 181 in the 1999-2000 school year to 1,059 in the 2012-13 school year.[5] Those years bookmark the previous assessment era, allowing us to calculate consistent measures of academic growth. Additionally, charters opened all over the state in a wide array of districts; over 100 districts experienced the opening of their first and only charter school between 2003 and 2013.

The second feature about California is that charter schools can be one of two types: locally funded or directly funded. While other states have only one of these charter types, both are significant in California; locally funded charter schools have consistently been approximately a third of the charter sector. This paper advances the literature by formulating and testing the hypothesis that the impact of a charter school depends on its funding type. Locally funded charters are informally referred to as "affiliated charters" by district staff and as "dependent charters" by charter advocates. Districts possess significant levers of control over these charter schools. Districts typically control their budgets, appoint a majority of their boards, and require staff to abide by the districts' collective bargaining agreements. In contrast, directly funded charters are informally referred to as "independent charters" because of their autonomy

[5] In fact, 181 charters in 1999-2000 is a high estimate. The state only began including which schools were charters in 2006-07 data files, so we obtain estimates for previous years using the open date for those 2006-07 charters. However, 74 of those charters have open dates before 1993, the year that California first issued charters. That means that these 74 must have opened as traditional public schools on their open date and converted to charter schools by 2006-07. We know 29 charters began in 1993, and since only 11 schools have open dates of 1993 we can assume that 18 of the 74 schools with early open dates became charters in 1993. That means the remaining 56 schools with early open dates converted to charters at some point between 1994 and 2006. Some of those 56 charters probably opened after 1999-2000, but we assume they all opened before then.
On the other hand, it is possible that a few charters that were open in 1999-2000 closed before 2006-07, preventing us from including them in our estimate.

from their districts. These charters control their own budget, structure their own board, and are not bound by the collective bargaining agreement of their district.

These two types of charter schools represent competing visions of the role of charter schools. Since the first California charter schools opened in 1993, people have debated whether charters should be internal or external forces of change. Those who prefer internal innovation see locally funded charters as a valuable tool of district-directed flexibility and innovation. Those who prefer external innovation see directly funded charters as helpful disruptors and incubators of experimentation.

Lastly, California has a relatively flexible legal environment concerning the authorization and oversight of charter schools. Since we are analyzing new charter schools, it would be important to note if the law caused our sample to be biased in terms of charter mission, composition, or quality. California charter law states that all school districts must allow charter schools to open as long as the charter petition fulfills generic legal requirements. While districts can interpret the law differently, any charter school that feels unfairly denied can appeal to the county and, if that also fails, even to the state. There is a cap on the number of charter schools that can exist statewide, but it is much higher than the current number of charters and increases by 100 each year – faster than the current growth rate. The National Alliance for Public Charter Schools, an advocacy organization, has one main critique of California's law: insufficient accountability provisions for charter schools. Therefore, California charter schools can remain open even if they provide very low quality educational outcomes for their students. By making it relatively easy to open charter schools and relatively hard to enforce a level of quality,

the law allows for a wide range of charters to open in California. This means that the charters in California do not face significant pressure to follow a particular model, serve particular types of students, or maintain a particular level of quality. If such filtering mechanisms existed, they would be important to consider because they would probably influence where charters tend to open.

Data and Main Analyses

We look at 1,404 district-gradespans in California, 72 of which had their first and only charter school open between 2003 and 2010.[6] These dates allow us to measure three years of growth data before and after the charter school opened. The independent variable is the opening of that first and only charter. The dependent variable is change in academic growth, reported by California's Department of Education as change in the state's Academic Performance Index.[7] We average growth across schools in a district-gradespan, excluding the one charter and any alternative schools.[8] Growth over three years provides a better estimation of the trend than relatively volatile single-year growth measures. We therefore calculate growth based on the three years before the charter

[6] This analysis excludes 1 case where there was no district school in the gradespan in which the charter opened (because there is no change in academic growth to analyze), 4 cases where the charter was a virtual school (because these charters attract only very particularly types of students), 11 cases where a locally funded charter opened in a district-gradespan with only one public school (because these are outlier cases), and 30 cases where that one charter school was an alternative school (because these charters serve special student populations).

[7] The Academic Performance Index (API) is based on standardized test scores, which themselves are limited measures of academic learning. Each year the state changed the exact calculation used to determine schools' scores: the list of included assessments could change, assessments could be given different weights, etc. However, the state produced growth scores each year that used the same API calculation in two consecutive school years. In order to determine three years of growth, we add up the growth scores from three years. The API (i.e. achievement) scores are not comparable across that many years.

[8] The state designates schools that serve a special population as "Alternative" and/or "ASAM" (Alternative School Accountability Model). For simplicity, this paper refers to schools with either of those designations as "alternative" and excludes them from analysis.

school opened and compare that to the three years after it opened. If anything, this may underestimate the impact of a charter school opening; charters may generate a one-year increase in growth that then disappears by the second year. To ensure that statewide growth trends do not bias the results, growth in each year is normalized to the state average for all students attending non-alternative schools. We then convert that growth to standard deviation units using the mean and standard error for the entire sample of non-alternative schools in the state.[9]

The main analysis utilizes a difference-in-differences design. (The following section details analyses that overcome omitted variable bias.) The first difference is the average change in academic growth for non-alternative district schools during the three years before and three years after their one charter school opened.[10] The second difference is between different types of district-gradespans. District-gradespans that have no charter open during this period experienced virtually no change in growth (1%).[11] District-gradespans where one charter opened experienced approximately the same change (3%). However, our two novel hypotheses would expect the change in academic growth to depend on charter type. This is exactly what we see in the last two columns of Figure 1. Locally funded charters do not lead to increase in academic growth; if

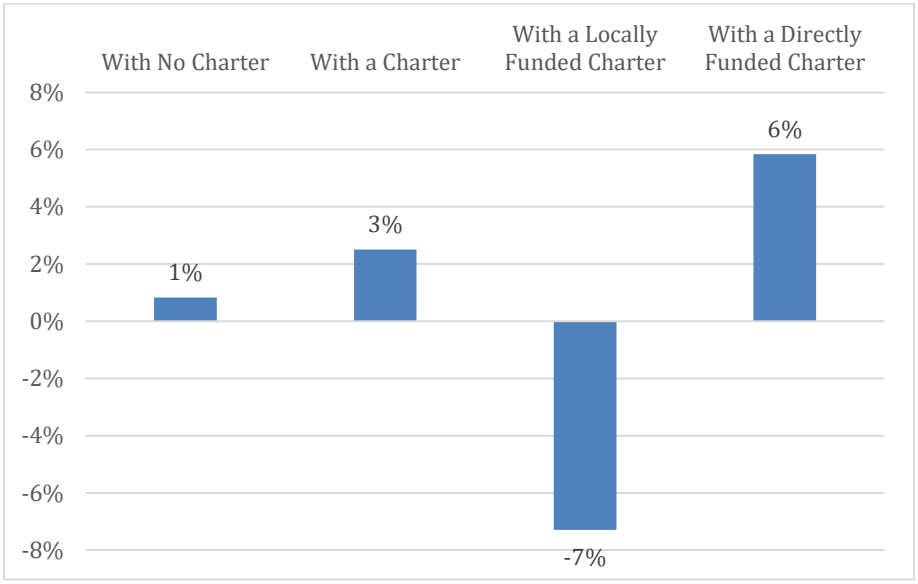
[9] This measure has an uncertain relationship with value added measures of academic growth, the gold standard in academic research. However, this measure has the benefit of being widely reported and understood among the public at the time. Therefore, this measure is an excellent measure of perceived improvement in academic quality. Given that the state did not report another type of academic growth during this period, this change in API measures only way in which all parents, school leaders, and district staff would estimate changes in school quality.

[10] It is worth noting that to the extent these charter schools are guilty of skimming high-performing students from their districts, this will depress the growth we measure in district schools.

[11] When districts have no charter school open, it is unclear what year we should use to calculate change in growth. We average the results from all eight possible calculations, with the "year" of a charter not opening ranging from 2003 to 2010.

anything, they may lead to a decrease (-7%). In contrast, directly funded charters lead to an increase in growth (6%).

Figure 1*:* Change in Growth for District-Gradespans



Note: Outcome is the three years of growth after a charter opened minus the three years of growth before a charter opened, averaged for all schools that are not alternative or charter. Excludes 1 case where there was no district school in that gradespan, 4 cases where the one charter that opened was a fully virtual charter school, 30 cases where that charter was an alternative school, and 11 cases where a locally funded charter opened in a district-gradespan with only one traditional public school. When the last 11 cases are included, the change in growth for locally funded charters drops to -29% of a standard deviation.

The small number of observations create a large amount of uncertainty around these estimates. A t-test comparing district-gradespans with directly funded charters to those with no charter produce a p-value of 0.47. T-tests comparing locally and directly funded charters have a p-value of 0.50 when we exclude the 11 district-gradespans with only one traditional school; however, the p-value drops to 0.05 when we include those 11 observations.

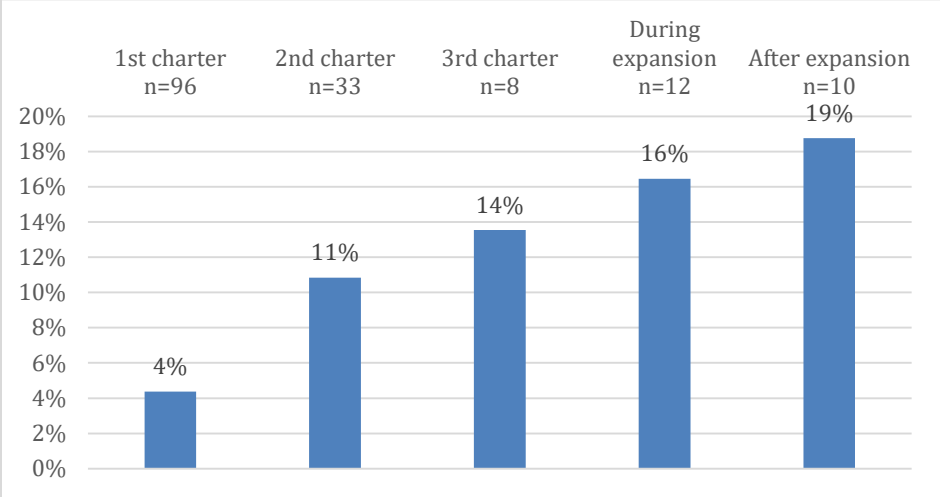
We then analyze the data in a second way. Ignoring locally funded charters, we find that 96 district-gradespans had the first directly funded charter open between 2003 and 2010. This larger group experienced a 4% increase in academic growth after the charter opened – almost exactly the same (6%) as the analysis above. We then find 33 district-gradespans that had their second directly funded charter school open between 2003 and 2010.[12] On average, we see an 11% standard deviation increase in growth after the second charter opens. There are only 8 district-gradespans where a third charter opened in that timeframe; this leads to an average growth of 14%.[13] Finally, there were 12 district-gradespans that experienced a period of charter expansion between 2003 and 2010, with their fourth or higher charter opening in that period.[14] These district-gradespans saw an average increase of 16%. The consistently positive results are suggestive of a general trend; that independent charters increase district academic growth. Since the number of observations for individual columns is very low, we hesitate to make claims about the differences between them. However, the evidence does not support the idea of a decreasing marginal return to each additional charter school; if anything, the evidence suggests the opposite.

[12] Since we measure growth using three years of data before and after a charter school opens, we exclude:
- 8 observations with were fewer than two years between the opening of the first and second charter
- 1 observation where the third charter opened in 2011, just one year after the second charter

[13] We exclude 3 observations with fewer than two years between the opening of the second and third charter.

[14] We exclude 5 observations where the number of charters opening during 2000-2003 prevent us from measuring academic growth before charter expansion.

Figure 2: Change in Growth for District-Gradespans with Directly Funded Charters



Note: Columns are determined based only on directly funded charters; we ignore the presence (or not) of locally funded charters. The outcome is the three years of growth after a charter opened minus the three years of growth before a charter opened, averaged for all schools that are not alternative or charter. We exclude observations with fewer than two years between charter openings.

The larger number of observations provides greater certainty about some estimates. The small numbers of observations in individual columns of Figure 2 means that the differences between them are not statistically significant. A t-test comparing district-gradespans with one directly funded charter to none finds a p-value of 0.26. Pooling all the district-gradespans in Figure 2 and comparing them to those with no directly funded charters produces a p-value of 0.15. While not at the standard level of statistical significance (usually $p<0.10$ or 0.05), the consistency of our findings coupled with the stringency of our tests suggest a non-random relationship.

Before moving to threats to inference, we use the sample of district gradespans with one directly funded charters school to test four hypotheses from the literature. We only find modest support for the Relative Hypothesis: *the first charter will impact district*

academic growth proportional to the percent of schools that are charters. To test this, we measure the correlation between change in district academic growth and charter market share – the share of non-alternative schools in a district that are charter. This correlation is 0.16 for the whole sample and 0.19 for directly funded charters. While this correlation is very low, its direction is consistent with the Relative Hypothesis. We find even weaker support for Qualified Hypothesis A: *charter impact is proportional to their relatively high academic performance*. We test this by measuring the correlation between change in district academic growth and the charter school's relative academic achievement. We measure relative achievement as the difference between the score of the charter school and the average achievement score of all other non-alternative schools in the district-gradespan during the two years after the charter school opened.[15] This correlation is 0.03 for the whole sample and 0.12 for directly funded charters. Again, the direction is consistent with the hypothesis.

As for the other hypotheses, evidence from directly funded charters directly counters the predictions that there would be no effect on district academic growth. This does not support the Absolute Hypothesis (i.e. that the absolute number of nearby charters drives competition) or Qualified Hypothesis B (i.e. that the charter sector needs to be large and growing to have a significant impact). While not statistically significant, Figure 2 shows that there might be larger increases in academic growth in the dozen

[15] Theory requires that when measuring relative academic achievement we look at two instead of three years after the charter school opens. The hypothesized mechanism is that people in the district see the relatively high performance of the charter school, which in turn motivates changes that lead to more growth in the district. School scores are not reported until the fall of the next school year, so there would be a one-year lag between observing relatively high performance and the impact of any subsequent reaction. In contrast, all other factors – enrollment, size, and charter market share – are observed at the beginning of each school year.

district-gradespans whose charter sectors experience large expansions. However, the very small number of observations prevents us from reaching strong conclusions for or against Qualified Hypothesis B.

Threats to Inference

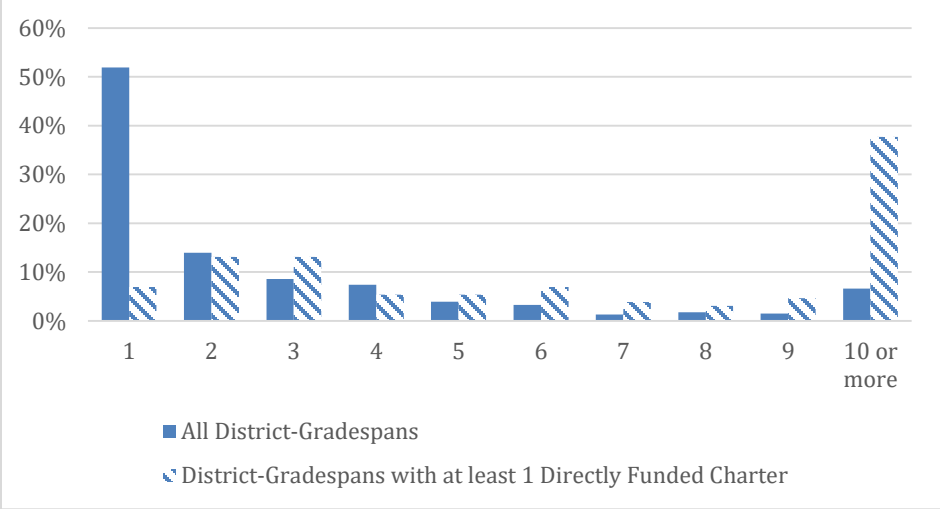
The main finding above – that the opening of directly funded charter schools increases district academic growth – faces several threats to inference. Our data was not produced in an experimental fashion. There is omitted variable bias: charter openings were not random, which means that other factors besides the opening of a directly funded charter school may explain the subsequent increase in academic growth. Additionally, there is selection bias: charters were not randomly assigned to districts, which means that the districts where charters opened may be categorically different than other districts. Even if charters did increase academic growth in the districts in our sample, charters may not have a similar effect in other districts. There also is the possibility that some districts strategically open locally funded charter schools in an effort to prevent the opening of a directly funded charter school. This means that the funding type of a district's first charter school is not randomly assigned. Districts with a locally funded charter may be different from other districts in a variety of unobservable ways – more strategic, better able to understand and respond to community needs, etc.[16]

We make a variety of efforts to alleviate those concerns. The difference-in-differences nature of the analysis means that we control for all factors that are constant

[16] It is important to note that California law states that districts should approve the opening of any charter school petition that meets particular conditions. As long as there is still community interest in a directly funded charter school, the existence of a locally funded charter does not prevent the approval of the directly funded one.

within district-gradespans. This eliminates a wide swath of omitted variable concerns. As for selection bias, charter schools opened across a wide range of districts in California, a large and diverse state with quite permissible charter school laws. One potential concern is district size. The literature has tended to focus on the experience of large, urban school districts; to what extent does the present analysis overcome this bias? Figure 3 shows that the district-gradespans with at least one charter have a size distribution somewhat similar to other district-gradespans. There are important differences at the extremes of the distributions; those without charter schools are much more likely to have only one school, while those with at least one charter are much more likely to have 10 or more schools. However, over half of district-gradespans with at least one charter had between 2 and 10 schools, and this portion of the distribution looks very similar to other district-gradespans. This suggests that at least in terms of size, the findings are likely to apply to all district-gradespans, with the possible exception of those containing just a single school.

Figure 3*:* Distribution of District-Gradespans by Number of Schools

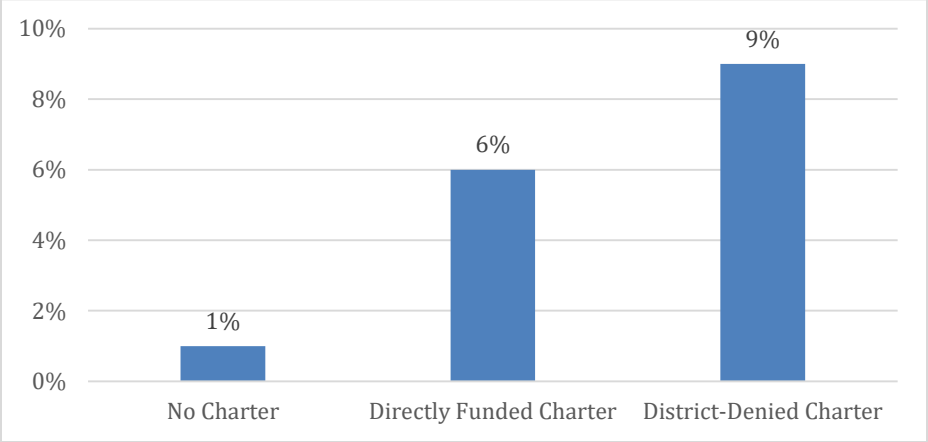


Omitted variable bias is likely to come from two sources, one being the district's willingness to authorize (i.e. open) the charter. For example, districts willing to authorize a charter school may also be increasingly open to innovation. This increasing openness to innovation (or some other district factor) may both improve academic growth and cause these charters to open. In order to prove that such an omitted variable is not driving our findings, we would need to untangle district willingness to authorize from the opening of a directly funded charter school.

The three-step process that determines charter school authorization in California allows us to do just that.[17] First, anyone seeking to open a charter school must submit a detailed petition to the local school board. If the district denies the charter petition, the second step is to appeal to the county board of education. If the county also denies the petition, then the third and final step is to appeal to the state board of education. There are 15 cases in our sample where districts actually denied the opening of a directly funded charter school, but the schools won an appeal to the county or state. If district willingness to authorize is related to the true cause of academic growth, then those 15 cases should have growth similar to district-gradespans with no charter school. Instead those 15 cases have growth comparable to directly funded charters approved by the district; in fact, Figure 4 shows that their growth is even higher. While there are not enough observations for statistical significance, it alleviates the concern that the main finding is driven by a factor related to district willingness to open a charter.

[17] For context, other U.S. states have very different authorization processes. At one extreme, states such as Massachusetts mandate that all charter schools be approved by one central entity. At the other extreme, states such as Ohio allow for a wide variety of actors – districts, universities, nonprofit organizations, etc. – to authorize charter schools anywhere in the state.

Figure 4*:* Change in Growth for District Gradespans, Testing District Willingness



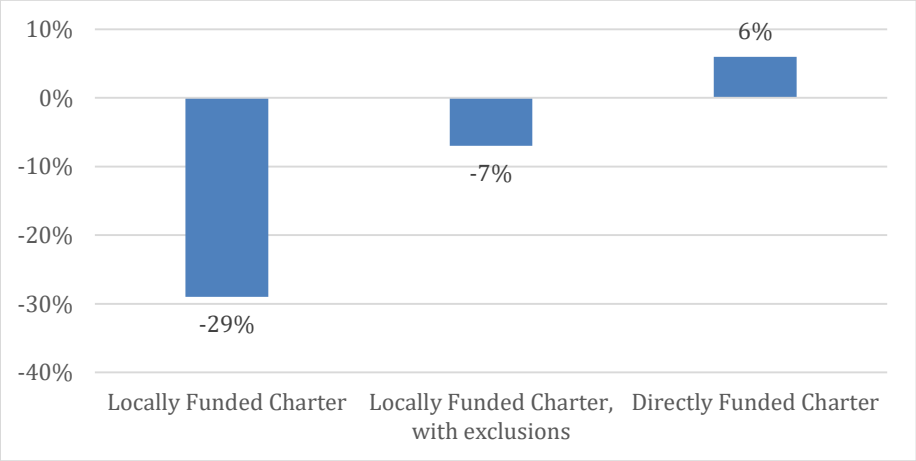
Note: Outcome is the three years of growth after a charter opened minus the three years of growth before a charter opened, averaged for all schools that are not alternative or charter.

The second likely source of omitted variable bias is the community's ability to generate a charter petition. An example of the latter would be that communities able to create charter petitions may have a growing number of educationally engaged parents, while communities that don't create such petitions do not. This growing number of engaged parents (or some other community factor) may improve academic growth and cause these charters to open. In order to prove that our main finding is not caused by such an omitted variable, we would need to untangle community ability to generate a charter petition from the opening of a directly funded charter school.

The best way to do this is to compare the impact of locally and directly funded charter schools. For both types of charters, communities had to create petitions and go through the same process to get a charter approved. Community ability is held constant. Nevertheless, only districts that opened directly-funded charters experienced higher academic growth. As stated above, the difference between directly and locally funded

charters is statically significant (p-value is 0.05) when we include all observations. There are not many cases of locally funded charters opening during the 2003 to 2010 window. Nevertheless, the available evidence – displayed in Figure 5 – suggests that they had, at best, a neutral effect on district academic growth. As long as locally funded charters did not have a positive impact, the implication is that factors associated with community ability to generate a charter petition cannot be the underlying cause of academic growth.

Figure 5*:* Change in Growth for District-Gradespans, Testing Community Ability



Note: Outcome is the three years of growth after a charter opened minus the three years of growth before a charter opened, averaged for all schools that are not alternative or charter. "With exclusions" means we excluded 11 observations where the district-gradespan contained only one traditional public school.

Taken together, these analyses suggest that the positive result for directly funded charters is not a result of bias. California is a large state with relatively lenient charter laws, allowing charters of wide ranging quality to open in a variety of contexts. The district-gradespans that experienced the opening of their first directly funded charter fall along the spectrum in terms of size. This suggests that these findings apply regardless of district size – an important addition to the existing literature. The impact of directly

funded charters occurs even when districts deny the charter's petition, revealing that factors associated with district willingness do not explain the results. Lastly, a similar impact does not appear to occur in locally funded charters, suggesting that factors associated with community ability to generate a charter petition do not explain the results. These additional analyses alleviate the primary threats to inference and bolster the case that these findings reflect a causal impact.

Conclusions and Implications

This paper finds that the opening of the first charter school in a district-gradespan causes an increase in academic growth – but only if the charter is directly funded (i.e. independent of district control). In contrast, the impact is not positive when the charter that opens is locally funded (i.e. affiliated with the district). The impact of directly funded charters is relatively small: approximately 5% of a standard deviation, or half of the difference between teachers in the top and bottom quartiles of performance. However, the impact is widespread: over three years and including all non-alternative schools in the district-gradespan where the charter opened. Using such a broad measure of growth may underestimate a larger and narrower impact on growth, such as over a single year in a smaller number of schools. However, it is possible to imagine mechanisms involving district staff that would cause such a widespread impact. Future research could provide insight into the exact mechanisms that lead from the opening of the first charter school to increased district academic growth.

This paper adds to the literature about the impact charter schools have on district schools in four ways. First, it highlights the distinction between locally and directly

funded charter schools. Charter schools only impact district academic growth when they are directly funded, which means they are independent in terms of collective bargaining, board governance, and budgetary decisions. To the extent that other states provide charter schools with varying degrees of independence, we are likely to see varying degrees of impact on district schools. Second, this paper shifts the focus towards smaller school districts. The extant literature is dominated by analyses of large urban districts, but approximately half of all students are in the 93% of districts serving fewer than 10,000 students. The districts in this analysis are a much more representative sample of all districts, which helps to alleviate concerns of selection bias.

Third, this paper advances the literature by alleviating some of the primary concerns about omitted variable bias. We see the same impact even when districts deny the charter school, implying that factors related to district willingness to open a charter school cannot explain the results. On the other hand, we do not see the impact when the charter school that opens is locally funded, implying that factors related to community willingness to open a charter school cannot explain the results. The extant literature only has one analysis with more internal validity: an instrumental variable analysis of one southwestern school district (Imberman 2011). However, the finding in this paper has greater external validity because it is based upon a much larger and more representative set of districts.

Fourth and lastly, this finding suggests that the literature should shift beyond an economic theory based on competition for students. Such a theory can only partly explain the patterns in the data. There is modest support for the hypotheses that charter

impact on their districts is proportional to their relative market share and to their relative academic performance. Additionally, the differential impact of locally and directly funded charters – this paper's novel theoretical contribution – supports the economic theory of competition. However, the correlations in the data are low enough to suggest that the economic theory is not the entire explanation. While beyond the scope of this paper, it could be that the mechanisms at play are more psychological than economic. It is not difficult to imagine that the opening of the first directly funded charter school is a watershed moment for most small- to medium-sized districts. This may prompt local leaders to acknowledge issues that were previously ignored and feel an urgency to take action. No such psychological effect occurs with locally funded charters because these schools do not really pose a threat to districts. Future researchers could explore this theory further.

The findings have interesting practical implications, as well. A powerful argument against charter schools has been that since they drain financial resources from districts, they harm district students academically. Charters typically do reduce district budgets, and research indeed links financial resources to academic growth (e.g. Jackson et al. 2014). The finding in this paper implies that the academic benefits from competition outweigh the harm caused by financial reductions.

Additionally, state leaders can encourage academic growth through charter school policy. They should ensure that charter schools have the independence necessary to increase district academic growth. They could make it easier for charters to open in districts that currently have none. While even the second or third charters may have

similar impact, the limited number of observations makes the evidence base strongest for the first charter. At the same time, state policymakers could prohibit or limit locally funded charters on the ground that they provide no external growth benefit and may even harm district growth.

The two charter types in California – directly funded and locally funded – reflect two distinct visions for the way charter schools were supposed to improve public education. Directly funded charters would innovate from the outside, while locally funded charters would innovate from the inside. Unfortunately, the evidence suggests that districts experience either static or declining growth when they open a locally funded charter. This may have the effect – intended or not – of moving the forces of innovation into a silo at one school. Future case studies of some of these districts could illuminate the exact reasons locally funded charters fail to generate academic gains in their districts.

The evidence presented here suggests that the opening of more directly funded charters would slightly improve academic growth in the thousands of school districts with no charters at all. Making this a reality, however, will require looking beyond the large urban districts that so frequently obtain our attention. Time will tell if charter schools shift from being primarily a solution to the problems of large urban districts to being a much broader source of innovation and improvement within our system of public education.

BANNING PROGRESS: SUSPENSION BANS AND SCHOOLWIDE ACADEMIC GROWTH

Overview

Of the approximately 50 million students in U.S. public elementary and secondary schools, close to seven percent – 3.5 million – are suspended each year. Descriptive statistics suggest that at least some suspensions are racially biased and unnecessary: African-American, Latino, and American Indian students are at least twice as likely as other students to be suspended (Wallace et al. 2008), and suspension rates have doubled over the past several decades (Losen 2011).[18] Simultaneously, many schools and districts have successfully managed student behavior while issuing few to no out-of-school suspensions (e.g. Christle et al. 2005; Luiselli et al. 2005; Skiba & Sprague 2008).

A variety of actors want to reduce suspensions. Notably, the U.S. Department of Education promotes alternatives to suspension: the Behavior Education Program (Crone et al. 2010), function-based interventions (e.g. Liaupsin et al. 2006), and school-wide positive behavior support (Luiselli et al. 2005; Putnam et al. 2006; Skiba and Sprague 2008). The National Association of State Boards of Education recommends that states include suspension rates in their accountability systems (Charis and Losen 2017).

---

[18] Factors beyond school discipline policy are likely to impact both of these trends:

In 1975 Congress passed Public Law 94-142, which is now known as the Individuals with Disabilities Education Act (IDEA). This law has helped bring hundreds of thousands of students with disabilities from state institutions into public schools, as well as brought millions of students from segregated instructional environments into integrated classrooms. In addition to its many benefits, Public Law 94-142 also could explain some of rise in suspension rates. (https://www2.ed.gov/policy/speced/leg/idea/history.html)

- African-American, American Indian, and Latino students are more likely to be low-income, English Learners, and students with disabilities than students in other subgroups. These students are also less likely to be taught by teachers who share their racial and cultural background. This may explain some of the disproportionality in suspension rates.

Likewise, advocacy organizations emphasize the public costs of suspensions (e.g. Rumberger and Losen 2017) and highlight schools that report high overall or subgroup suspension rates (Losen et al. 2015). Given this environment, it is unsurprising that many schools and districts are eager to reduce their suspension rates. The fastest way to do that is to ban suspensions for all but the most serious offenses.

Some suspension critics say that such suspension bans will produce schoolwide academic benefits. They point to research that suspensions reduce academic performance for both suspended students and, in some cases, even their non-suspended classmates (Arcia 2006; Perry and Morris 2014). However, the literature provides a second perspective: that suspensions can counter disruptive behavior that would otherwise reduce learning opportunities for all students (McFarland 2001). Given the competing hypotheses, how do suspension bans impact schoolwide academic performance?

The experience of Los Angeles Unified School District (LAUSD), the second largest school district in the country, provides an opportunity to answer that question. In May 2013, the LAUSD school board banned suspensions with the subjective rationale of "defiance," forcing a sudden and precipitous drop in its use. While not a true experiment, this policy shift can be called a natural experiment for determining the impact of a suspension ban. A difference-in-differences research design allows us to estimate how this sudden imposed change in suspension rates impacted academics. The first difference is temporal, comparing academic growth before and after the 2013 policy change. The second difference is geographic, comparing schools within LAUSD to those in the rest of California – i.e. schools that did not receive the "treatment" of a suspension ban. We test

three distinct hypotheses, and all of them support the conclusion that the LAUSD suspension ban harmed academic growth.

This paper proceeds in six sections. The first recounts the history of suspension bans. Next is an overview of the competing theories and hypotheses the behavior management literature provides concerning suspension bans. The third section details the data and measures used to measure suspensions and academic growth. The fourth sections walks through the analyses and results. The fifth section describes the remaining threats to inference and how additional analyses alleviate many of those concerns. The paper concludes with implications of the findings and next steps for practitioners and researchers.

Background

Suspension bans emerged as a reaction to the zero tolerance approach to school discipline. Zero tolerance policies administer strict punishment for relatively minor rule violations, often regardless of the circumstances and without formal due process (Cerrone 1999). The logic of this approach borrows from the broken window theory in the criminology literature (Wilson and Kelling 1982). Like broken window theory, zero tolerance makes two primary assumptions. One concerns the relationship between minor incidents (e.g. talking rudely to a teacher) and major incidents (e.g. fighting someone). Zero tolerance approaches assume that the presence of minor incidents makes major incidents more likely. The second assumption is that instituting policies of strict punishment for minor violations will make these incidents less frequent. This will occur through some combination of convincing students who get punished to not violate the

rule again and convincing other students to never violate the rule at all. If valid, then the zero tolerance approach should lead to fewer minor and major incidents.

Interestingly, zero tolerance policies became widespread in schools partly because of the passage of a federal gun law: the Gun-Free Schools Act of 1994. This law made federal education funding conditional on districts expelling students for a full calendar year if they brought a gun to school. This can be construed as a zero tolerance policy. Having a gun at school is relatively minor compared to brandishing or using a gun at school. This new law strictly punished the presence of guns as an effort to reduce the number of times any guns were brought onto or used on school campuses. Many states and districts used this law as a model to update their systems of discipline, adopting policies of automatic suspensions for a wide variety of relatively minor infractions (Skiba and Knesting 2001). By the early 2000s schools were suspending nearly twice as many students as in the 1970s (Wald and Losen 2003).

Suspension bans are an attempt to counter zero tolerance policies. Instead of automatic suspensions for minor infractions, suspension bans typically forbid schools from suspending students for all but the most serious infractions, such as those involving violence and drugs. LAUSD was the first large district to adopt such a ban for all grades. Surprisingly, the story of that first suspension ban starts with the federal tax reform law of 1986.

The 1986 tax law required nonprofit health organizations to start paying taxes. Blue Cross of California subsequently struggled to compete with for-profit competitors, so in 1996 it converted into an investor-owned for-profit called WellPoint Health

Networks, Inc. (Kane 1997). This required the creation of The California Endowment (TCE), a nonprofit charitable organization with an endowment of $2.3 billion. Newspapers across the state expressed concern that TCE's enormous endowment would allow it to dominate policy debates, and that the public would be unable to ensure that TCE acted in its interest.[19] Perhaps to allay these concerns, the website for TCE emphasizes that its 17-member board is extremely diverse and "is designed to reflect a cross-section of California's people and places."[20]

TCE produced a case study detailing its school discipline efforts in California (Martinez et al. 2013). While self-commissioned histories should always be read with caution, the events described suggest that TCE indeed played an unexpected and indispensable role in building the anti-suspension movement in California. In 2010, TCE launched a multi-million dollar effort called Building Healthy Communities to improve the health of 14 areas in California. While collecting input from stakeholders, TCE staff were surprised to hear that the abundant use of school suspensions was harming students' social and emotional health. TCE's statewide policy team looked into school suspensions and found that it was "an issue that framed correctly could have legs in Sacramento" because it was a widespread problem that could be remedied through relatively small changes to state education law (Martinez et al. 2013, p. 7). California *Education Code* Section 48900(k) allows schools to suspend students if they have "[d]isrupted school activities or otherwise willfully defied the valid authority of supervisors, teachers,

[19] For example, see Thelen J., "Charity or Advocacy for HMOs?" *The Recorder*, 30 September 1994; Editorial, "Blue Doublecross?" *San Jose Mercury News*, 11 November 1994; Editorial, "The Blue Cross Octopus," *Sacramento Bee*, 6 December1994.
[20] http://www.calendow.org/our-story/#leadership

administrators, school officials, or other school personnel engaged in the performance of their duties." Defiance is the most subjective basis on which schools can issue suspensions; other rationales concern various forms of theft, violence, and possession of illegal substances. Many schools have made frequent use of this flexibility to remove defiant students from school. When the state began collecting suspension data with these categories in 2011-12, it reported over 200,000 based on defiance, constituting 39% of all suspensions.[21]

Having decided to focus on school discipline, TCE used its resources to connect interest groups to one another. In May 2011, TCE convened a discussion of school discipline with community organizers from 8 of the 14 communities along with a statewide advocacy organization called Fight Crime: Invest in Kids. TCE then created the School Discipline Action Team, a coalition of three distinct groups:

1. Community organizers, such as Community Asset Development Re-defining Education (CADRE) and Labor Community Strategy Center (LCSC), who had worked for over a decade with families suffering from zero tolerance policies.
2. Legal advocates, such as Public Counsel and the American Civil Liberties Union, who knew the technical details necessary to know how to change school discipline law.

[21] For simplicity, "suspensions" without any qualifier refers to out-of-school suspensions.

3. State advocates, Children Now and Fight Crime: Invest in Kids, who were "new to the discipline issue" but also "were sophisticated, repeat players on the statewide scene" (Martinez et al. 2013, p. 8).

The School Discipline Action Team soon drafted 10 bills, but the community organizers lacked the expertise to keep up with the legal and state advocates; "even though CADRE and LCSC were formally involved when the legislative priorities were being hashed out in December 2011 and January 2012, they had limited ability to make substantive contributions" (Martinez et al. 2013, p. 11).

In the final phase, TCE executed a strategic communications plan designed to amplify and coordinate messages from a range of interest groups. TCE created a television commercial that aired in Sacramento as legislators considered the 10 bills. Additionally, TCE paid for a statewide poll about school discipline and strategically released those poll results simultaneously with recent research on suspensions. Seven of the ten bills passed both chambers, leaving it to Governor Jerry Brown to either sign or veto them. Weeks before Governor Brown made his decisions, TCE paid for all the facilities, rental, and travel costs for speakers to attend an event in Los Angeles to highlight the issue of harsh school discipline. Governor Brown ended up signing five of the seven bills into law. One veto was a bill that would have imposed a statewide ban on suspensions based on defiance.

That veto was only a temporary setback. In 2013, some of the community organizing groups that TCE had supported pushed for LAUSD to ban all suspensions based on defiance. The ban was backed by Superintendent John Deasy, an education

reformer, and approved by five out of seven members of the school board, which was frequently at odds with the superintendent. This surprising unity reflects the success of TCE's work at making the case for suspension bans in the court of California public opinion generally and among policymakers in particular.

Other districts and states soon started to adopt suspension bans. San Francisco Unified School District instituted its own ban in the summer of 2014, followed by Pasadena Unified School District in December 2014 and Oakland Unified School District in the summer of 2015 (Frey 2015). California state officials then started the trend of banning defiance suspensions only in younger grades; they banned these suspensions for grades K-3 starting in the 2015-16 school year. Since then the policy has spread across the country, with school boards issuing suspension bans for grades K-5 in Oregon[22] and grades K-2 in Texas (Reid-Cleveland 2017) and New York City (Berwick 2017). Additional bans are under consideration in cities such as Pittsburgh (Lindstrom 2017) and Philadelphia (Cline-Thomas and Chang 2017).[23]

Advocacy groups continue to pressure policymakers to reduce suspensions. The Civil Rights Project out of the University of California, Los Angeles is a leader of these efforts, using a multi-pronged strategy that includes appeals to potential legal action,

---

[22] For coverage in local news reports, see: http://gov.oregonlive.com/bill/2015/SB553/

[23] It is unclear whether policymakers are adopting suspension bans because they are facing similar incentives, or if policymakers are learning from each other through policy diffusion (Volden et al 2008). If it is policy diffusion, it would most likely be the imitation mechanism, with leaders copying the policy without considering its wider effects (Shipan and Volden 2008). The mechanisms of competition and coercion do not seem to apply, and the learning mechanism makes the assumption that subsequent adopters knew the impact of California's early suspension bans. The policy itself is very simple, which makes it easy to adopt (Makse and Volden 2011). Given the decentralized control of education policy, it is especially easy for state or local school boards to experiment with school discipline policies (Shipan and Volden 2012).

monetary savings, and public exposure. The organization's report about the discipline gap described the huge disparities in out-of-school suspensions as a "potentially unlawful denial of educational opportunity" (Losen et al. 2015). Another report estimated that the lifetime cost of suspensions for one cohort of California high school students was $2.7 billion (Rumberger and Losen 2017). The Civil Rights Project simultaneously published an online dataset of individual districts' suspension rates, cost of suspensions, and potential benefit from discipline reform.[24]

We conclude this section by noting that the federal government has contributed to school discipline reform. Under the Obama administration, agencies pushed hard against racial disproportionality in suspension rates. In 2014, the U.S. Department of Education and Department of Justice published a "Dear Colleague" letter containing a school discipline guidance package.[25] The package emphasized that school discipline must be done without being discriminatory, and the inclusion of the Department of Justice signaled that the federal government would pursue legal action against districts who failed to comply. Indeed, the Department of Education had previously reached voluntary legal settlements with a number of school districts with racially disproportionate suspensions, including LAUSD in 2011 and Oakland Unified in 2012.[26] While the Education and Justice Departments are not prioritizing these efforts under the Trump administration, anti-suspension information and resources are still available. The

---

[24] Available here: http://www.fixschooldiscipline.org/costsofdiscipline/
[25] Available here: https://www.ed.gov/news/press-releases/us-departments-education-and-justice-release-school-discipline-guidance-package-
[26] See https://www2.ed.gov/about/offices/list/ocr/docs/investigations/09105001-b2.pdf and https://www.ed.gov/news/press-releases/us-department-education-announces-voluntary-resolution-oakland-unified-school-di

Department of Education website still contains a section "Suspension 101" that includes the headlines "Suspensions don't work," "Suspensions have negative consequences," and "There are effective alternatives to suspension."[27]

Theory and Hypotheses

How does the literature expect a suspension ban to impact academic growth? Education researchers have not used the ideal research method of a randomized controlled trial to test behavior management policies. This is partly for ethical reasons: schools and districts cannot randomly suspend only some students for rule violations.[28] This has forced the literature to rely on sub-optimal methods to estimate the impact of a suspension ban.

One strand of the behavior management literature suggests suspension bans will increase academic growth through two mechanisms. First, students who would have been suspended without the policy should now have higher academic growth. Arcia (2006) makes this claim based on a matching analysis. Students who received suspensions are matched with students with similar observed characteristics – such as academics, demographics, and prior behavior issues – who did not receive suspensions. Suspended students had much worse academic outcomes in subsequent years. The problem is that this difference in outcomes may be caused at least partly by unobserved differences between matched students. For example, students who did not receive suspensions may have had better home environments or peer relationships. The second

[27] https://www2.ed.gov/policy/gen/guid/school-discipline/index.html
[28] However, it would be ethically sound for districts to randomly assign schools to one of two (or more) equally promising discipline reform programs.

mechanism argues that even students who would not have been suspended should experience higher growth because they suffer less from the distraction of a punitive environment created by issuing too many suspensions. This mechanism finds support in Perry and Morris's work in *American Sociological Review* (2014). They use fixed-effect regression models to find that fluctuations in suspension rates cause subsequent fluctuations in academic performance. This assumes that there are no omitted variables that might cause both changes.[29] However, omitted variables almost certainly are a problem. Factors such as sudden problems in students' home lives or peer relationships are likely to cause both an increase in suspensions and a decrease in academic growth.

A second strand of the literature proposes that a suspension ban will decrease academic growth. This perspective argues that suspensions are a tool teachers can use to remove defiant students, thereby providing more opportunities for the vast majority of students to learn. By observing classrooms repeatedly over an entire school year, McFarland (2001) sees that defiant behavior can harm both teachers and other students. Defiant behavior can derail teachers' plans, increase their stress, and – in the most extreme cases – even cause them to leave their positions. Teachers can take actions to prevent most defiant behavior from occurring, but this requires training and practice. If teachers have no tool other than suspensions to deal with defiance, then suddenly removing that tool would lead to classroom management problems. This would result in lost learning opportunities for all students. The main issue with this strand of the

[29] Prior research was cross-sectional, comparing suspension rates to academic achievement (Rausch & Skiba 2004). Even some recent papers on the subject sometimes use correlational evidence to support the claim that suspensions harm overall academic growth (Losen et al. 2015).

literature is limited external validity. The fact that one researcher observed negative outcomes caused by defiant students in a small set of schools does not mean that the same negative outcomes would occur everywhere.

In summary, theory provides opposing views about how a sudden decrease in suspensions would impact academic performance. The first strand views suspensions as completely harmful, so the more we reduce suspensions the more we will improve academic growth. The second strand views suspensions as a tool for teachers; the more we restrain their ability to use this tool (without providing a replacement), the more we will harm academic growth. Both strands of the literature have limitations based on the observational (i.e. non-experimental) nature of their evidence. Furthermore, the recent adoption of suspension bans means that no researchers to date have explicitly studied the academic impact of this school discipline policy.

The LAUSD suspension ban provides an opportunity to test three distinct hypotheses concerning the relationship between suspension reduction and academic growth:

H1: Change in academic growth should be higher (lower) in LAUSD than in the rest of California

H2a: The gap found in H1 should be even larger for schools that gave the banned suspensions before the LAUSD policy shift

H2b: There should be almost no gap for schools that did not give the banned suspensions before the LAUSD policy shift

H3: Among schools in LAUSD, change in academic growth should be highest (lowest) in schools that used to give the most of those suspensions before the LAUSD policy shift

H1 is the simplest hypothesis, comparing all schools that experienced the intervention (i.e. the 2013 suspension ban) to all schools that did not. H2a and H2b highlight the fact that we primarily expect the suspension ban to impact schools that gave defiance suspensions prior to the ban.[30] Finally, H3 exploits the fact that the suspension ban impacted schools within LAUSD differently based on how many defiance suspensions they gave prior to the ban. The next two sections describe the data and analyses we use to test these three hypotheses.

Data and Measurement

The data suggests that the LAUSD suspension ban had an enormous impact on suspensions.[31] Figures 6 and 7A show the number of suspensions based on defiance has dropped steadily state-wide in absolute and relative terms, while the absolute number given for other reasons has decreased slightly.[32] Graphs 6 and 7B show a different story for LAUSD: a precipitous drop in 2013-14 that leads to an extremely small number of

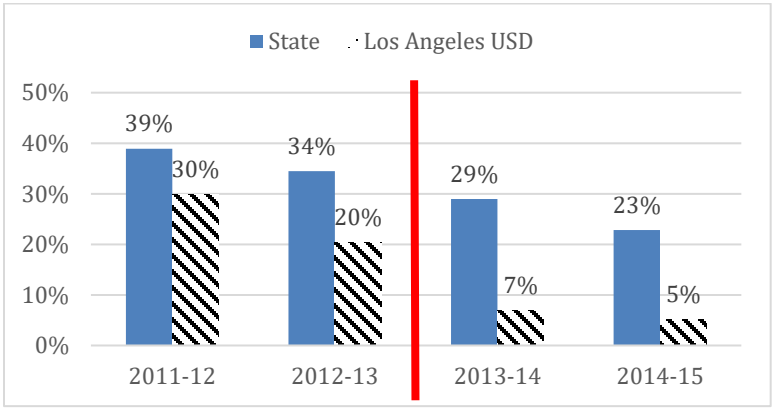
[30] We expect the suspension ban to have some impact on some schools that gave no defiance suspensions in 2013. New students, changes in current students' lives, and changes among school staff could all result in schools feeling a sudden need to give defiance suspensions. In our analysis there were 201 schools outside of LAUSD that gave no defiance suspensions in 2013; 81 of them (40%) gave at least one defiance suspension in 2015, and 11 (5%) gave at least a dozen. Likewise, some schools that gave defiance suspensions in 2013 may not feel a need to give any such suspensions after 2013, even in the absence of a ban. Of the 866 schools outside of LAUSD that gave defiance suspensions in 2013, 127 (15%) gave no such suspensions in 2015. This is why we expect the impact of the suspension ban to be driven mostly – but not entirely – by schools that gave defiance suspensions in 2013.

[31] Suspension data is available here: https://www.cde.ca.gov/ds/sd/sd/filessd.asp. One suspension "incident is defined as one or more students committing one or more offenses on the same date at the same time."

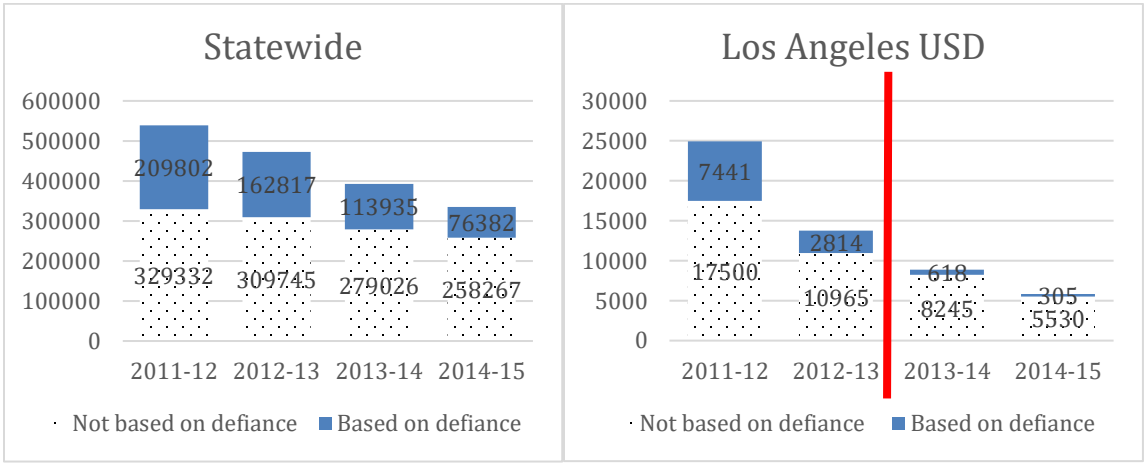
[32] Annually, the state has reported suspension data with the defiance category beginning with the 2011-12 school year.

suspensions based on defiance. This demonstrates the district-wide response to an explicit school board policy. In May 2013, the LAUSD school board voted to ban suspensions based on defiance during the upcoming school year (Watanabe 2013b). The overall compliance rate was high; from 2,814 suspension based on defiance in 2012-13 to just 305 in 2014-15. In general, schools did not compensate by suspending students for other reasons; the number of those suspensions decreased over the same period at a rate faster than the statewide average.

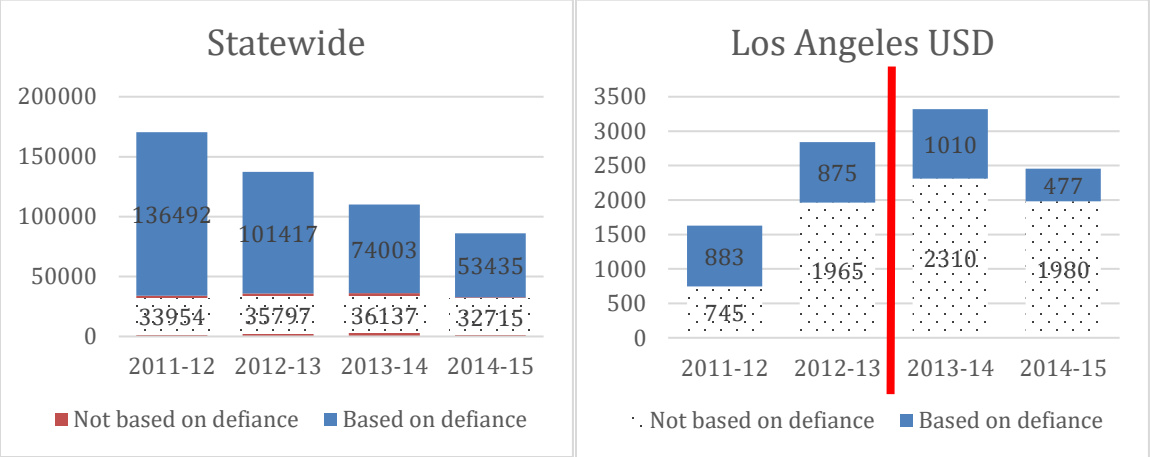
Figure 6: % of Out-of-School Suspensions Based on Defiance



Figures 7A and 7B: # of Out-of-School Suspensions



Figures 8A and 8B: # of In-School Suspensions



Some LAUSD schools appeared to compensate temporarily by issuing more in-school suspensions. Figure 8B reveals an initial uptick in 2013-14 in-school suspensions, both based on defiance and not. However, by 2014-15 the number of in-school suspensions based on defiance dropped to approximately half the level of 2012-13, and the number given for other reasons returned to the 2012-13 level. If we ignore 2013-14, the change in LAUSD from 2012-13 to 2014-15 mirrors that of the state.

Measurement of suspensions becomes more complicated when we shift to the school level. We use two types of independent variables depending on which hypothesis we are testing. For hypotheses one and two, we are comparing schools that experienced LAUSD's suspension ban to California schools that did not. The independent variable is therefore binary: did the California school experience the policy change (i.e. because it is located in LAUSD) or not. The third hypothesis compares LAUSD schools by the extent to which they were impacted by the suspension ban. This forces us to grapple with a variety of data and theoretical limitations described in detail below and in Appendix A.

Those constraints lead us to use a measure of the intent to treat: a categorical variable reflecting the number of suspensions based on defiance a school gave in 2013.

Table 1 shows basic information for the groups of schools used to test each hypothesis. Hypothesis one is the broadest, comparing all schools in LAUSD to all California schools outside of LAUSD. The only requirement for inclusion is that schools must have the relevant growth data: sixth grade data in 2011 and 2013, and eighth grade data in 2013 and 2015. Hypothesis two separates these schools based on whether they gave any suspensions based on defiance in 2013, before the LAUSD suspension ban took effect. Hypothesis three looks only at schools within LAUSD. Here we add one exclusion rule; we remove 19 schools that served high school students. We do this because we want our schoolwide suspension data to reflect middle school, the grades where we are measuring academic growth.

Table 1*:* Information about Schools in Analyses

	Schools	# of Schools	Avg. # 2013 Suspensions	Avg. # 2015 Suspensions	Low-Income	English Learner	Students w/ Disabilities
H1	LAUSD	113	6	1	81%	20%	12%
	Non-LAUSD	1,068	19	9	56%	20%	9%
H2a	LAUSD	79	8	1	83%	21%	12%
	Non-LAUSD	867	23	11	58%	20%	10%
H2b	LAUSD	34	0	0	79%	18%	11%
	Non-LAUSD	201	0	2	47%	19%	8%
H3	LAUSD: 0	29	0	0	77%	18%	10%
	LAUSD: 1-10	49	2	1	80%	19%	12%
	LAUSD: 11+	16	27	3	89%	20%	14%

Note: Demographic variables reflect all tested students in 2015

Available school-level suspension data has several important limitations.[33] While we would ideally like to have data by grade level to match academic growth, the state

[33] For ease of reading, "suspensions" from this point on refer to out-of-school suspensions based on defiance, unless explicitly qualified otherwise.

only reports data at the school level. Additionally, redaction prevents us from calculating exact suspension rates for a subset of schools. The state provides suspension data files containing data for each subgroup in each school, but in order to protect student confidentiality it does not report numbers between one and ten. Using the 2013 (i.e. pre-treatment) number of suspensions based on defiance produces the following categorization:

1. None (29 schools)
2. Between 1 and 10 (49 schools)
3. At least 11 (16 schools)

This can be viewed as a measure of the intent to treat. If schools in LAUSD had complied perfectly with the suspension ban, then this categorization would perfectly reflect the absolute change in suspensions. Officially, compliance was high enough to make this almost true. Of the 94 schools in our sample, 67 had zero suspensions based on defiance in 2015, another 20 had one redacted subgroup (i.e. on average 1.5 suspensions each), six schools had between two and twelve suspensions, and one school had thirty-five.

However, there are reasons to believe that the data after 2013 is not accurate. In the first year of implementation, parents claimed that their children were sent home without being officially suspended (Watanabe 2014). School staff became much more likely to call the police in order to deal with defiant students, forcing officers to remind school staff that "willful defiance is not a crime" (Watanabe 2015). There are also a variety of ways to remove students from class but keep them in school, and

administrators may not report all those instances as in-school suspensions. The street-level bureaucracy literature is founded upon the notion that the actual implementation of top-down policies is determined largely by the decisions of front line, or street-level, workers (Lipsky 1980). In this case, the street-level workers are school administrators and even teachers who decide how to deal with defiant students. School staff see themselves as accountable to their students, parents, and fellow staff as well as their school board (Hupe and Hill 2007). Staff therefore might report perfect compliance while finding ways to remove defiant students from school.

Importantly, reported compliance is a confounding factor because it impacts the change in suspension rate and could also be related to the change in academic growth. Twelve LAUSD schools reported giving more suspensions based on defiance after the suspension ban. These schools had relatively low growth rates in 2013 and grew even less in 2015 (see Appendix A for more details). Instead of the increase in suspensions causing the drop in academic growth, it seems more likely that these schools suffered from other issues that led to both the increase in suspensions and the drop in academic growth.

Data limitations place a variety of constraints on our ability to measure change in academic growth, the dependent variable. The main issue is that California switched to a new assessment regime immediately after LAUSD enacted its policy. Fortunately, the residual gain model of academic growth does not require that pre- and post-tests be on the same scale. The basic residual gain model is a bivariate regression where the post-test score is the dependent variable and the pre-test score is the independent variable.

The regression residuals estimate the extent to which students performed lower or higher than expected given their starting score. This is one form of value-added modeling, which is the best available metric of academic growth in the education literature (e.g. Kogan et al. 2016b).[34]

A related challenge is that the assessment regime prior to 2015 gave end-of-course assessments for some high school English classes as well as all Math classes starting with Algebra, which some students took in middle school. We can only calculate the residual gain model of academic growth when all students took the same assessment in the same grade. This prevents measuring growth for high schools at all, or for middle school Math. Elementary would be possible as well, but suspensions occur rarely among students below sixth grade. Therefore, all measures of academic growth are based on middle school English assessments.

Two other issues concern state reporting of academic data. First, California did not publish any statewide assessment results in English or Math for the 2013-14 school year. This means that the pre-test is sixth graders in spring 2013 and the post-test is eight graders two years later in spring 2015. [35] In order to calculate analogous growth prior to the policy intervention, we similarly span two years: sixth graders in spring 2011 to eighth graders in spring 2013. The second issue is that the state only makes data

[34] The value-add literature has primarily focused on its controversial use in teacher evaluation. Using value-add estimates for entire grades avoids some concerns, such as sorting difficult-to-teach students into particular teachers' classes (Rothstein 2009). However, even grade level value-add measures can experience significant variation across time (Goldhaber and Hansen 2008) and can be sensitive to the assessment used (Lockwood et al. 2007). On the positive side, value-add measures are highly correlated with principal evaluations of teachers (Kimball et al. 2004; Jacob and Lefgren 2008).

[35] Test results for 2010-11 to 2012-13 school years are here: https://star.cde.ca.gov/starresearchfiles.asp. Test results for the 2014-15 school year are here: https://caaspp.cde.ca.gov/sb2017/ResearchFileList.

available for entire grades at each school, not individual students. Ideally, pre-test sixth grade scores and eighth grade post-test scores would include the exact same sets of students. Reality is more complicated; some students leave and enter the sixth grade cohort because they change schools or are held back a grade. If the weighted average score of the students who leave the cohort matches the weighted average of the students who enter, then there is no bias. Bias occurs when students who enter and students who leave have different weighted average scores. For example, imagine a school where equal numbers of students leave and enter the sixth grade cohort. Students who leave are relatively high-achieving while students who enter are relatively low-achieving. In this case, our growth measure would underestimate the amount of academic improvement that really occurred. While California does not report what percent of students remain in a sixth grade cohort from year to year, it reported what percent of students are continuously enrolled at a school from early October to spring testing as recently as 2013. The median score was always near 95%, reflecting the fact that most schools have relatively stable student populations.

Analyses and Results

The primary analysis is a difference-in-differences research design. The first difference is temporal: outcomes before the suspension ban compared to outcomes after the suspension ban. In this case, the dependent variable is change in middle school academic growth. The second difference is the comparison between various groups of schools, depending on the particular hypothesis under consideration. While H1 and H2

compare LAUSD schools to schools in other parts of California, H3 compares groups of schools within LAUSD.

Table 1 shows that LAUSD schools differ from other California schools on a variety of observed characteristics. This means the suspension ban was not applied to schools in an "as if" random process, so we cannot consider the analyses of H1 or H2 to be natural experiments (Dunning 2012). H1 is particularly problematic because LAUSD enacted a variety of other policies during the period that could explain differences between the academic growth of LAUSD compared to the rest of the state. H2a and H2b alleviate this problem somewhat by looking at schools based on whether or not they gave defiance suspensions in 2013.

However, Table 1 also shows that LAUSD schools were very similar on observed characteristics regardless of the number of suspensions they gave in 2013. This allows us to argue that the suspension ban was applied to schools within LAUSD in an "as-if" random process, allowing us to consider the analysis for H3 a natural experiment. The case for a natural experiment is strengthened by the fact that schools within LAUSD were subject to any other policy changes that the board enacted between 2013 and 2015. Additionally, LAUSD schools share other unobserved characteristics with one another that they may not share with schools in other parts of California. For all these reasons, the comparison among schools in LAUSD should be considered a natural experiment, making this a very strong test of the relationship between suspension reduction and academic growth.

To test H1, we compare all LAUSD middle schools to all other middle schools statewide. Figure 9 shows that LAUSD schools experienced a 16% standard deviation decrease while other California middle schools remained essentially static. However, this comparison includes schools that were not impacted by the suspension ban because they did not utilize defiance suspensions. The middle columns support H2a by revealing a bigger gap between LAUSD (-22%) and the rest of the state (-1%). The right-hand columns show very slight changes in academic growth for schools – in and out of LAUSD – that did not give defiance suspensions. This supports H2b, providing confidence that the gap we see between LAUSD and the rest of the state is driven by this suspension ban as opposed to other possible factors. Bivariate regressions reveal that the difference between LAUSD and non-LAUSD schools is almost statistically significant for H1 (p=0.17) and H2a (p=0.12), supporting the strand of the literature arguing that a sudden drop in suspensions harms academic growth.

Figure 9*:* Change in Academic Growth by Policy Treatment

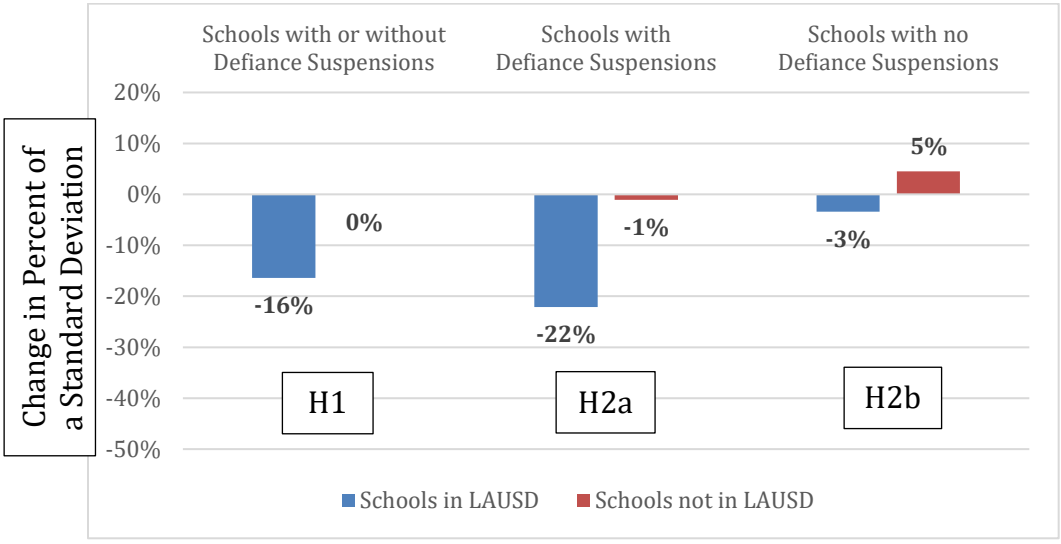
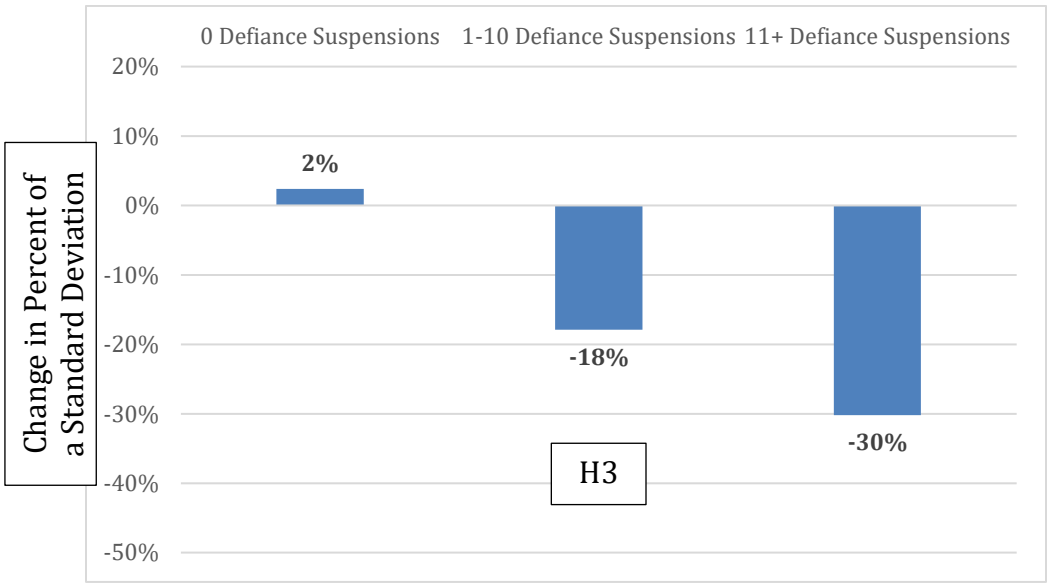


Figure 10: Within LAUSD, Change in Academic Growth by Pre-Policy Number of Suspensions



The comparisons in Figure 9, between LAUSD and the rest of California, are not a natural experiment. This is because there are a variety of observable differences between these schools, with key differences summarized in Table 1 above. In contrast, the comparisons in Figure 10 do reflect a natural experiment because the schools within LAUSD are nearly identical on a range of observable traits with one exception: the number of defiance suspensions they gave in 2013.

Looking within LAUSD reveals a linear relationship between the intent to treat and academic growth. Schools with no suspensions in 2013 had almost no change in growth. Schools with one to ten suspensions experienced an 18% standard deviation drop, while schools with at least eleven suspensions in 2013 experienced a 30% decrease in growth. This pattern also perfectly fits the expectations of the second strand of the literature: suddenly reducing suspensions harms academic growth. The small numbers of

schools in each category prevent any of the differences from being statistically significant. However, this comparison of schools within LAUSD is a natural experiment: the schools are very similar in observed and unobserved characteristics, with the exception of the extent to which they were impacted by the suspension ban (because they gave different numbers of defiance suspensions in 2013). Combined with hypothesis 2, these findings alleviate concerns of omitted variable bias, the concern that some other factor caused the observed drops in academic growth. In order for an omitted variable to explain the story, it would need to both be related to changes in academic growth and the number of suspensions based on defiance schools issued in 2013.

Threats to Inference

This paper falls short of the ideal of an experimental design leading to statistically significant results. However, additional analyses alleviate the main threats to inference. One major concern is measurement error in the outcome, change in academic growth. Each measurement of growth covers two years and reflects one cohort of students. Student mobility causes grade level cohorts on which growth is calculated change by unknown amounts over time at different schools. Academic growth from 2013 to 2015 spans two different assessment systems and standards, leading to concerns that we may be measuring variations in teacher preparation for this change more than changes in student learning.

Two analyses address this issue. One involves converting the academic performance of each school in the state into a percentile from 1 to 100 in both 2013 and 2015. This is the closest we can get to a consistent measure of achievement given the

change in assessment systems during this time. We then calculate the change in academic percentile from 2013 to 2015 for each school. Replicating graphs 4 and 5 with this academic measure reveals the same trends reported in this chapter. The other analysis concerns suspension bans that took place after California adopted new assessments in 2014-15 and is described below.

A second major concern is omitted variable bias. As explained above, the difference-in-differences methodology partly addresses this concern. By looking at the change over time, we hold all time-invariant factors constant. The top-down nature of the suspension ban provides additional assurances; the intended drop in suspensions is unrelated to sudden changes in students' lives that would also cause a drop in academic growth. Additionally, we run six regressions to see if school-level traits can explain the strongest finding, hypothesis 2A. We control for prior academic growth, racial demographics, non-racial demographics, and then all those factors combined. The variable for being in LAUSD always has a negative coefficient (i.e. LAUSD schools experienced decreased academic growth), and is similar in size across most models. The exceptions are models 3 and 5, which include non-racial demographics: the percent of students who are low-income, English Learners, and students with disabilities. Model 6 shows that when these non-racial demographics are excluded, the LAUSD dummy variable is negative and statistically significant.

The rows for H2A in Table 1 show that LAUSD has relatively high percentages of students in all three of these non-racial demographic categories. It is possible that these factors caused approximately half of LAUSD's relative decrease in academic

growth. However, it is also possible that the correlation between these factors and being in LAUSD is coincidental.

Table 2: OLS Regressions as Robustness Tests for Hypothesis 2A

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
In LAUSD	-0.209 [0.133]	-0.235** [0.107]	-0.116 [0.139]	-0.231* [0.138]	-0.078 [0.112]	-0.197* [0.109]
Additional Independent Variables	None	Prior Academic Growth (2011 to 2013)	Non-Racial Demographics (% with Disabilities, % Low-Income, % English Learner)	Racial Demographics (% African American, % Asian, % Latino, % Other)	All	All but Non-Racial Demo-graphics
Observations	946	946	946	946	946	946
Adjusted R^2	0.002	0.359	0.007	0.003	0.387	0.369

Notes: The outcome variable is change in growth from 2011-2013 to 2013-2015. Observations include all schools with change in growth data and at least one suspension based on defiance in 2013: 79 in LAUSD and 867 not in LAUSD. Demographic data reflects all test-takers at each school in spring of 2015.

The final concern is selection bias: LAUSD schools were not randomly selected to impose a suspension ban. We are assuming that LAUSD schools are comparable to all other California schools. However, LAUSD is different from the rest of the California in many ways, and perhaps some of those unique characteristics allowed the suspension ban to harm academic growth. If another school district implemented a suspension ban, would academic growth be similarly hurt? Additionally, the LAUSD analysis only includes middle schools. Would suspension bans also be harmful in elementary grades?

Four subsequent suspension bans in California suggest that the experience of Los Angeles in generalizable.[36] First, San Francisco Unified School District adopted the same suspension ban in 2014, one year after Los Angeles. We can conduct a similar

[36] A fifth suspension ban took place in Azusa Unified School District, but the paper cannot similarly measure the impact of that ban because it was implemented over three years.

difference-in-differences analysis for San Francisco, comparing 2011-2013 growth to 2013-15 growth. Students in San Francisco only experienced the ban during 2014-15, not 2013-14, so they effectively received half the "dosage" of the ban that students in Los Angeles received. Nevertheless, columns one and two in Table 3 show a significant decline in academic growth compared to the rest of the state, even after controlling for prior growth rate and demographics. While the coefficient is even larger than the observed impact in Los Angeles, the confidence interval is fairly wide because of the small number of schools.

Table 3: Four Subsequent Suspension Bans

	San Francisco Unified (n=20) vs. CA		**Pasadena Unified (n=30) vs. CA**		**Oakland Unified (n=134) vs. CA**		**CA K-3 Sites: Some (n=21) vs. No (n-37) Suspensions**	
Experienced Ban	-0.890*** [0.268]	-0.506*** [0.214]	-2.027 [2.127]	-1.372 [2.096]	-5.328*** [1.175]	-2.784** [1.196]	-4.347 [4.486]	-4.503 [4.938]
Additional Independent Variables?	No	Yes	No	Yes	No	Yes	No	Yes
Observations	1181	1181	7040	7037	7040	7037	58	58
Adjusted R^2	0.008	0.426	0.000	0.036	0.003	0.036	-0.001	-0.035

Notes: The outcome variable for San Francisco Unified is change in growth from 2011-2013 to 2013-2015. The outcome variable for Pasadena and Oakland is actual academic growth minus expected growth after controlling for prior average scale score. The outcome variable for CA K-3 schools is the change in third grade average ELA scale score. Additional Independent Variables include test taker demographics (% low-income, % students with disabilities, % English Learners, % African American, % Asian, % Latino, % Other) and prior achievement – or in the case of San Francisco, prior growth.

The next bans were in Pasadena in December 2014 and Oakland in 2015. These occurred after the state transitioned to new assessments, which eliminates concerns that the assessment switch somehow biases the results. However, the bans occurred during or immediately after the first year California administered its new assessment. A

difference-in-indifference analysis is not possible because we cannot measure growth before these bans. Therefore, the best we can do is compare their academic growth during 2015-16 to other schools statewide.

Pasadena schools experienced the suspension ban for roughly half of 2014-15 and then all of 2015-16. Growth is based on the difference between the current year's achievement (2015-16) and the prior year's achievement (2014-15). That means the Pasadena analyses in Table 3 are comparing a full year of suspension ban to a half year of the suspension ban. In contrast, Oakland schools only experienced the suspension ban during 2015-16. The Oakland analyses therefore compare a full year of suspension ban to a year with no suspension ban. We would expect the Pasadena "Experienced Ban" variable to be approximately half the size of the Oakland one. The results in Table 3 fit that expectation exactly. The fact that the Pasadena "Experienced Ban" variables are not statistically significant is unsurprising given that the effect is smaller (as we would expect) and there are only 30 Pasadena schools in the analysis.

Finally, the entire state of California banned suspensions for students in grades K-3 in 2015. To analyze the impact on academics, we exploit the fact that just over a third of K-3 schools (21 of 58) gave defiance suspensions in 2014-15, before the ban took effect. In the analysis of LAUSD, we see that the impact of the suspension ban was concentrated in schools that previous issued defiance suspensions. If the suspension ban harms academic growth even for third graders, then we would expect those 21 schools to grow less from 2014-15 to 2015-16 than the other 37 K-3 schools. That is exactly what we see. The 21 K-3 schools most impacted by the suspension ban improved their percent

Met by 2.5% (32.2% to 34.7%). That is only half as much improvement as the 37 other K-3 schools (44.0% to 49.1%).[37] Table 3 shows that this gap is not statistically significant, but that is primarily a function of the small number of schools in the analysis. The fact that the "Experienced Ban" coefficient is similar in both K-3 columns indicates that the gap is not the result of differences in demographics or prior achievement.

Analyses of these four subsequent suspension bans are not as robust as LAUSD's. With the exception of San Francisco, we cannot conduct a full difference-in-differences analysis because we lack data to measure the change in academic growth. It is possible that Pasadena, Oakland, and the twenty-one K-3 schools that gave defiance suspensions would have had relatively low growth even without the suspension ban – for reasons not accounted for by the independent variables in the Table 3 regressions. But the cumulative evidence makes that exceedingly unlikely. The analyses of LAUSD and these four subsequent bans combine to form a strong case that suspension bans have a significant and negative causal impact on academic growth.

Conclusions and Implications

Despite lingering uncertainty, the evidence in this paper is important for three primary reasons. First, it is closer to the experimental ideal than the current literature. It is a significant improvement on work that correlates suspension rates and achievement (Losen et al. 2015). It does not rely on matched comparisons between students who get suspended and students who do not, as does Arcia (2006). Nor does it rely on the assumption that natural fluctuations in suspension rates cause subsequent changes in

[37] We see similar but slightly smaller gaps when we measure achievement in Math or use average scale scores.

academic performance, as does Perry and Morris (2014). This paper relies on the LAUSD board decision to ban suspensions based on defiance to impose a sudden, intended shift in suspension rates among a particular subset of California schools. The standard of quality for social science research is not perfection, but providing new information with the best available method (Gerring 2012).[38]

This paper also complements recent research on discipline reform. Philadelphia, Pennsylvania saw a decrease in academic performance after its reforms (Steinberg and Lacoe 2017). New York City saw a worsening school climate after discipline reforms during the 2014-15 school year, with many more schools seeing higher shares of students reporting violence, drug use, gang activity, and disrespect (Eden 2017). Teacher surveys from school districts as diverse as Baton Rouge, Louisiana, Syracuse, New York, and Oklahoma City all suggest that discipline reforms make the average teacher feel less safe (Eden 2018). The harm to school climate is likely an important mechanism by which suspension bans harm academic growth. As school climate deteriorates, it becomes more difficult for teachers to teach and students to learn. In at least one district, discipline reform appears to have increased teacher turnover (D'Orio 2018). This, too, may harm academic growth.

[38] Student-level data could enable a more accurate estimate of the LAUSD suspension ban's impact. Ideally, the cohorts used to measure two-year growth would only include students continuously enrolled in schools during that time. Additionally, exact counts of the numbers of students – ideally in those cohorts, not the entire school – who received out-of-school suspensions for defiance would allow the calculation of 2013 suspension rates for the same groups of students for whom we measure academic growth. Student-level data would enable use of a multilevel model and make it more likely that estimates are statistically significant.

This chapter's results have implications for the theoretical disagreement within the literature. None of the analyses support the more prominent strand of the literature claiming that sudden reductions in suspensions will cause academic growth. In contrast, all of the analyses provide evidence in favor of the strand of the literature claiming that sudden reductions in suspensions will reduce academic growth. Perry and Morris' (2014) finding that reducing suspensions would improve academic growth for non-suspended students appears to be driven by omitted variable bias. Unmeasured factors that caused drops in suspension rates in their sample also caused increases in academic growth. Arcia's (2006) finding that suspensions academically harm suspended students may or may not suffer from similar bias. Our analysis of LAUSD lacks the data granularity to differentiate between types of students. It could be that the suspension ban in fact improved academic growth for students who would have been suspended and simultaneously harmed academic growth for all other students. However, research on discipline reform from Philadelphia found that previously suspended students did not experience improved academic growth after the policy change (Steinberg and Lacoe 2017). If this is also true in LAUSD, the implication would be that suspensions by themselves may not significantly harm academic growth for suspended students.

Such potential problems with previous research suggests a possible need to re-conceptualize suspensions. The literature has assumed that suspensions have important causal power. For example, suspensions are seen as a piece of the school-to-prison pipeline because suspensions are assumed to cause dropouts, which then facilitates trouble with law enforcement and prison. According to this theory, suspension bans

should cause substantial declines in dropout rates. Future research will need to test the relationship between suspension bans and dropout rates, but the preliminary data does not look promising. While suspension bans preceded declines in dropout rates in Los Angeles and San Francisco, bans were followed by only a minor reduction in dropout rates in Oakland and an increased dropout rate in Pasadena. Additionally, it could be that the declines we see in Los Angeles and San Francisco are the result of other changes in district policy. This suggests that at least some suspensions may be better conceptualized not as a root cause, but as a symptom of other issues. While suspension bans address the symptom of suspensions, bans will not reliably improve student outcomes if the root causes remain untouched.

Table 4*:* Annual Dropout Rates for Districts that Banned Suspensions

	2011-12	2012-13	2013-14	2014-15	2015-16
Los Angeles	6.1	6.3	3.9*	4.3	3.4
San Francisco	4	6.3	7.4	1.4*	2.5
Pasadena	5.5	4.5	4.7	3.4	4.6*
Oakland	8.5	7.4	5.3	6.7	5.8*

Note: In each row, the number with an asterisk indicates the first year after the suspension ban.

The second reason this evidence is important is that the impact on academic growth might be substantial. A full school year of academic growth is roughly equivalent to one standard deviation; the analyses suggest the suspension ban had an impact of approximately 20% of a standard deviation on schools that gave those suspensions. If true, this would be an enormous impact. This would be a larger impact than shifting from a bottom quartile teacher to a top quartile teacher (2012 Gathering Feedback for

Teaching).[39] The same suspension ban would have been imposed statewide in 2012 if California Governor Gerry Brown had not vetoed it (Watanabe 2013a). Although there is some uncertainty around this 20% estimate, the fact that it could have such a large negative impact should make people very hesitant to encourage the adoption of suspension bans. Even significant amounts of money simultaneously spent on teacher training in behavior management, as in Oakland, does not appear to prevent the academic harm.

This should give pause to the variety of actors promoting suspension bans. At best, the lack of statistical significance in some analyses means that LAUSD's suspension ban may have had no causal impact. There is no evidence allowing us to say that the ban improved academic growth, and the totality of evidence heavily favors the conclusion that the ban harmed academic growth. Furthermore, bans will harm growth in schools that previously gave defiance suspensions, and these schools disproportionately serve African America and Latino students. In other words, suspension bans appear to widen the racial achievement gap. Bans had the benefit of being simple to implement, had virtually no immediate financial cost, and produced immediate results. Other approaches to suspension and dropout reduction will be relatively complicated, cost time and money (e.g. for staff training), and may take several years to see results. Nevertheless, this approach seems much preferable to widening the racial achievement gap, harming academic growth for a huge number of students, and not reliably reducing dropout rates.

[39] The Measures of Effective Teaching (MET) project rigorously evaluated the performance of nearly 3,000 teachers. It found that teacher performance is most accurately predicted by a combination of classroom observations, student surveys, and value added measures from standardized tests.

The final reason this evidence is important is that school discipline is an active area of policy debate and experimentation. Betsy Devos, President Donald Trump's pick for Education Secretary, has held meetings to discuss whether to change the Department of Education's stance on school discipline. Conservative groups are advocating that the federal government should resume its limited role of investigating particular complaints of unfair practices. This would be quite a departure from the Obama-era policy of pressuring school districts to reduce suspensions if rates between racial groups were not equitable. This paper indirectly weighs in on this debate: if the federal government wants school districts to reduce suspensions, districts should also evaluate their changes to ensure that their efforts reliably reduce dropout rates and do not have the unintended consequence of harming academic growth.

What should districts do that have already implemented suspension bans? LAUSD School board member Richard Vladovic voted to ban suspensions based on defiance "as an experiment, saying he would be 'the first to stop it' if it proved disruptive to learning" (Watanabe 2013c). This brings up an important point: just because a suspension ban decreases academic growth does not mean that reversing a ban will cause an increase. Unfortunately, it is probably easier to harm academic growth than to help it. Also, it is possible that training in restorative justice or other behavior management practices might cause an increase in academic growth without having to resume giving

suspensions for defiance.[40] Districts would be wise to try a variety of more gradual options, evaluate their impacts, and then make an informed choice as to how to proceed.

This highlights the broader need for policy evaluation at the school district level. School board members and district leaders rarely know with certainty how a policy is likely to impact academic growth, and they often do not conduct evaluations of the policies they implement. An array of factors contribute to this problem, ranging from the relatively weak research base in education to the political cost of having to admit that past decisions led to bad outcomes. We are far from making evaluation a routine component of district policy decisions. The aim of this paper is to nudge us in that direction. Future researchers may find ways to reduce suspensions that also increase academic growth. Even more important would be if future researchers regularly help to inform school district leaders as they make important decisions concerning the education of our children.

[40] Recent research finds that restorative justice training led to further reductions in suspension rates (Hashim et al. 2018). Future research will need to look at the impact of this training on academic growth.

THE GROUPING DILEMMA: HOW BETWEEN-CLASS ABILITY GROUPING IMPACTS GROWTH AND EQUITY

Overview

Between-class ability grouping is the practice of using prior academic achievement to place students within a grade in different classes for either a single subject or the full school day.[41] The core curriculum remains constant, but higher ability classes often cover content at greater depth or breadth because they can move at a faster pace. United States public schools have utilized between-class ability grouping in various forms for over a century, starting in Santa Barbara, California sometime soon after 1900 and Detroit in 1919 (Otto 1941; Courtis 1925).

It is helpful to conceptualize between-class ability grouping as the midpoint between two other methods of sorting students. One is within-class ability grouping, which occurs when teachers sort students in a single class into groups and then provide distinct instruction to each group at different times. This is a version of differentiated instruction. Such groups tend to be extremely flexible, changing as frequently as the teacher can reevaluate students' abilities. The other extreme is tracking, where students sorted into different classes that experience different curricula. Tracked groups tend to be very inflexible because even if students are motivated to change tracks, they have missed curricular content that the other track covered. Between-class ability grouping is less flexible than within-class grouping but more flexible than tracking. The practice gives

[41] Between-class ability grouping is also known as ability-grouped class assignment (Slavin 1987), multilevel grouping (Kulik and Kulik 1992), multitrack grouping (Miles 1954), and XYZ classes (Kulik 1992). The most common term in current use is "between-class ability grouping," so we use either that term or simply "ability grouping" with no qualifier to refer to this practice.

the same curriculum to all students like within-class grouping, but it separates students into distinct classes like tracking. In other words, between-class ability grouping is differentiated instruction at the classroom level.

For nearly all of its history, the literature has concluded that between-class ability grouping has no impact on academic growth (e.g. Miller and Otto 1930, Ekstrom 1961, Kulik 1992, Steenbergen-Hu et al. 2016). However, research from the past decade suggests that between-class ability grouping increases academic growth (e.g. Duflo et al. 2011; Collins and Gan 2013). These researchers posit that the cause is a clustering mechanism: decreased classroom dispersion in prior academic ability allows teachers to target their instruction more narrowly.

This paper advances the literature by specifically testing this clustering mechanism. It does so with student-level data from Fortune School of Education, an organization that operates a network of elementary and middle charter schools that adopted an evolving policy of ability grouping during the 2016-17 and 2017-18 school years. Approximately half of the observations are of students experiencing between-class ability grouping; the other half are of students who were not sorted.[42] However, the primary independent variable is not whether or not there is between-class ability grouping, but the degree to which students in a class are clustered by prior ability. There is significant variation in clustering across classes whether or not they experienced between-class ability grouping. Instead of comparing a treated group to a control, this

[42] In the sample, students in grade 1 were never sorted into between-class ability groups and students in grade 2 were only sorted in one of the four trimesters. Students in grades 3 through 8 could not be sorted when there was only one class in that grade at that site. Every grade had some students who were not sorted.

paper exploits variation in the strength of the treatment. This allows us to partially assuage concerns of omitted variable bias because we do not need treatment and control groups to be equivalent on all observed and unobserved characteristics. We exploit the fluctuating strength of the treatment, which is hopefully quasi-random, for the same students over time.

Across a range of models and specifications, we find a substantially small but statistically significant effect: a one standard deviation decrease in clustering causes a 7% decrease in overall academic growth. Contrary to the expectations of the literature, between-class ability grouping slightly harms overall academic growth.

However, between-class ability grouping also improves equity by boosting growth for the lowest-ability students. When between-class ability grouping is not implemented, students in the third quartile benefit the most from classroom dispersion. This suggests that teachers confronting classrooms with wide variation in student ability focus their instruction at students in the third quartile, just above the class average. When there is ability grouping, students in the bottom quartile see a substantial boost in academic growth, helping them catch up to their peers. This is probably because teachers with only lower-ability students in a classroom are able to focus their instruction at a level that is much more appropriate for students in the bottom quartile. Unfortunately, between-class ability grouping also lowers growth for students in the top quartiles. If schools were able to implement between-class ability grouping in a way that did not decrease growth for higher-ability students, this practice would improve equity as well as overall academic growth.

The paper proceeds in four parts. First we provide background on ability groups, both the history of the practice in the US and how this practice came to be adopted at the network of Fortune School of Education charter schools we analyze. Then we discuss the competing evidence in the extant literature about the impact of ability groups. Third, we describe the data and analyses used to evaluate the hypothesized mechanism of ability groups: decreased classroom dispersion in prior ability allows teachers to target their instruction more narrowly. Finally, we discuss the limitations of the results as well as the implications for practitioners.

Background

The first recorded use of between-class ability grouping is from an undated year soon after 1900 in Santa Barbara, California (Otto 1941). However, it appears that few people learned about ability grouping through this or other early adoptions by relatively small school districts. Instead, Detroit popularized the practice after adopting it in 1919 (Courtis 1925). Detroit school leaders used standardized assessments to divide first graders into three types of all-day classes: the top 20%, the middle 60%, and the bottom 20%. This particular format became the most common form of ability grouping in the United States (Kulik 1992).

Unfortunately, we have limited data about the proportion of schools nationally that have utilized between-class ability groups. National surveys have either focused on within-class ability groups or posed questions that could apply equally to partial sorting, between-class ability groups, or even tracking. The best we can do is look at those trends and speculate on what that is likely to mean for between-class ability groups. The earliest

data we have is from 1961, when a survey found that 80% of elementary schools utilized within-class ability grouping in English (Austin and Morrison 1961). This rate appears to be relatively consistent up through the mid-1980s (Loveless 2013). However, the rate was only 28% in 1998 when the National Association of Educational Progress (NAEP) began surveying fourth grade teachers about their use of within-class ability groups. In that year, another 33% of fourth grade students were in reading groups based on "interest," "diversity," or "other" factors. The remaining 39% were not placed into groups at all.

Did within-class ability grouping in elementary English classes really decline from 80% to 28% from the mid-1980s to 1998? NAEP surveys about eighth grade English classes suggest that such a change is plausible. The eighth grade survey asked school leaders whether students were "typically assigned to classes by ability." Unfortunately, this phrasing could reflect three distinct practices: partial sorting, full between-class ability grouping, or tracking (if the classes had different curricula). We will say the survey reflects "sorting/tracking" to reflect this ambiguity. Nevertheless, in the earliest survey from 1990, 60% of students were sorted/tracked in English classes. The rate dropped to 50% in 1992 and 37% in 1994. By 1998, only 32% of eighth grade students experienced English sorting/tracking. Since the survey reflects multiple practices, it is likely that less than 32% of eighth graders experienced between-class ability grouping in that year. That is very similar to the fourth grade ability grouping rate (28%) in 1998.

Why did practices in both these grades experience such sharp declines? Loveless argues that the media published criticisms of these practices during that period, and that "educators are aware of public debates and are influenced when particular school practices become controversial" (2013, p. 19). As evidence, Loveless calculates the number of times "ability grouping" was mentioned each year in the magazine *Education Week*. While most years mention that term no more than three times, the term was mentioned between five and twenty times each year from 1989 to 1998 (Loveless 2013). This is exactly the timeframe in which surveys show that eighth grade sorting/tracking declined, as well as when fourth grade within-class ability grouping declined.

After 1998, however, the practices rebounded. NAEP data shows within-class ability grouping steadily increasing from 28% of fourth graders in 1998 to 71% in 2009.[43] This is almost a return to the 80% level observed in the mid-1980s and previous decades. Such an increase is especially impressive given the fact that within-class ability grouping tends to be most common in first grade, declining in use as students move to higher elementary grades (McPartland et al. 1987). Therefore, the NAEP survey is likely to underestimate the percent of US elementary students who experience within-class ability grouping. We only have comparable eighth grade sorting/tracking data from one year: 43% in 2003. The rate for fourth grade ability grouping in that year was 47%, suggesting that sorting/tracking were at least on a similar trajectory from 1998 to 2003. NAEP surveys did not ask about sorting/tracking after 2003, so we do not know whether or not this trend continued. Even if the trend did continue like within-class ability grouping, we

[43] NAEP surveys after 2009 have not asked about ability grouping.

would not know whether this reflects partial sorting, between-class ability grouping, or tracking.

The remainder of this section describes how Fortune School of Education (FSE), a charter management organization in California, implemented between-class ability grouping. While the responses of the school leaders and teachers in this system are only anecdotal evidence, it suggests that between-class ability grouping was far from a widespread, accepted practice by 2016.

In the 2016-17 school year, FSE operated six charter schools serving primarily low-income African-American and Latino students from Kindergarten to eighth grade. In an effort to increase academic growth, FSE central office leaders and principals decided to group students by ability into English classes in grades three through eight. At the beginning of the 2016-17 school year, schedulers sorted students into English classes based on their score on either (1) the interim assessment from the end of the prior school year or, if that was unavailable, (2) a placement exam given at the very beginning of the school year. The lowest level class was called "Rising," the highest was called "Benchmark," and the middle level (when there were three classes in a grade) was called "Strategic." When grades were divided into two classes, schedulers tried to find one cut score (e.g. the 40th percentile) for each school that would result in reasonable class sizes for all grades. Principals determined when the cut score needed to be changed because certain class sizes were too large or small. Principals needed to obtain central office approval in order to place students in a different class.

Elementary students were in mixed ability groups for all other subjects and then moved into ability groups for English class. Middle school students took all their classes with their English ability groups. All ability groups utilized the same curricula and administered the same interim assessments at the end of each trimester. Scores on those interim assessments could result in students moving into a higher or lower level English class for the next trimester; an average of 17% of students changed levels in each of the four transitions for which we have data. By design, lower level classes had more instructional aides and fewer students. During the school year, all English teachers received the same professional development regardless of whether they taught lower or higher level classes.

This shift to ability groups was controversial internally. Recent academic research (Collins and Gan 2013) as well as anecdotal evidence from some high-performing schools suggested that this could result in higher academic growth, especially for low-achieving students, by allowing teachers to target their instruction more precisely. However, many FSE principals and central office staff had been taught that such sorting was akin to tracking and would harm student growth. Additionally, staff understandably were concerned about unintended negative consequences: lower-achieving classes could suffer from lower expectations, lower quality resources, and psychological stigma. FSE therefore attempted to implement ability groups in a way that avoided or minimized these potential problems. Use of standard curricula and assessments helped to ensure that all students were held to the same high expectations. The fact that lower-level classes were smaller and had instructional aides ensured that

these students received a resource advantage compared to higher-level classes. Class names (Rising, Strategic, and Benchmark) and the fluidity of the system attempted to prevent negative psychological impact on students. Instead of feeling stigmatized and stuck, students in low level English classes could feel comfortable or motivated to advance into a more difficult English class the next trimester.

Enrollment patterns dictated the number of ability groups in every grade (third through eighth) at each school. FSE charter schools were still building up to full enrollment because they were relatively new – between two and seven years old. Therefore, schools had anywhere from one to three classes per grade. Grades with only enough students for one class could not implement ability groups at all; grades with two or three classes would have two or three ability groups, respectively. In other words, enrollment patterns conditioned the implementation of ability groups.

FSE made three policy changes for the 2017-18 school year. Two were relatively minor. One was to use ability groups in grade two as well as grades three through eight. The other change was to more rigorously incorporate class size into the placement process. Schedulers still started by trying to find one cut score (e.g. the 40th percentile) for each school that would result in reasonable sizes for Rising and Benchmark classes. However, instead of relying on principals to determine when to adjust for class size, we automated the process with a two part policy. First, Rising classes had to be smaller than Benchmark classes. Second, the difference in size between Rising and Benchmark classes could be no more than 10 students. This new policy essentially formalized the

input that principals provided during the first year of implementation, thereby saving principal time.

The most substantial change was to have teachers specialize in particular subjects. The policy was called "departmentalization," and it applied in grades two and higher whenever there were two classes in a grade at an elementary school.[44] Instead of teaching all subjects, staff taught two subjects: either English and social studies or math and science. They instructed the Rising class for half the day and the Benchmark class for the other half. This had a number of benefits. First, it substantially reduced lesson planning time because teachers led two similar lessons each day. Second, teachers were able to focus on their stronger content areas. Many teachers prefer one pair of subjects to the other, and departmentalization allowed them to build on their strengths. Third, this avoided the concern that the lower ability class may have a weaker instructor than the other; all students were taught the same subjects by the same teachers.

Changes between the 2016-17 and 2017-18 school years highlights an important theoretical point: ability grouping can be implemented in a variety of ways. Implementation details could have important effects on academic growth. Other schools may implement ability grouping in a very different way, perhaps leading to significantly better or worse results. However, the main purpose of this paper is not to evaluate the entire practice of ability grouping. Instead, this paper aims to evaluate the proposed clustering mechanism that connects ability grouping to academic growth: decreased classroom dispersion in academic ability allows teachers to better target their instruction.

[44] In fall 2017-18, there were no schools that had more than two classes for grades two through eight.

It is a strength of this paper that the data includes classes that experienced different implementations of ability groups as well as classes that experienced no ability groups at all. We will analyze these contexts separately to see if the clustering mechanism only appears in certain contexts. But before we get to the analyses, we need to look at the existing literature to see what evidence exists to support or oppose the existence of a clustering mechanism.

Theory and Hypotheses

The literature presents two competing perspectives on how between-class ability grouping impacts academic growth: a dominant view, and a recent view. The dominant view has been that that there is no significant effect on student learning. The recent view is that between-class ability grouping has a positive effect on student learning. Proponents of the recent view claim that the mechanism underlying their findings is a decrease in classroom dispersion. They hypothesize that by decreasing the range of academic ability in a classroom, between-class ability grouping allows teachers to target their instruction more narrowly for their students, thereby increasing academic growth. The main contribution of this paper is to directly measure and analyze this mechanism.

Miller and Otto (1930) first articulated the dominant view after reviewing twenty studies from the 1920s. Decades later, Ekstrom (1961) reached the same conclusion based on nine studies. Slavin (1987) and Kulik (1992) also find a null effect in their separate reviews of over fifty studies.[45] Most recently and comprehensively,

[45] This is corroborated by a review of the ability grouping literature in both the United States and Great Britain (Wynne and Malcolm 1999).

Steenbergen-Hu et al. (2016) synthesize 13 meta-analyses to conclude that between-class ability grouping has no significant effect.

One potential concern is that these reviews include studies of grouping across a wide range of grades and subjects. It is possible that the overall null finding masks a real effect when grouping occurs in English classes at the elementary or middle school level, the focus of this paper. Table 5 summarizes the only five studies that share that focus. The results is the same: three find a positive overall effect and two find a negative effect. The average effect size is an insignificant -0.02.

Table 5*: Analyses of Elementary/Middle English Class Ability Grouping*

Study	Effect Size	Grade(s) Analyzed	Methodology
Berkun, Swanson, & Sawyer (1966)	0.40	3-5	Grouped classes compared to non-grouped classes in nearby schools, adjusted for initial ability
Bremer (1958)	-0.10	1	Grouped students matched to non-grouped students from the subsequent year in the same schools
Moses (1966)	0.07	4-6	Grouped students matched with non-grouped students in nearby schools
Nichols (1969)	-0.95	1	Grouped students matched with non-grouped students in nearby schools
Tobin (1966)	0.46	2-6	Grouped students compared to non-grouped students from previous years in the same schools

Note: Table is adapted from Kulik (1992).

However, close examination of those five studies reveals significant methodological concerns. The most prominent problem with these studies is omitted variable bias. While the analyses control for students' initial ability and sometimes a few additional variables, many other factors could differ between the comparison groups. Berkun et al. (1966) compares classes from different schools and uses statistical adjustments in an effort to compensate for different average initial abilities. Moses (1966) and Nichols (1969) are slightly more rigorous, using student-level matching to

compare students at different schools. But all three analyses rest on the assumption that teacher, school, or uncontrolled student effects do not significantly bias their effect size. The two remaining studies take a different approach, making comparisons over time to the same schools and teachers (Bremer 1958; Tobin 1966). This controls for teacher and school effects, but it introduces the possibility that temporal effects bias their results. Additionally, differences between students not reflected by initial ability remain a concern.

In opposition to this traditional strand of the literature, two recent studies argue that ability groups may cause increases in academic growth. The more rigorous analysis is Duflo et al. (2011), which conducts a randomized controlled trial in Kenya. Researchers provided an additional teacher for 121 schools that had only one first grade class, and then randomly divided those schools into two groups: 60 grouped students by ability into two first grade classes, and 61 randomly assigned students to the two classes. The 60 schools with grouped classes experienced 0.14 standard deviations more academic growth. This effect was consistent across students with different prior ability levels, and it persisted a year after tracking ended. The authors contend that the primary mechanism was that decreased classroom dispersion allowed teachers to more narrowly target their instruction.

While the internal validity of Dulfo et al. (2011) is extremely strong, its external validity is limited. The authors readily acknowledge this, and they admit that addition research is necessary to determine the extent to which their findings are generalizable. Additionally, they discuss two systemic conditions that are likely to impact ability

grouping: initial heterogeneity in student ability, and teacher incentives to focus on students at a particular level. More initial heterogeneity among students would make ability grouping more consequential in allowing teachers to target their instruction. In contrast, limited initial heterogeneity would curtail the impact of ability grouping. Second, systems that incentivize focus on the highest ability students would likely see the benefits of ability grouping go primarily to low ability students. This is because ability grouping would allow teachers of lower ability classes to tailor instruction to their level. While these conditions may not hold in the United States in general, the next section shows that they hold in much of our sample.

The other recent study of ability groups looked at fourth grade students in Dallas Independent School District (Collins and Gan 2013). None of the schools utilized between-class ability grouping, so Collins and Gan measure the extent to which students are sorted by prior ability into different classes at each school. This is calculated using the difference in means between classes in the same grade; it is correlated with classroom dispersion, but quite a distinct measure. Collins and Gan then look for a relationship between the extent of sorting and subsequent academic growth. In an effort to avoid omitted variable bias, they use the extent of sorting in fifth grade as an instrumental variable to predict the extent of sorting in fourth grade. They find positive effects for high and low ability students in a range of model specifications. Following Duflo et al. (2011), the authors argue that a clustering mechanism explains their findings. However, Collins and Gan do not specifically test this clustering mechanism, and none of the schools explicitly utilized between-class ability grouping.

In summary, the literature on between-class ability grouping is divided between a dominant view (no significant effect) with weak internal validity and a recent view (a positive effect) with weak external validity. The next section describes our effort to advance the literature with an analysis of students at Fortune School of Education (FSE), a network of charter schools that implemented between-class ability grouping in ways that varied across grades and over time. Instead of attempting to evaluate between-class ability grouping as a whole, we aim to look for the clustering mechanism that the recent literature believes underlies its positive impact.

Data and Analyses

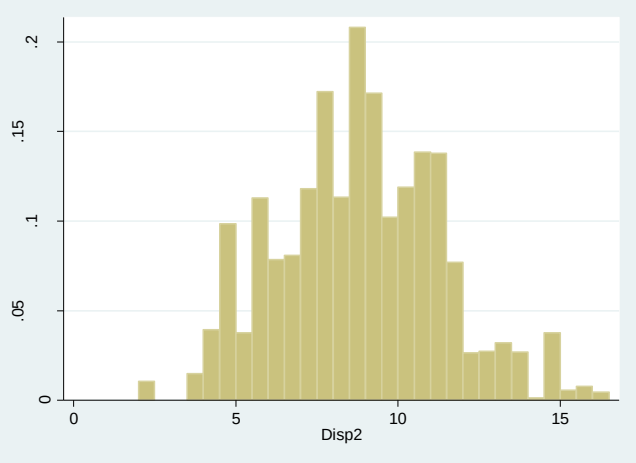
We use anonymized student-level data from Fortune School of Education (FSE) from all three trimesters of 2016-17 and the first trimester of 2017-18. Students take an assessment at the end of each trimester, and that score is used to re-assign students to classes. We therefore have up to four observations per student, one from each trimester in the dataset.

The data is in a four-tier hierarchical structure. The bottom tier is composed of 1,530 students: 311 with one observation (i.e. one trimester), 238 with two, 403 with three, and 578 with four. The result is a total of 4,308 unique data points. Students are nested within 234 distinct English classes, with each class lasting one trimester. Those classes are taught by 75 teachers, who teach one or more of eight possible grades (1 through 8) and are nested within six schools. The nested nature of the data along with multiple observations over time allow us to control for distinct student, teacher, grade, and school effects on academic growth.

We have two important independent variables. One is a dummy variable indicating whether or not between-class ability grouping occurred. Between-class ability grouping took place for approximately half of the observations; no sorting took place in the other half. Students in grade one were never sorted by ability, and students in grade two were only sorted in 2017-18 (for which we have one trimester of data). Students in grades three through eight were sorted as long as there were at least two classes in that grade at that school. Across all schools in our sample, each grade – except for grade one – has a mix of sorted and unsorted students.

The novel independent variable is classroom dispersion, which is the spread of prior achievement levels within a classroom. If this is the only mechanism by which between-class ability grouping impacts academic growth, then this variable should be significant while the dummy variable should be insignificant. We measure classroom dispersion by calculating the mean deviation of scale scores on the previous interim assessment. This includes all students currently in the class, regardless of which class they were in during the prior trimester. Figure 11 shows that the distribution of classroom dispersion is close to normal. The standard deviation of classroom dispersion is 2.5 scale score points, which is exactly the average difference between students who experience between-class ability grouping and those who do not (10.1 and 7.6, respectively). The distribution of classroom dispersion is also close to normal when we look at only classes where between-class ability grouping occurred or classes where this practice did not occur.

Figure 11: Histogram of Classroom Dispersion



Focusing on classroom dispersion directly tests the mechanism the recent literature argues is at work with between-class ability grouping: teachers can better target their instruction the more students are clustered around the same ability level. Theoretically, this mechanism should hold true whether or not students are formally grouped by ability into classes. This builds on the insight of Collins and Gan (2013), who measure the extent to which classes within a grade are sorted by ability. While the rest of the literature treats between-class ability grouping as a binary, they conceptualize it as a continuum. We conceptualize the mechanism underlying between-class ability grouping – classroom dispersion – as a continuum. While it is possible to partially sort students by prior ability, the observations in our sample either were fully sorted or not sorted at all.

For interim assessments, FSE uses the Northwest Evaluation Association's Measure of Academic Progress assessment (i.e. NWEA MAP) in both Reading and Math each trimester. FSE began administering these assessments in the second trimester of the

2015-16 school year, so by 2016-17 staff had familiarity with the assessment and understood that it would be important for internal decision-making. NWEA MAP is nationally normed and vertically aligned, allowing students in any grade to receive achievement scores at any point on the scale. In every grade level, the distributions of NWEA MAP scores in our sample approximate a normal distribution and have a standard deviation between 14 and 16. As expected, the average scale score increases in every grade. NWEA also reports each score as a national percentile ranging from 1 (lowest) to 99 (highest). Scores in our sample are fairly evenly distributed across this distribution, with two exceptions: a cluster of students in the bottom 5% and a relatively small portion of students in the top 5%.

Our goal is to measure the relationship between classroom dispersion and academic growth. We follow the methodology of value-added models, the best available metric of academic growth in the education literature (e.g. Kogan et al. 2016b).[46] Value-added measures of academic growth include the prior scale score as an independent variable to predict the outcome – the end of trimester (i.e. current) scale score. Adding additional covariates and controls to the model allows us to see if other factors impact academic growth even after accounting for the prior English score.

In Model 1 of Table 6, we see that classroom dispersion is associated with significantly higher academic growth. Since between-class ability grouping reduces

[46] The value-add literature has primarily focused on its controversial use in teacher evaluation. Using value-add estimates for entire grades avoids some concerns, such as sorting difficult-to-teach students into particular teachers' classes (Rothstein 2009). However, even grade level value-add measures can experience significant variation across time (Goldhaber and Hansen 2008) and can be sensitive to the assessment used (Lockwood et al. 2007). On the positive side, value-add measures are highly correlated with principal evaluations of teachers (Kimball et al. 2004; Jacob and Lefgren 2008).

classroom dispersion (by one standard deviation, on average), this implies that the practice lowered growth. Such a result defies the expectations of the traditional view (no effect) and recent view (positive effect) of the literature. Model 1 also shows that in a regression with just these current and prior test scores and classroom dispersion, the adjusted R-squared is 0.77 and the coefficient on prior score is 0.86. Removing classroom dispersion (not shown) causes the adjusted R-squared to drop to just 0.76. As the value-added literature would predict, prior and current NWEA MAP scale scores are highly correlated with each other.

The next four models in Table 6 add independent variables that control for other factors that may impact academic growth. Following Collins and Gan (2013), Model 2 controls for a variety of student and classroom factors that might impact student academic growth.[47] Three classroom-related factors are worth discussing in detail. Approximately half of our observations are of students in classrooms that were created via between-class ability grouping. Experiencing between-class ability grouping may have an impact independent from the classroom dispersion mechanism, so we control for that possibility. We also control for class size because past research has shown that class size impacts growth (e.g. Word et al 1990; Angrist and Lavy 1999; Molnar et al. 1999). Lastly, we control for the average prior score in the classroom because there is a large and growing literature about the importance of peer effects for academic growth (e.g.

[47] Unlike Collins and Gan (2013), we do not include independent variables for teacher experience or salary because we control directly for individual teacher effects. Similarly, we do not control for average school scores or size because we control directly for individual school effects. Our models include separate variables for free and reduced lunch status, while Collins and Gan (2013) combine these into one category. Our models do not control for gender or race, but including them leads to similar results (available upon request).

Sacerdote 2011; but for evidence against peer effects see Abdulkadiroğlu et al. 2014). After controlling for all these factors, the coefficient on classroom dispersion is slightly higher and continues to be statistically significant (p-value < 0.001) – not at all the prediction of the literature.

Model 3 includes all the covariates from Model 2 and adds prior Math score, a variable for which we have imperfect coverage. The classroom dispersion coefficient is nearly identical; subsequent models do not include prior Math score in order to avoid losing observations. Model 4 adds four types of fixed effects: grades, teachers, sites, and terms (i.e. the four trimesters). Adding these fixed effects shrinks the coefficient on classroom dispersion by half, but it remains positive and significant.

Table 6*:* OLS Models of Academic Growth

Model	**1. Basic OLS Model**	**2. Add Student and Class Factors**	**3. Add Prior Math Score**	**4. All but Student Control**	**5. Student Control**
Classroom Dispersion	0.437*** [0.063]	0.676*** [0.069]	0.569*** [0.067]	0.261*** [0.100]	0.377*** [0.085]
Prior Score	0.857*** [0.007]	0.699*** [0.013]	0.518*** [0.016]	0.695*** [0.013]	-0.177*** [0.020]
Ability Grouping	n/a	1.275*** [0.371]	0.257 [0.363]	-3.991*** [1.461]	6.160 [6.392]
Student and Class Factors	No	No	Yes	Yes	Yes
Prior Math	No	No	Yes	No	No
Grade, Teacher, School, & Term Controls	No	No	No	Yes	Yes
Student Control	No	No	No	No	Yes
Observations	4308	4308	4190	4308	4308
Adjusted R^2	0.766	0.776	0.793	0.794	0.882

Notes: Outcome is current NWEA MAP scale score. Student factors include dummy variables for being an English Learner, a student with a disability, a recipient of free lunch, and a recipient of reduced price lunch. Class factors include class size, average prior English score, and a dummy variable for ability grouping.

Model 5 includes all the variables in Model 4 and adds student fixed effects. This means each student is being compared only to themselves in different trimesters, not to other students. Model 5 measures the impact of differences in classroom dispersion that individual students experience from trimester to trimester. Such a test of classroom dispersion is quite different than Models 1 through 4. Impressively, the coefficient on classroom dispersion remains positive and significant; larger than Model 4, but not as large as Models 1 through 3. Even this test of the relationship between classroom dispersion and academic growth defies the expectations of the literature.

The contrast with the ability grouping dummy variable is stark. That variable is statistically insignificant in two models, and has opposite signs in the other two models. It appears that between-class ability grouping did not have any effect beyond the classroom clustering mechanism.

Given the nature of our data, mixed effects multilevel models are more appropriate than OLS regressions (Gelman and Hill 2007). One reason is that multilevel models explicitly deal with the nested nature of our data. Multilevel models can divide our observations into groups based on four nesting variables: terms (i.e. trimesters), followed by schools, followed by English teachers, followed by grades. This creates 206 groups with an average of 21 observations each.[48] The other advantage is that multilevel models apply random effects for each of the four nesting variables. This allows the intercept for academic growth to vary by each of those 206 groups.

[48] The number of groups (206) is slightly less than the number of distinct English classes (234) because in the one trimester of 2017-18 data, departmentalization caused many elementary teachers to teach two English classes in the same grade.

Table 7 shows the results of four multilevel models. The first includes all the same variables as Model 4 in Table 6. The difference is that while the OLS model includes the four nesting variables – term, school, teacher, and grade – as fixed effects, the multilevel model includes them as random effects. The coefficient on classroom dispersion continues to be positive and statistically significant (p-value < 0.001), with a magnitude nearly identical to Models 4 and 5 from Table 6. The second model in Table 7 is identical to the first except for one change: it nests teachers within grades instead of the other way around. Either approach leads to the same number of groups, and either approach is defensible; in some cases one English teacher spans multiple grades, while in other cases one grade contains multiple English teachers. Since our data covers two school years, a number of teachers change what grade they teach. Ultimately, this swayed us to use Model 1 as our main analysis, but it is reassuring that the coefficient for classroom dispersion is nearly identical in Model 2.

Table 7*:* Multilevel Models of Academic Growth

Model	**1. Random Effects, Grades within Teachers**	**2. Random Effects, Teachers within Grades**	**3. Fixed and Random Effects, Grades within Teachers**	**4. Fixed and Random Effects, Grades within Teachers**
Classroom Dispersion	0.391*** [0.108]	0.349*** [0.106]	0.203* [0.115]	0.173 [0.115]
Ability Grouping	1.474** [0.626]	1.179* [0.696]	-3.939** [1.950]	-4.159** [1.976]
Observations	4308	4308	4308	4308
Log likelihood	-15797	-15794	-15716	-15716

Notes: All models are mixed-effects multilevel regressions. Outcome is current NWEA MAP scale score. Models also include the following independent variables: prior English scale score, dummy variables for being an English Learner, a student with a disability, a recipient of free lunch, and a recipient of reduced price lunch, class size, and the class' average prior English scale score.

Models 3 and 4 in Table 7 mirror the first two but make one important change. They include the four nesting variables as fixed effects as well as random effects. Doing so dramatically reduces the power of random effects. This explains why the coefficients in these two closely match those in Model 4 of Table 6, an OLS model with fixed effects for our nesting variables. The classroom dispersion coefficient drops slightly and even becomes statistically insignificant in Model 4. Model 4 is one of the only analyses that does not find a positive and significant relationship between classroom dispersion and academic growth.

Table 8 looks for differences between important subsets of the data.[49] All the models in this table use Model 1 from Table 7, which we will refer to this as our main analysis: a mixed-effects multilevel model with only random effects for the four nesting variables. The first two columns of Table 8 separate students who were experiencing between-class ability grouping from students who were not. Each of those groups happens to be almost exactly half the dataset, and they span all grades and schools. The effect of classroom dispersion is similar for both groups, demonstrating that classroom dispersion is positively related to growth regardless of whether ability grouping is in place. This is an important confirmation of our theory that the mechanism of classroom dispersion operates at all times, not just when between-class ability grouping is implemented.

[49] Analyses by grade level (not reported) reveal similarly positive and significant effects for grades 1 and 2 and 3 through 5. However, the result for grades 6 through 8 is negative and significant. This is based on 719 observations, and 61% of those observations are from students taught at one middle school. That middle school happened to experience turnover of English instructors during the 2016-17 school year. While it could be that classroom dispersion has a different impact in middle school grades than in other grades, there is no obvious theoretical reason to expect this to be the case. Also, the factors mentioned above gives us reason to be suspicious of the relatively small amount of middle school data we have.

Model 3 in Table 8 isolates term 4, the first trimester of the 2017-18 school year. As mentioned previously, several changes occurred at the start of this school year concerning the implementation of between-class ability grouping. It is possible that these changes could alter the relationship between classroom dispersion and academic growth. However, Model 3 suggests the relationship remained unchanged: positive and significant. We will update this analysis to confirm that this is still true when more data from 2017-18 is available. Also, it is still possible that implementing between-class ability grouping in a different manner or context would result in a different coefficient for classroom dispersion. However, the consistency of our results across these subsets of data suggest that at least in this context, there is a real relationship between classroom dispersion and academic growth.

Table 8*:* Results by Ability Grouping and for Term 4

Model	**1. Ability Grouping**	**2. No Ability Grouping**	**3. Only Term 4**
Classroom Dispersion	0.327** [0.158]	0.393** [0.170]	0.347* [0.192]
Ability Grouping	n/a	n/a	0.677 [1.325]
Observations	2204	2104	1069
Log likelihood	-8098	-7685	-3860

Notes: All models are mixed-effects multilevel regressions that utilize random effects for grade, teacher, school, and term (i.e. Model 1 from Table 3). Outcome is current NWEA MAP scale score. All models also include the following independent variables: prior English scale score, dummy variables for being an English Learner, a student with a disability, a recipient of free lunch, and a recipient of reduced price lunch, class size, and the class' average prior English scale score.

In addition to overall academic growth, researchers and the public increasingly care about equity of outcomes across students. Before concluding the analysis section, we therefore want to see if classroom dispersion impacts students differently depending

on their prior ability level. Table 9 conducts the main analysis four times, once for each different quartile of prior English ability. The first quartile is the lowest performing and the fourth quartile is the highest performing. Every grade in the dataset was divided as closely into quartiles as possible. The average national percentile is 9 for the first quartile, 32 for the second quartile, 53 for the third quartile, and 78 for the fourth quartile.

The classroom dispersion coefficient for elementary grades are relatively consistent and always positive. The coefficient is largest and significant for students in the third quartile, which appears to reflect the incentives facing these – and many other – schools. The No Child Left Behind federal education law taught schools to measure their performance in terms of the percent of students who are at or above the standard for proficiency. This is approximately the 60^{th} national percentile, which is near the average for the third quartile of students in our sample. Therefore, we would expect teachers and administrators to focus particularly on achieving academic growth with this third quartile. The more classroom dispersion there is, the more this focus is at the expense of other students. There is a large negative coefficient on ability grouping for the third quartile because when ability grouping occurs, those third quartile students no longer receive the same focused attention from all teachers.

Results for the first two quartiles reflect the opposite story. Their coefficients for ability grouping are positive, and for the first quartile are significant. This is most likely because between-class ability grouping allows these students to receive more focus than they would in a non-sorted class. Therefore, ability grouping improves equity by enabling more teachers to focus on the lowest achieving students. This part of our

hypothesized clustering mechanism appears to be true. However, ability grouping reduces overall academic growth because of diminished growth among other students – especially those in the third quartile. It is unclear why higher ability students do not similarly benefit from the clustering mechanism. Whatever the reason, the result is that ability grouping leads to a more equitable, but slightly lower on average, distribution of academic performance.

Table 9*:* Results by Grade Level and Student Ability

	Quartile of Prior ELA Ability			
Quartile	**1st**	**2nd**	**3rd**	**4th**
Classroom Dispersion	0.160 [0.210]	0.211 [0.189]	0.360** [0.180]	0.112 [0.210]
Ability Grouping	3.313*** [1.071]	0.892 [1.036]	-1.584 [1.097]	-0.849 [1.181]
Observations	1124	1094	1054	1036
Log likelihood	-4301	-4000	-3752	-3667

Notes: All models are mixed-effects multilevel regressions that utilize random effects for grade, teacher, school, and term (i.e. Model 1 from Table 3). Outcome is current NWEA MAP scale score. All models also include the following independent variables: prior English scale score, dummy variables for being an English Learner, a student with a disability, a recipient of free lunch, and a recipient of reduced price lunch, class size, and the class' average prior English scale score.

Conclusions and Implications

The evidence from our sample shows that more classroom dispersion in prior ability increases academic growth. This is exactly the opposite of what the recent literature would expect, and defies the null hypothesis of the traditional literature. With a few small exceptions, the results for classroom dispersion are remarkably consistent across a variety of model specifications and subsets of the data. In contrast, our ability grouping dummy variable has a wide range of positive and negative coefficients across

different models. Therefore, between-class ability grouping does not appear to impact overall academic growth independent of the clustering mechanism.

Why does between-class ability grouping decrease overall academic growth? While it is difficult to know for sure, the analysis in Table 5 does suggest a hypothesis. Ability grouping increased academic growth for students in the bottom quartile, but decreased growth for students in the top two quartiles. In other words, teachers were able to make use of decreased classroom dispersion to help low-ability students, but not high-ability students. Perhaps teachers did not have the training or tools to push high-ability students especially quickly. Since the No Child Left Behind Act in 2000, schools in the United States have been incentivized to help students get to proficiency, not any higher. This incentive will go away as states transition to accountability systems that value growth in all parts of the achievement spectrum. Perhaps targeted professional development in this area would help teachers accelerate the growth of high-ability students.

The fact that minimizing or maximizing classroom dispersion – a practice that is virtually free – could impact academic growth is noteworthy. Additionally, the impact estimates are small but similar to other factors important to academic growth. The difference between top and bottom quartile teachers is approximately 10% of a standard deviation (Kane and Staiger 2012).[50] In our sample, implementing between-class ability

[50] The Measures of Effective Teaching (MET) project rigorously evaluated the performance of nearly 3,000 teachers. It found that teacher performance is most accurately predicted by a combination of classroom observations, student surveys, and value added measures from standardized tests.

grouping decreases classroom dispersion by an average of one standard deviation. That translates to an estimated 7% standard deviation decrease in academic growth.

This paper advances the literature on between-class ability grouping by focusing on one mechanism: decreasing classroom dispersion in prior ability allows teachers to focus their instruction more narrowly. When recent studies find a positive effect from ability grouping, they claim this mechanism explains their results (Duflo et al. 2011; Collins and Gan 2013). Our finding does not support their claim. This clash is likely because this paper studies a very different context. Indeed, Duflo et al. (2011) readily admit that their analysis of Kenya may not generalize to developed countries such as the United States. For Collins and Gan (2013), the difference may be caused by the change to Common Core standards. Instead of memorization and recall, these standards emphasize reading comprehension and critical thinking.

Of course, there may be factors specific to Fortune School of Education (FSE) that limit the generalizability of our findings. As a network of charter schools, FSE has the ability to implement and tweak policies much more quickly than many school districts. Additionally, FSE students are primarily low-income African-American and Latino, and their families had to choose to have their children attend these schools. Teachers tend to be relatively young, and both teacher and student turnover is relatively high. These factors, either singly or in combination, could condition the way classroom dispersion impacts academic growth.

It is also important to note that in some contexts, between-class ability grouping may affect academic growth through mechanisms besides this clustering mechanism. In

an effect to detect this, we report the results of an ability group dummy variable for each analysis. The coefficient on this variable is frequently insignificant, occasionally attaining significance in different directions in different model specifications. This pattern suggests that at FSE, ability grouping did not have an impact on academic growth outside of the clustering mechanism.

We therefore can conclude that ability grouping at FSE has not had a null effect, as the traditional literature would predict. There are methodological reasons to trust our findings over that of the traditional literature. We employ multilevel models in order to allow for random effects from grades, teachers, schools, and terms. We test a variety of model specifications and find consistent, significant effects from classroom dispersion. We even find a similar effect when we control for student fixed effects.

There is a possibility that when implemented in particular ways, between-class ability grouping could have an average effect beyond clustering students by prior ability. On the negative side, students may disengage if they feel that they are trapped in lower-level classes. On the positive side, allowing elementary teachers to "departmentalize" and teach only half of the subjects in their grade may eventually result in higher academic growth when this practice is combined with ability grouping. Future research could look for these and other possible mechanisms linking between-class ability grouping to academic growth.

Turning to implications for practitioners, there are two important sources of uncertainty. One is about the generalizability of our findings: the relationship between classroom dispersion and academic growth may be influenced by the context of FSE.

The second source of uncertainty is about ways of implementing ability grouping: there could be other mechanisms by which ability grouping could influence academic growth. Given these issues, the best course of action is for school leaders focused on overall academic growth to be cautious about adopting between-class ability grouping and to test its impact in their specific context.

However, we care about both growth and equity. Our findings suggest that while maximizing dispersion would boost overall academic growth in elementary grades, it would do so primarily by helping students in the middle of the distribution. In contrast, between-class ability grouping would boost growth for students in the bottom half – and especially the bottom quartile – of the distribution, while decreasing growth for higher-ability students. In other words, ability grouping would make a school's scores more equitable, but at a cost of approximately 7% of a standard deviation in overall academic growth. There is no easy way to reconcile tradeoffs between the important values of equity and overall growth. Ideally we will find less costly ways to help the lowest-ability students catch up to their peers.

Perhaps there is a way to implement between-class ability grouping that has an overall positive impact in elementary grades. It is possible that specific curricula or professional development could change the relationship between classroom dispersion and academic growth. Maybe with additional training on how to push a higher ability classroom, teachers could help top quartile students benefit from ability grouping as much as bottom quartile students. This could allow between-class ability grouping to

improve both equity and overall academic growth in elementary schools. Hopefully researchers and practitioners will work together to find such a solution soon.

APPENDIX A

The 2014-15 data for LAUSD is either zero or redacted, with just four exceptions.[51] Supplemental manual data collection allows us to obtain exact suspension rates for all but 83 LAUSD schools.[52] This reveals that the 83 remaining schools have an average of 1.5 suspensions each.[53]

There is a bit more uncertainty about the number of suspensions in 2012-13. Data files contain exact information for schools that report either zero or more than ten suspensions based on defiance for all subgroups. The more difficult cases are schools where some or all of their suspension data is redacted: 459 redacted data points account for 1166 suspensions. Manual data collection allows us to know the exact number of suspensions – totaling 674 – for another 107 LAUSD schools. The 204 schools with one redacted subgroup therefore share the remaining 492 suspensions (1166 minus 674) – an average of 2.4 suspensions each. The larger number of schools with unknown suspension counts (204 vs. 83) combined with those schools' higher average number of suspensions (2.4 vs. 1.5) results in a little more uncertainty surrounding the estimates for 2012-13 than for 2014-15.

---

[51] Four LAUSD schools – only one included in the analysis, Edwin Markham Middle – had at least one subgroup with over ten suspensions based on defiance in 2014-15, for a combined total of 75 suspensions. Subtracting these 75 suspensions from the 305 we know occurred district-wide leaves 230 possible suspensions.

[52] Schoolwide suspension counts are only available from individual school pages a state website (not in a dataset format), and it is redacted whenever reporting it would reveal a redacted subgroup. Looking up each of the twenty-one schools that had two or three redacted subgroups allows us to explain 106 of the remaining 230 suspensions that occurred district-wide (see above footnote).

[53] The remaining 124 suspensions must be spread across eighty-three schools that have only one redacted subgroup each. While it is possible that a very small number of these schools have up to ten suspensions, mathematical constraints assure us that the vast majority of these school have only one or two.

The uncertainty motivates us to measure "change in suspension rate" categorically rather than continuously. Replacing unknown redacted data with the average number of suspensions per subgroup would allow us to estimate a continuous variable, but this measure would contain errors for schools with unknown redacted data. One approach uses those estimated suspension rates to place LAUSD middle schools into one of one of four categories:

1. Increased suspension rate (12 schools)
2. Maintained suspension rate (25 schools)
3. Decreased suspension rate <=1% (44 schools)
4. Decreased suspension rate >1% (13 schools)

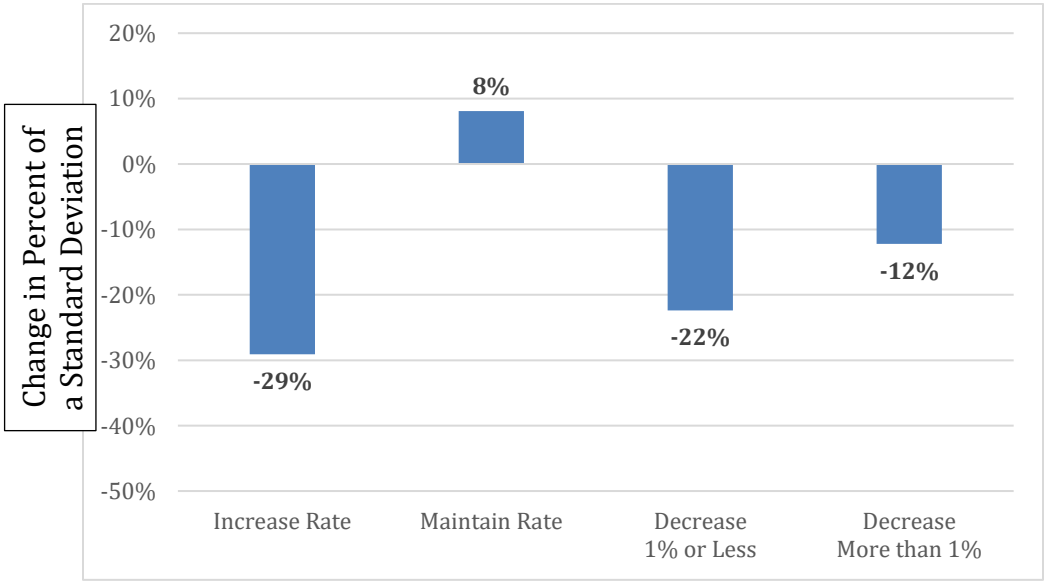
For 13 LAUSD schools in our sample, we are uncertain whether they experienced an increase or decrease in suspensions because they had one redacted subgroup in one year and between 1 and 10 suspensions in the other year. We run robustness tests without these schools and find very similar results.

There is significant overlap between this categorization and the one (in the paper) based on the number of suspensions given in 2013. All but 1 of the 13 schools that decreased suspension rates more than 1% had at least eleven suspensions in 2013. All but 3 of the 44 schools that decreased 1% or less had fewer than eleven suspensions in 2013. All 25 schools that had no change in suspension rate had no suspensions in 2013. The main difference is that the first categorization identifies and groups the 12 schools that experienced increases in suspension rates; 6 had one redacted subgroup (i.e. probably

one or two suspensions), five had between two and twelve reported suspensions, and one had thirty-five suspensions in 2013.

To test hypothesis three, we conduct analyses of schools within LAUSD. The two strands of the literature predict that the decrease (or increase) in academic growth should be correlated to the decrease in suspension rate. Figure 12 shows a non-linear relationship between suspensions and growth. Schools that maintained a suspension rate saw an increase in academic growth. Schools with a decrease of 1% or less had a 22% drop in growth. Schools with at least a 1% decrease in suspension rate experienced a decrease approximately half the size: 10% of a standard deviation. The pattern for these three right-hand columns does not perfectly fit either strand of the literature, but it is closer to the one that predicts suspension bans harm academic growth.

Figure 12: Change in Academic Growth by Change in Suspension Rate



The largest decrease – 29% of a standard deviation – was among the dozen schools that had an increase in suspension rate. Four of these schools had 0 defiance suspensions in 2013, seven of them had between 1 and 10, and one school had 40. Although this finding is not driven by a particular outlying school, evidence does not suggest that giving more suspensions in 2015 caused this academic decline. Most of these schools had extremely small changes in suspension rate, and the five schools with changes above 0.2% experienced almost no change in academic growth (-5%). Additionally, it is noteworthy that these twelve schools had the lowest 2013 growth rate of all the groups of LAUSD schools we analyze, as shown in Table 10. This suggests a different explanation: that at least some of these schools had other problems which led to both their non-compliance with the suspension ban and a further decrease in their academic growth rate.

Table 10*:* Academic Growth Rates by Change in Suspension Rate

	Academic Growth 2011 to 2013	**Academic Growth 2013 to 2015**
Increased Suspension Rate	-21%	-51%
Maintained Suspension Rate	5%	13%
Decreased Suspension Rate 1% or Less	-10%	-32%
Decreased Suspension Rate More than 1%	-9%	-21%

Note: Academic growth rates are reported in standard deviation units.

BIBLIOGRAPHY

(2007). History: Twenty-five years of progress in educating children with disabilities through IDEA. U.S. Office of Special Education programs, accessed https://www2.ed.gov/policy/speced/leg/idea/history.pdf.

(2012). Gathering Feedback for Teaching. *Bill and Melinda Gates Foundation*, MET Project research paper, accessed file:///C:/Users/dzarecki/Downloads/MET_Gathering_Feedback_Research_Paper1.pdf.

(2014). U.S. Departments of Education and Justice Release School Discipline Guidance Package to Enhance School Climate and Improve School Discipline Policies/Practices. Jan 8, accessed https://www.ed.gov/news/press-releases/us-departments-education-and-justice-release-school-discipline-guidance-package-.

Abdulkadiroğlu, A., Angrist, J., & Pathak, P. (2014). The elite illusion: Achievement effects at Boston and New York exam schools. *Econometrica*, *82*(1), 137-196.

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

Arcia, Emily. (2006). Achievement and Enrollment Status of Suspended Students: Outcomes in a Large, Multicultural School District. Education and Urban Society 38(3):359–69.

Asen, R., Gurke, D., Conners, P., Solomon, R., & Gumm, E. (2013). Research evidence and school board deliberations: Lessons from three Wisconsin school districts. *Educational Policy*, *27*(1), 33-63.

Austin, Mary C. & Morrison, Coleman. (1961). *The Torch Lighters: Tomorrow's Teachers of Reading* (Cambridge: Harvard University Graduate School of Education.

Berkun, M. M., Swanson, L. W., & Sawyer, D. M. (1966). An experiment on homogeneous grouping for reading in elementary classes. *Journal of Educational Research, 59*, 413-414.

Berwick, Carly. (2016). "Ban school suspensions!" *The Week,* Aug 9.

Bremer, N. (1958). First grade achievement under different plans of grouping. *Elementary English, 35*, 324-326.

Bifulco, Robert, and Helen F. Ladd. (2005). "Institutional change and coproduction of public services: The effect of charter schools on parental involvement." *Journal of Public Administration Research and Theory* 16.4: 553-576.

Booker, Kevin, and Brian Gill. (2012). "School Competition and Student Outcomes," in Ladd, Helen F., and Edward B. Fiske, eds. *Handbook of research in education finance and policy*. Routledge.

Buddin, Richard, and Ron Zimmer. (2005). "Is Charter School Competition in California Improving the Performance of Traditional Public Schools?" Santa Monica, CA: RAND Corporation, WR–297–EDU, 2005. Online only: http://www.rand.org/publications/WR/WR297/.

Cerrone, K. M. (1999). The Gun-Free Schools Act of 1994: Zero Tolerance Takes Aim at Procedural Due Process. *Pace Law Review*, 20, 131.

Charis, Kimberly, Losen, D. J. (2017). Fifth Indicator: School Climate and Student Discipline. *National Association of State Boards of Education*, Policy Update 24(4).

Christle, C. A., Jolivette, K., & Nelson, C. M. (2005). Breaking the school to prison pipeline: Identifying school risk and protective factors for youth delinquency. *Exceptionality*, *13*(2), 69-88.

Cline-Thomas and David Chang. (2017). Should the Philly School District End Suspensions for Elementary School Students? NBC10, May 3.

Collins, C. A., & Gan, L. (2013). *Does sorting students improve scores? An analysis of class composition* (No. w18848). National Bureau of Economic Research.

Cordes, Sarah. (2016). *In Pursuit of the Common Good: The Spillover Effects of Charter Schools on Public School Students in New York City*. Working paper, Temple University.

Courtis. S. A. (1925). Ability-grouping in Detroit schools. In G. M. Whipple (Ed.), *The ability grouping of pupils*, 35th Yearbook of the National Society for the Study of Education (Part I, pp. 44-47). Bloomington, IL: Public School Publishing.

Cremata, Edward J., and Margaret E. Raymond. (2014). "The competitive effects of charter schools: Evidence from the District of Columbia." *Association for Education Finance and Policy Conference Working Paper*.

Crone, D. A., Hawken, L. S., & Horner, R. H. (2010). *Responding to problem behavior in schools: The behavior education program*. Guilford Press.

d'Hombres, B., Elia, L., & Weber, A. (2013). Multivariate analysis of the effect of income inequality on health, social capital, and happiness. *JRC Econometrics and Applied Statistics Unit, Institute for the Protection and Security of the Citizen (IPSC), European Commission (version: April 2013)*.

D'Orio Wayne. (2018). Is School Discipline Reform Moving Too Fast? *The Atlantic*, Jan 11.

Dabla-Norris, M. E., Kochhar, M. K., Suphaphiphat, M. N., Ricka, M. F., & Tsounta, E. (2015). *Causes and consequences of income inequality: a global perspective*. International Monetary Fund.

Duflo, E., Dupas, P., & Kremera, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *The American Economic Review*, *101*(5), 1739-1774.

Dunning, T. (2012). *Natural experiments in the social sciences: a design-based approach*. Cambridge University Press.

Eden, M. (2017). School discipline reform and disorder: Evidence from New York City Public Schools, 2012-16. *The Education Digest*, 83(1), 22.

Eden, M. (2018). Teachers Nationwide Say Obama's Discipline 'Reform' Put Them in Danger. So Why Are the Unions Fighting DeVos on Repeal? *The 74*, Apr 2.

Ekstrom, R. B. (1961). Experimental studies of homogeneous grouping: A critical review. *School Review*, 69, 216-226.

Epple, Dennis, Richard Romano, and Ron Zimmer. (2015). *Charter schools: A survey of research on their characteristics and effectiveness*. No. w21256. National Bureau of Economic Research.

Frey, Susan. (2015). Oakland ends suspensions for willful defiance, funds restorative justice. *EdSource*, May 14.

Galindo, C., & Sheldon, S. B. (2012). School and home connections and children's kindergarten achievement gains: The mediating role of family involvement. *Early Childhood Research Quarterly*, *27*(1), 90-103.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* (Vol. 1). New York, NY, USA: Cambridge University Press.

Gerring, John. (2012). *Social Science Methodology: A Unified Framework*, 2nd edition. Cambridge University Press.

Grissom, J. A. (2012). Is discord detrimental? Using institutional variation to identify the impact of public governing board conflict on outcomes. *Journal of Public Administration Research and Theory*, *24*(2), 289-315.

Goldhaber, D. D., & Hansen, M. (2008). *Is it Just a Bad Class?: Assessing the Stability of Measured Teacher Performance*. Center on Reinventing Public Education.

Hanushek, E. A., Ruhose, J., & Woessmann, L. (2015). Human capital quality and aggregate income differences: Development accounting for US states.

Hanushek, E. A., & Woessmann, L. (2010). *The economics of international differences in educational achievement* (No. w15949). National Bureau of Economic Research.

Harris, D. N., & Herrington, C. D. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education*, *112*(2), 209-238.

Hashim, A. K., Strunk, K. O., & Dhaliwal, T. K. (2018). Justice for All? Suspension Bans and Restorative Justice Programs in the Los Angeles Unified School District. *Peabody Journal of Education*, (just-accepted).

Hughes, C., Warren, P. Y., Stewart, E. A., Tomaskovic-Devey, D., Mears, D. P. (2017). Racial Threat, Intergroup Contact, and School Punishment. *Journal of Research in Crime and Deliquency*, 54(5) 583–616.

Hupe, P., & Hill, M. (2007). Street-Level bureaucracy and public accountability. *Public Administration*, 85(2), 279-299.

Imberman, Scott A. (2011). The effect of charter schools on achievement and behavior of public school students. *Journal of Public Economics* 95.7: 850-863.

Jackson, C. K., Johnson, R., & Persico, C. (2014). The effect of school finance reforms on the distribution of spending, academic achievement, and adult outcomes (No. w20118). *National Bureau of Economic Research*.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of labor Economics*, *26*(1), 101-136.

Jinnai, Yusuke. (2013). The Impact of Charter Schools' Entry on Traditional Public Schools: New Evidence from North Carolina. *Rochester, NY: University of Rochester, Job Market Paper*.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. Bill & Melinda Gates Foundation.

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education, 79*(4), 54-78.

Kogan, V., Lavertu, S., & Peskowitz, Z. (2017). Direct Democracy and Administrative Disruption. *Journal of Public Administration Research and Theory*. 27(3): 381-399.

Kogan, V., Lavertu, S., & Peskowitz, Z. (2016a). Do School Report Cards Produce Accountability Through the Ballot Box? *Journal of Policy Analysis and Management*. 35(3): 639-661.

Kogan, V., Lavertu, S., & Peskowitz, Z. (2016b). Performance federalism and local democracy: Theory and evidence from school tax referenda. *American Journal of Political Science*, *60*(2), 418-435.

Kulik, J. A. (1992). An Analysis of the Research on Ability Grouping: Historical and Contemporary Perspectives. Research-Based Decision Making Series.

Kulik, J. A., & Kulik, C. L. C. (1992). Meta-analytic findings on grouping programs. *Gifted child quarterly*, *36*(2), 73-77.

Liaupsin, C. J., Umbreit, J., Ferro, J. B., Urso, A., & Upreti, G. (2006). Improving academic engagement through systematic, function-based intervention. *Education and Treatment of Children*, 573-591.

Lindstrom, Natasha. (2017). Pittsburgh Public Schools examines suspension practices. *TribLive*, June 22.

Lipsky, M. (1980). *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. New York: Russell Sage Foundation.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47-67.

Lorentzen, I. J. (2013). *The relationship between school board governance behaviors and student achievement scores* (Doctoral dissertation, University of Montana).

Losen, Daniel. (2011). Discipline Policies, Successful Schools, and Racial Justice. UCLA: The Civil Rights Project / Proyecto Derechos Civiles.

Losen, Daniel J.; Hodson, Cheri L.; Keith II, Michael A.; Morrison, Katrina; & Belway, Shakti. (2015). Are We Closing the School Discipline Gap?. *K-12 Racial Disparities in School Discipline*. UCLA: The Civil Rights Project / Proyecto Derechos Civiles.

Losen, Daniel J.; Keith II, Michael A.; Hodson, Cheri L.; Martinez, Tia E.; & Belway, Shakti. (2015). Closing the School Discipline Gap in California: Signs of Progress. UCLA: The Civil Rights Project / Proyecto Derechos Civiles.

Loveless, T. (2013). The 2013 Brown Center Report on American Education: How Well are American students learning? Washington, DC: The Brown Center on Education Policy. *The Brookings Institution*.

Luiselli, J. K., Putnam, R. F., Handler, M. W., & Feinberg, A. B. (2005). Whole-school positive behaviour support: effects on student discipline problems and academic performance. *Educational Psychology*, *25*(2-3), 183-198.

McFarland, D. A. (2001). Student resistance: How the formal and informal organization of classrooms facilitate everyday forms of student defiance. *American Journal of Sociology*, *107*(3), 612-678.

McPartland, James M., Coldiron, J. Robert, & Braddock II, Jomills H. (1987). School Structures and Classroom Practices in Elementary, Middle, and Secondary Schools, Report No. 14, Baltimore: The Johns Hopkins University.

Miles, C. C. (1954). Gifted children. In L. Carmichael (Ed.), *Manual of child psychology* (pp. 984-1063). New York: John Wiley & Sons.

Miller, W. S., & Otto, H. J. (1930). Analysis of experimental studies in homogeneous grouping. *Journal of Educational Research*, 21, 95-102.

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

Moses, P. J. (1966). A study of the effects of inter-class grouping on achievement in reading. *Dissertation Abstracts, 26,* 4342. (University Microfilms No. 66-741)

Nichols, N. (1969). Interclass grouping for reading instruction. *Educational Leadership Research Supplement, 26*, 588-592.

Nisar, Hiren. (2012). "Heterogeneous competitive effects of charter schools in Milwaukee." *Draft, Abt Associates Inc.*

Otto, H. J. (1941). Elementary education - II. Organization and administration. In W. S. Monroe (Ed.), *Encyclopedia of Educational Research* (1st ed., pp. 428-446). New York: Macmillan.

Penuel, W. R., Briggs, D. C., Davidson, K. L., Herlihy, C., Sherer, D., Hill, H. C., ... & Allen, A. R. (2017). How School and District Leaders Access, Perceive, and Use Research. *AERA Open*, *3*(2).

Perry, B. L., & Morris, E. W. (2014). Suspending progress: Collateral consequences of exclusionary punishment in public schools. *American Sociological Review*, *79*(6), 1067-1087.

Putnam, R. P., Horner, R. H., & Algozzine, R. (2006). Academic achievement and the implementation of school-wide behavior support. *Positive Behavioral Interventions and Supports Newsletter*, *3*(1), 1-6.

Rausch, M. K., & Skiba, R. (2004). Disproportionality in School Discipline among Minority Students in Indiana: Description and Analysis. Children Left Behind Policy Briefs. Supplementary Analysis 2-A. *Center for Evaluation and Education Policy, Indiana University*.

Reardon, S. F. (2013). The widening income achievement gap. *Educational Leadership*, *70*(8), 10-16.

Reid-Cleveland, Keith. (2017). Texas Governor Greg Abbott signs bill to ban out-of-school suspensions for young students. *Black Youth Project*, June 13.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education*, *4*(4), 537-571.

Rumberger, R. W., & Losen, D. J. (2017). The Hidden Costs of California's Harsh School Discipline: And the Localized Economic Benefits from Suspending Fewer High School Students. *Civil Rights Project-Proyecto Derechos Civiles*.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far?. In *Handbook of the Economics of Education* (Vol. 3, pp. 249-277). Elsevier.

Sass, Tim R. (2006). "Charter Schools and Student Achievement in Florida." *Education Finance and Policy* 1: 91-122.

Shipan, C. R., & Volden, C. (2008). The mechanisms of policy diffusion. *American Journal of Political Science*, 52(4), 840-857.

Shipan, C. R., & Volden, C. (2012). Policy diffusion: Seven lessons for scholars and practitioners. *Public Administration Review*, 72(6), 788-796.

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of educational research*, *57*(3), 293-336.

Skiba, R. J., & Knesting, K. (2001). Zero tolerance, zero evidence: An analysis of school disciplinary practice. *New Directions for Student Leadership*, *2001*(92), 17-43.

Skiba, R. J., & Sprague, J. (2008). Without Suspensions. *Educational Leadership*.

Steenbergen-Hu, S., Makel, M. C., & Olszewski-Kubilius, P. (2016). What one hundred years of research says about the effects of ability grouping and acceleration on K–12 students' academic achievement: Findings of two second-order meta-analyses. *Review of Educational Research*, *86*(4), 849-899.

Steinberg, M. P., & Lacoe, J. (2017). The Academic and Behavioral Consequences of Discipline Policy Reform: Evidence from Philadelphia. Mathematica Policy Research.

Tobin, J. F. (1966). *An eight year study of classes grouped within grade levels on the basis of reading ability.* Unpublished doctoral dissertation, Boston University. (University Microfilms No. 66-345)

Volden, C., Ting, M. M., & Carpenter, D. P. (2008). A formal model of learning and policy diffusion. *American Political Science Review*, 102(3), 319-332.

Wald, J., & Losen, D. J. (2003). Defining and redirecting a school-to-prison pipeline. *New Directions for Student Leadership*, 2003(99), 9-15.

Wallace Jr, J. M., Goodkind, S., Wallace, C. M., & Bachman, J. G. (2008). Racial, ethnic, and gender differences in school discipline among US high school students: 1991-2005. *The Negro educational review*, *59*(1-2), 47.

Watanabe, Teresa. (2013a). LAUSD board could ban suspensions for 'willful defiance.' *Los Angeles Times*, May 12.

Watanabe, Teresa. (2013b). L.A. Unified bans suspension for 'willful defiance.' *Los Angeles Times*, May 14.

Watanabe, Teresa. (2013c). L.A. schools will no longer suspend a student for being defiant. *Los Angeles Times*, May 15.

Watanabe, Teresa. (2014). L.A. Unified suspension rates fall but some question figures' accuracy. *Los Angeles Times*, May 31.

Watanabe, Teresa. (2015). Why some LAUSD teachers are balking at a new approach to discipline problems. *Los Angeles Times*, Nov 07.

Whitehurst, Grover J., Matt Chingos, and Michael R. Gallaher. (2015). "School Districts and Student Achievement," *Education Finance and Policy* 10(3): 378–398.

Whitehurst, Grover J., Matt Chingos, and Michael R. Gallaher. (2013). "Do School Districts Matter?" *Brookings Institution*.

Whitehurst, Grover J., Matt Chingos, and Katharine M. Lindquist. (2014). "School Superintendents: Vital or Irrelevant?" *Brookings Institution*.

Winters, Marcus A. (2012). "Measuring the effect of charter schools on public school student achievement in an urban environment: Evidence from New York City." *Economics of Education review* 31.2: 293-301.

Word, E. (1990). Student/Teacher Achievement Ratio (STAR) Tennessee's K-3 Class Size Study. Final Summary Report 1985-1990.

Zimmer, Ron, et al. (2009). *Charter schools in eight states: Effects on achievement, attainment, integration, and competition*. Vol. 869. Rand Corporation.

CURRICULUM VITAE