UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Essays on Finance, Accountability, and Publication Bias in the Economics of Education.

Permalink
https://escholarship.org/uc/item/9d6748sz

Author
Tanner, Patrick Sean

Publication Date
2018

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Essays on Finance, Accountability, and Publication Bias in the Economics of Education.

By

Patrick Sean Tanner

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Public Policy

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Rucker C. Johnson, Chair
Professor Jesse Rothstein
Professor Hilary Hoynes

Summer 2018

Abstract

Essays on Finance, Accountability, and Publication Bias in the Economics of Education.

By

Patrick Sean Tanner

Doctor of Philosophy in Public Policy

University of California, Berkeley

Associate Professor Rucker C. Johnson, Chair

School quality exerts a significant impact on children well beyond their school-age years. High quality primary and secondary schools improve educational attainment, reduce criminal behavior, and increase labor market performance. Decades of federal and state policy efforts have sought to ensure broad access to effective schools. Two of the most consequential reform efforts of the past thirty years, school accountability and financial adequacy reforms, have been moderately successful at raising average performance on standardized tests of literacy and numeracy, yet critics contend that these gains mask serious limitations. In the pages that follow, I address salient concerns of each policy by analyzing the impact of No Child Left Behind on social emotional development and the impact of California's recent school finance reform, the Local Control Funding Formula. I conclude with an assessment of publication bias in policy-relevant scholarship.

Dedication

This work is dedicated to my wife Lindsay and son Sullivan. Having a graduate student for a husband and father is not a condition I would wish on anyone. The stress, late nights, and parsimonious standard of living could not have been easy to endure. You both nevertheless gave unconditional love, which I will spend the rest of my life gladly repaying. Thank you.

Table of Contents

Introduction 1

The Impact of No Child Left Behind on Non-cognitive Skills 2

Section I- Introduction 2

Section II- No Child Left Behind 3

School Accountability and Non-Cognitive Skills 6

Section III- Estimation Strategy 7

Section IV- ECLS-K 1998 Data 10

Section V- Results 12

Section VI- Discussion 14

Money and Freedom: The Impact of California's School Finance Reform on Academic Achievement and the Composition of District Spending 40

Section I- Introduction 40

Section II – The Local Control Funding Formula 42

Section III- School Finance Reforms 46

Section IV- Empirical Strategy 50

Section V- Data 55

Section VI- Results 57

Section VII- Discussion 60

Evidence of False Positives in Research Clearinghouses and Influential Journals: An Application of P-Curve to Policy Research 84

Section I - Introduction 84

Section II - False Positive Social Science 85

Section III - Assessing the Evidential Value in a Body of Literature 86

Section IV- Study Design 87

Section V- Data Collection 88

Section VI - Results 90

Section VII- Discussion 91

Conclusion 95

References 97

Introduction

The quality of children's education has marked effects on their future financial and social well-being. In helping to shape such quality, education policy can have a significant role in altering patterns of human capital accumulation and reducing income inequality in the United States. Of the numerous policies that have been enacted over the last thirty years, two of the most prominent have been increasing the financial resources of impoverished school districts and leveling sanctions on schools that fail to meet certain performance criteria. Despite the fact that these efforts, school finance reform and school accountability, respectively, have had modest success in bolstering student performance on standardized tests of literacy and numeracy, questions remain about the overall efficacy of each policy. While school finance reforms have led to meaningful increases in student performance generally, critics argue that the new revenues are insufficiently targeted to the students most in need and have also tended to increase the state's role in dictating how the money can be spent. As for school accountability, there is a concern that the focus on literacy and numeracy has come at the expense of non-cognitive skills – a set of social emotional behaviors and traits that exert more influence on adult well-being and labor market success than do standardized tests of academic skills. At the same time as these debates continue, the overall credibility of social science research has received renewed attention with a series of findings that have raised the specter of widespread publication bias. Most troubling is a practice colloquially known as "p-hacking" – analyzing the data in various ways until a significant result is found. If p-hacking pervades policy research, then the literature that informs school finance and accountability policies may be inaccurate and, consequently, the debates that surround them untethered from empirical reality.

In this work, I add to these three strands of the economics of education literature. In the first chapter, I test whether or not the nation's most significant accountability policy, the No Child Left Behind Act of 2001, has harmed non-cognitive skills. Based on an index of child, parent, and teacher responses to surveys within the Early Childhood Longitudinal Study of Kindergarten 1998, I find little effect of NCLB on non-cognitive skill development. In a second chapter, co-author Rucker Johnson and I assess the degree to which student performance has increased due to California's recent school finance reform, the Local Control Funding Formula (LCFF), a highly progressive system of revenue allocation based on student-level disadvantage that is also largely free of restrictions on expenditures. LCFF has enhanced literacy, numeracy, and high school graduation rates, especially for disadvantaged children. In the final chapter, I use a novel statistical technique to test for various forms of publication bias. Using statistical data from published literature in three major policy-oriented journals and two federal research clearinghouses, I find no evidence of p-hacking or other publication bias – suggesting that policy research is not unduly plagued by such behavior. I conclude with recommendations for future research on all three fronts.

The Impact of No Child Left Behind on Non-cognitive Skills

Section I- Introduction

School accountability polices are among the more recent waves of school reform in the United States. Focusing on the lack of performance incentives in public K-12 school systems, supporters of these policies believed that accountability would enhance student achievement if state governments illuminated disparities in achievement and levied rewards or sanctions on schools based upon explicit, mandated measures of student performance (Hanushek & Raymond, 2001, pp. 368–369). These "high-stakes" tests incentivized schools to focus narrowly on the tested subjects, and especially on the performance of groups of students whose test scores would have a meaningful impact on the schools' accountability rating (Manna, 2011, Chapters 2, 6). An abiding concern for critics of accountability is how the focus on standardized tests of literacy and numeracy impacts domains of child development that are known to be key determinants of adult success but are not included in the accountability testing regime (Duckworth, Quinn, & Tsukayama, 2012).[1]

By the time the No Child Left Behind Act (NCLB) mandated such systems for all public schools nation-wide, half of states had already implemented similar systems (Dee & Jacob, 2011, p. 423). Several studies have attempted to assess the impact of this federal mandate by using the states that had already adopted accountability ("early adopters") to construct a counterfactual for states that were newly compelled to do so ("late adopters") (Dee & Jacob, 2011; Lee & Reeves, 2012; Reardon, Greenberg, Kalogrides, Shores, & Valentino, 2013; Wong, Cook, & Steiner, 2011). This body of work has produced somewhat mixed results, with most research showing positive effects on math and little effect on reading.

My research contributes to this literature by providing the first national estimates of the impact of NCLB on non-cognitive skills. The study analyzes data from the Early Childhood Longitudinal Study of Kindergarten 1998 (ECLS-K), a nationally-representative panel data set of 21,000 children who attended kindergarten in 1998 and have been followed through eighth grade. The data contain detailed measures of both cognitive and non-cognitive skills at school entry and in each survey period. I compare the developmental trajectories of children residing in early adopter states to those living in late adopter states. I employ a propensity score weighting technique so that these two groups of children are balanced with respect to race, socioeconomic status, early childhood education, and cognitive achievement scores at the time they entered kindergarten. The identification of NCLB effects comes from the differentially binding nature of the law's accountability mandate between these two groups of students. Only students in the late-adopter states (those without prior accountability regimes) experienced their first exposure to accountability at the onset of NCLB- which took effect in fourth grade for the birth-cohort in the ECLS-K sample. By tracking the ECLS-K students as they progress through elementary school, this design accounts for pre-treatment (NCLB) differences in student achievement growth between groups of students who had the same cognitive achievement and socioeconomic profile at the onset of their public schooling career. The impact of NCLB is estimated as the change in non-cognitive skills, relative to pre-NCLB trends, between children from the early- and late-adopter states, evaluated at the same ages. This comparative interrupted time series with student fixed-effects is robust to linear and quadratic pre-NCLB trends as well as the inclusion of contemporaneous changes in paternal employment, socioeconomic status, and family structure.

[1] For a representative example of such criticism in the popular press, see Gewertz (2003) and Darling-Hammond (2007).

As I elaborate upon in the estimation strategy section, this research design is similar to that used in prior work on NCLB, but with a key distinction: this analysis uses data from a single cohort and compares growth in cognitive and non-cognitive domains between two groups of students that were differentially exposes to the NCLB mandate as they progress through elementary school, whereas the prior literature analyzed trends in average standardized test scores across successive cohorts of students.

In a departure from prior work, I examine broader measures of non-cognitive skills and traits rather than focusing only aberrant behaviors such as truancy and delinquency. Specifically, the nine measures are teachers' assessment of students' self-control, approaches to learning, peer relationships, and externalizing and internalizing problem behaviors, and parents' assessment of independence, ability to solve problems, attention span, and overall behavior. Furthermore, I consider both measures of cognitive and non-cognitive development to provide a more comprehensive assessment of the impacts of school accountability policies on broader dimensions of student performance.

To preview the results, I find little effect of NCLB on any of these measures of non-cognitive skills through fifth grade on average nationally. However, as the law specifically mandated that test scores be disaggregated by socioeconomic sub-groups, national averages obscure potentially meaningful heterogeneity in NCLB's effects. When analyzing the data by socioeconomic sub-groups, I find mixed results. Teachers' ratings of several non-cognitive skills increased for underrepresented minorities (African-American and Hispanic), but decreased for white and Asian students. The parents' assessments are mostly negative or indistinguishable from zero for all groups on average, and for each of these subgroups. Aggregate indices of these measures support this general pattern. The data allows changes in teachers' child development training and emphasis on non-cognitive assessment to be ruled out as potential mechanisms.

The remainder of the chapter is organized as follows. Section II begins with a description of No Child Left Behind and the prior work analyzing its impacts on cognitive achievement, then briefly summarizes the literature on non-cognitive skills and their relation to educational achievement and labor market performance. Section III provides a detailed discussion of the identification strategy followed by a description of the ECLS-K data in Section IV. Section V presents the main regression results, which includes estimates of the impact of NCLB on each of the nine measures overall and for white, Asian, African-American, Hispanic, free-lunch, and limited English proficiency students. Two tests of potential mechanisms are examined, along with a discussion of potential reasons for the heterogeneity in results. The final section concludes with summary discussion, policy implications, and fruitful directions for future research.

Section II- No Child Left Behind

Accountability systems shifted the focus to short-run educational outcomes—particularly, math and reading standardized test scores -- rather than school inputs, fundamentally altering the metrics by which school quality is judged. The central motivation of these policies is the perceived lack of progress in American students' achievement since the early 1970s (Hanushek & Raymond, 2001, pp. 366–368). Since the mid-1950s, per-pupil spending has increased markedly as a percentage of GDP (Springer, Houck, & Guthrie, 2015, p. 4). Consequently, traditional markers of school quality had been increasing, such as teacher-pupil ratios, the percentage of teachers with master's degrees, and teacher tenure. From 1960 to 2000, teacher salaries more than tripled (in constant dollars) while the proportion of teachers with at least a master's degree doubled and average tenure increased 30% (Hanushek, 2003, p. F68). However,

aggregate measures of student progress have failed to increase commensurately with these significant increases in spending, leading many to question the wisdom of increased resources without concurrent incentives to use resources wisely (McGuinn, 2006, pp. 134–135).

To address this concern, accountability policies measure student performance, publish school performance data, and subject schools to rewards and sanctions based on this data (Figlio & Ladd, 2015). These performance measures are standardized and disaggregated by demographic characteristics such as race/ethnicity and socioeconomic status (SES), so that the educational deficiencies of disadvantaged students are not masked by school-wide averages or inflated grades (Manna, 2011, pp. 121–122). Operationally, this means that schools are primarily judged on the basis of their students' performance on state-administered tests of literacy and numeracy.

Accountability began as state-level reforms, with Illinois being the first to adopt such a system in 1992. By the 2002-2003 school year, when NCLB required schools in all states to implement accountability plans, roughly half of states had already adopted such plans (Dee & Jacob, 2011, p. 423). Several key elements of the law bear directly on this analysis. First, tests were standardized by grade and state, so that all students were assessed by the same tests of literacy and numeracy within each grade and state. Second, each test had a score above which students were considered to have adequately mastered the subject by grade level- the "proficiency threshold." Third, school-level accountability scores were calculated for each demographic sub-group as the percentage of students in that group who scored above the proficiency threshold. These pass rates were then used to determine which schools were not making "adequate yearly progress" (AYP) toward 100% of students scoring above the proficiency threshold by 2014. An escalating series of punitive measures would be levied on schools that continued to fail to make AYP year after year, ranging from a requirement to offer public school choice to massive staff layoffs and outright closure (Manna, 2011, pp. 26–29). In order to underscore the severity of this new regime and distance the law from the ineffectual regulations of prior administrations, Secretary of Education Roderick Paige warned states that they were expected to heed the letter of the law and that the Bush administration would not tolerate noncompliance (McGuinn, 2006, p. 183).

A large body of research has been devoted to assessing the impact of accountability, and NCLB in particular, on student achievement and school resources (see Figlio and Ladd (2015, pp. 204–207) for a concise overview). The literature summarized below focuses on studies that analyze the impact of NCLB rather than the impact of accountability in early adopter states. As is made clear in the estimation strategy section and the following literature review, studies of the impact of NCLB are designed so that the precise treatment is the federal accountability *mandate*, not aspects of accountability systems themselves or various other features of NCLB. Due to jurisdictional and political tensions, NCLB allowed states substantial freedom in crafting their own accountability systems (McGuinn, 2006, pp. 177–179; Vinovskis, 2009, pp. 158–170). This state freedom led to consequential policy variation in how much pressure was put on schools across states, primarily through variation in proficiency thresholds and ability of schools to discard certain students' tests (Davidson, Reback, Rockoff, & Schwartz, 2013). Because this variation is endogenous to student achievement, valid causal inference is limited to the effect of the NCLB mandate rather than states' choices in crafting systems. Moreover, NCLB contained numerous provisions in addition to the accountability mandate, such as the requirement that

teachers be "highly qualified"[2] (Manna, 2011, pp. 29–30). These provisions affected both states that had and had not adopted accountability prior to the law, rendering the identification strategy favored here and in the prior literature inappropriate for assessing those effects.

Prior research on NCLB has found mostly positive effects on math and weak to null effects on reading. Of the five separate research papers summarized here, four have used variants of a comparative interrupted times series (CITS) design on data from the National Assessment of Educational Progress (NAEP). The control group is typically those states that had adopted consequential accountability prior to the onset of the national mandate (the early adopters) and the treated group is the states that were compelled to enact new systems in the 2002-2003 school year (the late adopters). The CITS design leverages the variation in the timing at which children were exposed to accountability, using a causal framework similar to that of a difference-in-difference model, but allowing pre-treatment differences in the (usually) linear trends in outcomes across treatment and control groups. The ability to control for differences in pre-treatment trends rather than merely pre-treatment levels is crucial in analyses of NCLB, where pre-2002 trends in NAEP scores are demonstrably different in early and late adopter states (Dee & Jacob, 2011, pp. 432–433). In the CITS models, the impact of NCLB is the difference in the difference in achievement trends (and levels) between the early adopter and late adopter states before and after the onset of the law in 2002.[3] This design has been used on NAEP data, which contains nationally-representative tests of literacy and numeracy given to successive cohorts of fourth and eighth grade students from 1990 to the present. As these scores are not used in accountability regimes and are thus "low-stakes," they are plausibly immune to the "teaching to the test" phenomenon (Jacob & Levitt, 2003) and so provide credible measures of students' cognitive achievement.

Using this strategy, Dee & Jacob (2011) find strong increases on fourth grade math scores (between .15 to .47 standard deviations), fewer and more moderate increases on eighth grade math scores (around .22 standard deviations), and little to no effects on 4th grade reading achievement. These results are stronger for underrepresented minority (URM) and lower-income students (Dee & Jacob, 2011, pp. 438–440). Wong, Cook, and Steiner (2011) further separate states into those that had relatively high vs. low proficiency thresholds. They find positive effects on fourth and eighth grade math overall (in the range of effects found by Dee and Jacob (2011)), with positive effects on reading limited to states that had high proficiency thresholds. Lee and Reeves (2012) use propensity score weights that place greater analytic weight on early adopter states that were similar to late adopters states on aggregate levels of high school attainment, poverty rates, Caucasian proportion of the population, state education revenue, SAT scores, and a political culture variable. They also include data on fidelity of implementation (data tracking capacity, funding, and stringency of proficiency thresholds) as explanatory variables. They find an effect on eighth grade math scores, but none on fourth grade math and reading scores. Finally, Reardon et al. (2013) find that NCLB narrowed achievement gaps (a small effect size of roughly 3%[4] decrease in the magnitude of the gap per year) between white

[2] Highly qualified meant, essentially, that teachers had formal training in and a manifest command of the subjects in which they taught.

[3] A full treatment of this model is provided in the estimation strategy section

[4] Reardon et al. (2013) use the V-statistic outlined in Reardon and Ho (2015), a measure of between-group disparity more appropriate to ordinal, non-Gaussian data such as the NAEP. The baseline V for the gaps was between .7 and .6 and the effect of NCLB per year ranged from -.02 to .02, depending on state characteristics.

and URM students only in states with larger pre-NCLB gaps, a greater degree of between-school segregation, and more accountability pressure focused on URM sub-groups.

Reback, Rockoff, and Schwartz (2014) use an alternate identification strategy to assess the impact of NCLB pressure for students in schools at risk of failing to meet proficiency thresholds. Using the same ECLS-K data employed in the present study, they analyze the cross-sectional variation in schools' risk of failure conditional on observable school and student characteristics. Their logic is that similar students face differing levels of accountability pressure due to state policy choices (ex. proficiency thresholds) and school-level demographics. They find positive effects of being near the failure boundary on reading scores, with positive but statistically insignificant effects on math and science tests.

Several analyses of changes in school resource allocation and pedagogical activities induced by NCLB support this set of findings of effects on achievement gains. Using the Schools and Staffing Survey (SASS), Dee and Jacob report that schools increased teacher salaries and hired more teachers with master's degrees, while teachers re-allocated their time toward reading and (to a lesser extent) math (2010, pp. 184–187). Further evidence from the SASS suggests that the onset of NCLB and the resulting reallocation of school resources was not associated with a decrease in teacher morale or increase in hours worked (at least as measured by survey responses in the SASS) (Grissom, Nicholson-Crotty, & Harrington, 2014).

On balance, these results provide evidence that NCLB shifted resources toward instruction in literacy and numeracy, increased performance on certain tests of those subjects, and did little to impact overall teacher morale.

School Accountability and Non-Cognitive Skills

In addition to the traditional focus on cognitive achievement, teachers also impact their students' development of non-cognitive skills, such as self-control, adaptability, and motivation. The category labeled "non-cognitive"[5] encompasses a broad range of attributes such as those related to concentration, social interactions, and judgement. In a school setting, they manifest themselves in children's ability to complete tasks, pay attention, and cooperate with teachers and peers, to name but a few. These skills are complementary to those traditionally labeled cognitive skills, such as numeracy and literacy, yet are not well captured by narrow aptitude or achievement tests. For example, in a study undertaken in direct response to NCLB, Duckworth, Quinn, and Tsukayama (2012) find that teacher and parent ratings of children's self-control in fourth grade are far better predictors of ninth grade G.P.A. than are fourth grade IQ tests. Conversely, fourth grade IQ is a far better predictor of ninth grade standardized achievement tests than is self-control. This suggests that a narrow focus on standardized testing will fail to account for the impact of schools and teachers on the elements of academic achievement more closely related to non-cognitive skills.

Non-cognitive skills are typically not in the large, state data sets that are used to monitor student achievement in the U.S. Nevertheless, the data that is available reveals that such non-cognitive skills predict acquisition of cognitive skills and are rewarded in the labor market independently of IQ and test scores (Almlund, Duckworth, Heckman, & Kautz, 2011). These skills are not stable over a lifetime, suggesting a role for the educational system in manipulating them. Heckman, Pinto, and Savelyev (2013) provide compelling evidence that the long-run gains from the Perry Preschool program are largely attributable to its impact on externalizing

[5] These skills have also been referred to as character, grit, social and emotion learning competencies, disposition, and temperament (Duckworth & Yeager, 2015, p. 238).

problems- aggressive, anti-social behavior. A similar pattern is found in children who participated in Project STAR, the canonical experiment of class size reduction on student achievement. The experimental impact of higher quality classrooms on test scores fades out in later grades, yet impacts on non-cognitive skills remain and are likely mechanisms through which the treatment affects long-run success (Chetty et al., 2011).

Adopting the methodology used in typical value-added models to control for observable characteristics and prior achievement, Jackson (2012) shows that teachers have meaningful impacts on their students' non-cognitive skills. Interestingly, teacher quality in regards to non-cognitive skills appears orthogonal to a measure of teacher quality based on gains in cognitive skills. In other words, a teacher that is above average in her effect on test scores is no more or less likely to improve her students' non-cognitive ability than an average (or below average) teacher. This has severe consequences for accountability policies whereby teachers and schools are rated only by their students' cognitive scores. There is little evidence of how teachers respond to those policies in terms of promoting cognitive vs. non-cognitive skills. Critics of accountability policies worry that the focus on preparation for high stakes testing hinders the acquisition of non-cognitive skills by crowding out pedagogical activities that encourage such skills in favor of those that increase scores on standardized achievement tests (Darling-Hammond, 2007; Duckworth et al., 2012).

Section III- Estimation Strategy

This chapter identifies the impact of NCLB's school accountability requirements by exploiting the differentially binding nature of the law between states that had already adopted accountability policies by 2001 (the control or "early adopter" group) and those states that were newly compelled by NCLB to adopt such policies (the treatment group). The preferred estimation strategy, a propensity-score weighted comparative interrupted time series, is robust to endogenous selection into treatment groups and both linear and non-linear differences in outcome dynamics due to heterogeneity in observed and unobserved determinants of students' non-cognitive skills. The basic intuition is similar to that of a difference-in-difference (DID) approach, but does not require the assumption of parallel pre-treatment time trends in the outcome. Additionally, the panel structure of ECLS-K allows the analysis of student achievement growth within the same cohort of students rather than changes in averages across successive cohorts of students as are analyzed in the NAEP studies. Though the research designs in both this analysis and the prior NAEP studies can be appropriately labeled comparative interrupted times series designs, the student fixed effect panel vs. pooled cohort model is a key distinction. The panel study design employed here has the benefit of avoiding threats to validity from endogenous selection in and out of the tested sample as well as increased precision from using student fixed-effects.[6] Analyses of the public school NAEP data are reliant on an assumption that schooling decisions (public vs. private) are exogenous to the onset of NCLB conditional on the demographic data available from NAEP. This is a strong assumption, one that is not necessary when tracking the same cohort of children across grades as ECLS-K does. However, this validity and precision come at a price- the divergent trends being captured in the CITS model are within-student growth rather than changes in between-student averages. The validity of the CITS design rests on correctly modeling pre-treatment divergence between early adopter and late adopter states. Modeling student growth over time presents challenges that are

[6] The increase in precision over pooled cohort models is conditional on similar sample sizes and objects of measurement.

likely not present in successive cohort models, where the assumption of stability or linearity in averages over time is easier to maintain. To deal with this challenge, I standardize the teacher ratings of non-cognitive outcomes by grade-level to net out any year to year changes in overall student growth. The raw measures from ECLS-K suggest that teachers are already informally standardizing their ratings, as there is very little year to year change in the central tendencies of all five constructs.

A simple DID strategy would model the impact of NCLB as the change in non-cognitive skills from third to fifth grade in the treated states minus the change in the control states. This would be operationalized as b_3 in equation 1.

$$Y_{ist} = b_0 + b_1 NCLB_t + b_2 Treat_s + b_3(NCLB_t \times Treat_s) + \varepsilon_{ist} \qquad (1)$$

Where Y_{ist} is the outcome of individual i in state s at time t, NCLB is a dummy variable equal to one in fifth grade and zero in third, Treat is a dummy variable equal to one if that state had not implemented school accountability prior to the onset of NCLB in 2002, and ε is a stochastic error term. One of the strongest identifying assumptions necessary for b_3 to be the causal impact of NCLB is that the students in treated and control states would have had parallel growth if not for the policy change. This is a strong assumption that is likely to be violated if the treated and control states differ in ways that affect the growth and levels of students' non-cognitive skills. As Figure 1 reveals, this is almost certainly true in the case of NCLB, where the early adopters are highly clustered in the South. As can be seen in the third row of table Table 9, teachers' ratings of students' externalizing problem behaviors and approaches to learning diverge between the two groups of states in the pre-NCLB period (the coefficients on Year*Treat are statistically significant and negative). A DID would incorrectly include a continuation of this pre-treatment trend in b_3.

This problem of non-parallel trends can be attenuated with a comparative interrupted time series (CITS) design, which explicitly controls for pre-treatment trends across treated and control groups. The basic linear strategy models the impact of the treatment as the change in level *and* slope across the treatment boundary in the treated states minus the change in level *and* slope in the control states.

A linear CITS model specification with individual panel data is below:

$$Y_{ist} = b_0 + b_1 YEAR_t + b_2 NCLB_t + b_3(Year_t \times NCLB_t) + b_4(T_s \times YEAR_t) + b_5(T_s \times NCLB_t) \qquad (2)$$

$$+ b_6(T_s \times Year_t \times NCLB_t) + \mu_i + \varepsilon_{ist}$$

Where Y_{ist} is the outcome of individual i in state s at time t, Year is the calendar year minus 1998, NCLB is a dummy variable equal to one at 2002 and after (when Year ≥ 4), T is a dummy variable equal to one if that state had not implemented school accountability prior to the onset of NCLB in 2002, μ is an individual fixed effect, and ε is a stochastic error term. The total effect of NCLB at any given year after its onset is $b_5 + b_6 \times t$, where t is the number of years since 2002. The total effect of NCLB at the final survey wave in 2007 is thus b_5 (the shift in level from 3rd to 5th grade) + $b_6 \times 3$ (the coefficient on the linear trend from 5th to 8th multiplied by 3 years). This strategy has the benefit of accounting for linear, non-parallel pre-treatment trends in the treated and control states. For instance, the states that adopted accountability policies may have also been more likely to attract and retain high quality teachers, who could have increased student test scores in addition to the gains from accountability. A basic CITS model allows for both time-invariant differences such as these as well as time-dependent differences so long as the

functional form (linearity) of the time-dependency does not change from the pre-NCLB era to the post-NCLB era. Similarly, the onset of NCLB is allowed to co-vary with secular changes in the determinants of student achievement. However, this covariance must be similar across treatment and control groups. This design has been used extensively in education research, particularly to estimate the impact of NCLB and accountability generally on student achievement (Jacob, 2005; Somers, Zhu, Jacob, & Bloom, 2012).

Heterogeneous, non-linear trends in cognitive achievement are an empirical regularity in many social experiments such as the Perry Preschool project (Heckman et al., 2013). In these experiments, treated children have initially higher cognitive achievement scores, but those scores decay and collapse to control group means around 3rd and 4th grades- the precise point at which the ECLS-K children experience the onset of NCLB. An estimation strategy that does not account for such differential early childhood education (ECE) would attribute the lack of fade out in those not exposed to ECE as an impact of a 3rd or 4th grade intervention rather than the fading out of ECE-related gains in those exposed to it. To account for potential differences such as these, I weight the early adopter group to fit the late adopter group by multiple baseline characteristics such as early childhood education, socioeconomic status, and cognitive assessments of literacy and numeracy. Specifically, I fit a saturated logistic regression of treatment status on binary indicators for Head Start attendance, non-Head Start preschool attendance, free/reduced price lunch and limited English proficiency status, ethnic category (African-American, Hispanic, White, other), and terciles of socioeconomic status (an index of parental education, income, and job categories) and cognitive scores on tests of literacy and numeracy taken upon entrance to kindergarten. These variables are completely interacted with one another to allow a saturated regression. The predicted treatment status ($\hat{t}$) is used to calculate a propensity score p wherein the late adopter group is assigned a value of 1 and the early adopter group is assigned $p = \hat{t}/(1-\hat{t})$. This weighted CITS design places more analytic weight on control group (early adopter) students with demographic and cognitive profiles that are most similar to the students in the treatment group (late adopter). Similar weighting strategies have been used for DID designs to overcome the problem of non-parallel trends (Abadie, 2005), Lee and Reeves' CITS analysis of NCLB (2012, pp. 214–215), and two CITS designs found to produce similar causal estimates to those from a regression discontinuity analysis (Somers et al., 2012) and a randomized field trial (St.Clair, Cook, & Hallberg, 2014) of two separate educational programs. As an addition guard against spurious causal inference, I control for contemporaneous changes in family structure, socioeconomic status, and paternal employment. Specifically, I include indicator variables for a father's presence in the home and fulltime employment status as well as a continuous measure of family socioeconomic status provided by ECLS-K.

The actual years observed in the ECLS-K data when both teacher and parent data on non-cognitive skills are available are years 1 (spring Kindergarten), 2 (spring 1st grade), 4 (spring 3rd grade), and 6 (spring 5th grade), so only the initial level shift can be estimated (the impact of ≈2 years of accountability). The full model is presented in equation 3.

$$Y_{ist} = b_0 + b_1 YEAR_t + b_2 NCLB_t + b_3(Year_t \times Treat_t) + b_4(NCLB_s \times Treat_t) + \sum_{j=5}^{k} \beta_j X_{ist} + \mu_i + \varepsilon_{ist} \quad (3)$$

Where Y_{ist} is the outcome of individual i in state s at time t, Year is the calendar year minus 1998, NCLB is a dummy variable equal to one at 2002 and after (when Year $\geq$ 4), Treat is a dummy variable equal to one if that state had not implemented school accountability prior to the onset of NCLB in 2002, X_{ist} is a vector of socioeconomic variables for individual i in state s at time t, μ is an individual fixed effect, and ε is a stochastic error term. The effect of NCLB is b_4, which represents the level shift in non-cognitive skills after approximately two academic years under NCLB.

In a less parametric model, I create grade-level dummy variables and estimate saturated regressions of the outcomes on grade-level dummies and their interactions with the treatment identifier. This has the advantage of being completely transparent and non-parametric, however it does not readily yield an estimate of the impact of NCLB under observed differences in the pre-NLCB outcome trends.

Section IV- ECLS-K 1998 Data

This analysis uses data from the Early Childhood Longitudinal Study of Kindergarten 1998 (ECLS-K), a nationally representative panel study that began surveying children in the fall of their kindergarten year in 1998 and concluded with a survey of the same cohort in the spring of their eighth grade year in 2007. Seven rounds of data were collected, six of which are used in this analysis.[7] The ECLS-K data has several features that are advantageous for an analysis of the impact of NCLB. First, the timing of the study is such that the children's elementary school career spanned the onset of NCLB: K-3 prior to NCLB and 4-8 under NCLB. Second, unlike alternate data sets of student achievement such as NAEP or state-level achievement tests, ECLS-K contains non-cognitive assessment data on the same cohort of children as they develop through elementary school. Third, ECLS-K contains a rich set of data on students, teachers, schools, and parents- both upon entrance to kindergarten and throughout the students' elementary school career.

The analysis sample includes data on approximately 10,000 children sampled each round.[8] Table 1 contains descriptive data across early and late adopter states in the fall of 1998 – the children's first semester of kindergarten and the first survey period. The states compelled by NCLB to adopt accountability are considerably less ethnically diverse, have a larger share of children attending Head Start, and have fewer children who receive free or reduced price lunch than the states that adopted accountability prior to NCLB. Additionally, schools in the late adopter states are less likely to have kindergarten teachers who completed post-baccalaureate training and are far more likely to receive Title 1 funding. These differences indicate likely differences in additional, unobserved determinants of student achievement and non-cognitive growth, suggesting that the early adopters are a poor counterfactual for the late adopter states in a simple, cross-sectional comparison design.

This analysis primarily focuses on two sets of non-cognitive domains that are measured from kindergarten to fifth and eighth grade in the ECLS-K surveys: the Social Rating Scale (SRS) from the teacher questionnaires (K to 5th) and parents' rating of their own children's behavior (K to 8th). Additional analyses are conducted on the children's self-reported non-cognitive skills. Focusing on teachers' and parents' ratings of the children as they mature

[7] The fall 1st grade round was dropped, leaving fall and spring kindergarten, spring 1st, spring 3rd, spring 5th, and spring 8th in the analysis sample. The fall 1st grade survey round was conducted only on a small sub-sample.
[8] The complete dataset contains information on 21,409 children. Attrition, missing data, and private school attendance whittle the analysis sample down to roughly 10,000.

through elementary school is advantageous for two reasons. First, the parent and teacher surveys are capturing distinct elements of children's non-cognitive growth in school and home settings. This allows tests of whether changes in classroom behaviors spill over into the children's behavior elsewhere. The second advantage of the teacher and parent surveys is that they are more likely to be reliable in the early grades than children's self-report (Junttila, Voeten, Kaukiainen, & Vauras, 2006). As the children mature and their responses become more reliable, ECLS-K begins to collect self-reported socio-emotional data. In the final year, eighth grade for most of the sample, the SRS is no longer collected from teachers as the children are not with a single teacher throughout the school day as they are in earlier years. The parent surveys of non-cognitive skills are given throughout the entire panel periods- kindergarten through eighth grade. Table 4 contains the details of when each construct is measured and from whom.

The four measures from the parents survey, independence, ability to pay attention, problem solving skills, and overall behavior are each collected from a single survey item on a 1 (is better than peers) to 4 (is much worse than peers) scale. Table 2 contains the distribution of responses in each category for all four variables over the entire survey period (kindergarten through eighth grade). For each item, more than half of parents report their child as being average ("as well as peers") for their age and at least a quarter reported their child as being better than average. Few parents labeled their child as below average and fewer still labeled their child as far below average. Given this distribution of responses, these ordinal scales have been transformed into dummy variables for this analysis, with "better than peers" coded as 1 and all other responses coded as 0. Doing so focuses the analysis on a readily interpretable boundary (better/not better than average), one that is also likely to yield more informative results that apply to a larger share of children in the sample.

The SRS from the teacher survey is composed of five constructs measured on a continuous one to four scale that is the combination of multiple, ordinal items in the questionnaires. The constructs are self-control, interpersonal skills, approaches to learning, internalizing problem behaviors (ex. anxiety), and externalizing problem behaviors (ex. class disruptions). The first three capture positive aspects of child development (higher values are normatively better) and the latter two capture problematic behavior (higher values are normatively worse). The child surveys contain self-description questionnaires (SDQs) on externalizing and internalizing problem behaviors as well as interpersonal relationship skills in the third and fifth grades (on either side of the onset of NCLB). Like the SRS, these measures are on a continuous one to four scale that is the combination of multiple items in the questionnaires. For each item from which the five SRS variables and three SQD variables are constructed, the respondent is asked how often particular behavior is observed (teachers) or experienced (students): never, sometimes, often, very often. The scale scores for the SRS and SDQ are the mean for all items in each scale. Table 3 contains summary statistics from all five constructs in the SRS and all three in the SDQ across all survey periods. Overall, the teachers and children reported relatively higher levels of positive constructs and relatively lower levels of the problem behaviors. For the causal analysis, the problem behaviors in the SRS and SDQ have been reverse coded in a manner that preserves ordinal relationships and standard deviations but eases interpretation so higher values are normatively better across all non-cognitive constructs in the teacher and child surveys. The SRS and SDQ measures are also standardized by grade, so that each variable in each survey period has mean zero and unit variance.[9] Table 5 contains the

[9] The means of the non-standardized variables remain fairly stable across all survey periods, suggesting that teachers were already conceptually standardizing their responses by grade level. Standardizing merely formalizes this.

non-cognitive skills, their labels for this analysis, the number of survey items from which they are constructed, and examples of each skill.

Table 6 contains the correlations amongst the non-cognitive skills in the teacher and parent surveys in the fall kindergarten period. As is apparent, there are higher correlations amongst skills within surveys than between them. Correlations of constructs range from .22 to .28 within the parent survey and from .24 to .78 within the teacher survey. However, many of the between-survey correlations are less than .1. This is further evidence that these two surveys are picking up distinct elements of children's behavior across two different environments: home and school. As is expected, the measures of non-cognitive skills from teachers are more highly correlated with children's cognitive assessments than are the skills measured by parents. Table 7 contains the correlations amongst the SRS and cognitive assessments of literacy, numeracy, and general knowledge in the fall kindergarten period. Approaches to learning correlates more strongly than the other four SRS constructs with the three cognitive assessments (.34 to .41 vs. .13 to .25). This is also commensurate with the prior research demonstrating that non-cognitive skills measure aspects of child development that are separate from and only modestly correlated with cognitive development (Almlund et al., 2011).[10] These non-cognitive skills also vary significantly over the course of a child's elementary school years. Table 8 contains the proportion of the variance in each construct due to variation within children over time, as well as the white-black, white-Hispanic, and treat-control gaps. Between one third and one half of the variation in non-cognitive skills is within a child over time rather than between children. Though this is less within-child variation than that of the ECLS-K literacy assessment, is it enough to suggest malleability and response to policy interventions. The ethnic achievement gaps are merely the percentage point difference for the parent survey measured and for the SRS are calculated using the non-parametric method preferred by Reardon and Ho (2015, p. 159), which is equivalent to Cohen's d for normally distributed variables. The SRS ethnicity gaps are considerably smaller that the white-black and white-Hispanic literacy gaps, yet many are still troubling large.

Taken together, the data supports the prior evidence that non-cognitive skills vary over childhood, are distinct from measures of cognitive skills, and are thus likely malleable by education policy.

Section V- Results

Nationally, I find scant evidence that NCLB had any impact on students' non-cognitive growth, as measured by parents or teachers. Figure 2 contains the mean, standardized self-control ratings from the SRS by grade-level and treatment group. As is apparent, the control (early adopter) and treatment (late adopter) groups do not significantly diverge in either the pre- or post-NCLB periods. Figure 3 reveals similar patterns across the remaining four SRS skills- little change due to NCLB. The formal CITS analysis confirms this visual pattern. Table 9 contains the results from the full regressions, with the treatment effect captured by NCLB*Treat- the shift in levels due to approximately two years of treatment. In no case is the impact statistically significant even at the .1 level. Negative impacts of one fortieth to one tenth of a standard deviation can be rejected and positive impacts of one seventieth to one seventh of a standard deviation can be rejected.

[10] A natural question is how these skills correlate with students' academic grades. ECLS-K collected report cards from teachers in each survey period, but has not yet transferred the grades to the data. Such an analysis is therefore not possible at present.

The skills measured by parents display the same pattern, both visually and statistically, as those measured by teaches: very little impact from NCLB. Figure 4 contains the time series graphs for the four measures from the parent survey and Table 10 presents the statistical analysis. The impact of NCLB (NCLB*Treat) is not statistically significant for any of the four variables at the .1 level. Negative impacts of 1.6 to 4.1 percentage points can be rejected and positive impacts of 1.2 to 3.5 percentage points can be rejected.

To provide a sense of scale, the point estimates and confidence intervals are given in Figure 5 along with the standardized effect from an RCT of twelve months of exposure to a pedagogy focused on non-cognitive skills (Barnett et al., 2008), in the case of the SRS, and the female percentage point advantage in the case of the parent survey. As this figure reveals, NCLB had very little impact at the national level.

When the results are separated by socioeconomic categories, the results become more mixed- with suggestive evidence of some positive effects for African-American and Hispanic students from the teacher surveys and negative effects for white, Asian, and LEP students. Table 11 reveals small negative effects on self-control and peer relationships for white students, a modest negative impact on externalizing problem behaviors for Asian students, and a modest negative impact on approaches to learning for students with limited English proficiency. Conversely, African-American students displayed fewer internalizing problem behaviors and Hispanic students had improved peer relationships. Beyond statistical significance, all of the coefficients for African-American students and four out of five for Hispanic students are positive; four out of five for white and Asian students are negative. This is reflected in the final column of Table 11, which provides the mean effect of NCLB on all SRS variables with accompanying standard errors.[11] Though none of the mean effects can be distinguished from zero at the .1 level of significance, the directions suggest that NCLB had weak positive effects on the classroom behavior of certain students who were more likely to be targets of the policy, contrary to the concerns of its critics.

The sub-group specific results from the parent survey reveals mostly negative effects. Table 12 shows weak to modest negative effects for African-American, Asian, free lunch, and limited English proficiency students. Conversely, Hispanic students displayed greater independence due to NCLB. The two statically significant mean effects are both negative (African-American and LEP). Assuming that these results are not due to differential changes in psychometric validity of either the parent survey or SRS, NCLB appears to have had divergent impacts on school and home behavior. In the classroom, the policy improved non-cognitive skills for ethnic groups that were explicitly targeted by it, and worsened those skills for others. At home, the impact of NCLB was a decline in certain skills for most sub-groups, with overall negative or statistically insignificant effects.

Tables Table 13 through Table 20 provide the results from the supplementary analysis of students' ratings of their own non-cognitive skills as well as standardized tests of literacy and numeracy. Table 13 contains the DID results for the entire sample and reveals little change due to NCLB. Table 14 through Table 16 uncover significant heterogeneity in the results. White and free lunch students rated themselves worse and Hispanic students rated themselves better for both sets of problem behaviors. LEP and Asian students reported no significant change, while

[11] The mean effect is the arithmetic mean of the five standardized beta coefficients for each subgroup. Standard errors were calculated by simulating the sampling distribution of the mean effects with 1000 samples (taken with replacement) clustered at the student level. For recent uses of mean effects in economic literature, see Kling, Liebman, and Katz (2007) and Casey, Glennerster, and Miguel (2012).

African-American students reported a decline in peer relations. In tables Table 17 through Table 20, I find effects on student achievement that are qualitatively commensurate with the prior literature. Overall, I find a statistically significant increase in math scores of one tenth of a standard deviation and no change in reading scores. I find statistically significant increases in both scores for white and African-American students of roughly one tenth (math) to one twentieth (reading) of a standard deviation in each group. I find an increase in math scores and decrease in reading scores for free lunch students. No significant changes are present for Hispanic, Asian, and LEP students. Though the nature of the heterogeneity is somewhat different, taken as a whole these results support the conclusion from the prior literature that NCLB had an overall stronger effect on math scores and a weaker or non-existent effect on reading scores.

The ECLS-K data allows tests of several potential mechanisms through which NCLB's accountability requirements impacted non-cognitive skills. The stock of teachers might have changed with respect to pedagogical training in or emphasis of non-cognitive skills development. ECLS-K asks teachers how many child development courses they have taken in college as well as how much emphasis they place on non-cognitive skills and standardized testing when evaluating students. Figure 6 lists the coefficients and 90% confidence intervals for the impact of NCLB on the number of child development courses teachers report taking in college. Overall, the policy appears to have decreased teacher training in this area, though only two coefficients are negative and significant and one is positive and significant. A systematic increase can be ruled out.

Table 21 contains the impact of NCLB on how much importance teachers place on students' effort, completion of homework, participation, and behavior as well as students' achievement on standardized tests. The variables are ordinal scales (not important to extremely important) that have been dichotomized into extremely important (1) and not extremely important (0). The results are not suggestive of non-cognitive skills becoming less important due to NCLB. African-American students have teachers who place more emphasis on completion of homework and less on participation, Hispanic students have teachers who place more emphasis on standardized tests and less on behavior, Asian students have teachers who place more emphasis on completion of homework and less on standardized tests. Together, these results rule out increases in post-secondary child development courses and shifts in evaluative emphasis (as reported in the ECLS-K surveys) as mechanisms through which NCLB might have impacted non-cognitive skills.

Section VI- Discussion

What to make of the divergence in effects on teacher vs. parent ratings is unclear. Table 22 provides suggestive evidence that no single domain should be privileged over the others. The table contains the adjusted R^2 statistic taken from regressions of standardized test scores and teachers' ratings of overall academic proficiency in reading and math on three sets of non-cognitive skills: those measured from surveys of teachers, parents, and the children themselves. The data comes from fifth grade tests and surveys. The only clear pattern is that teachers' ratings of non-cognitive skills are more closely correlated with their ratings of academic proficiency than are parents' and children's ratings of non-cognitive skills.

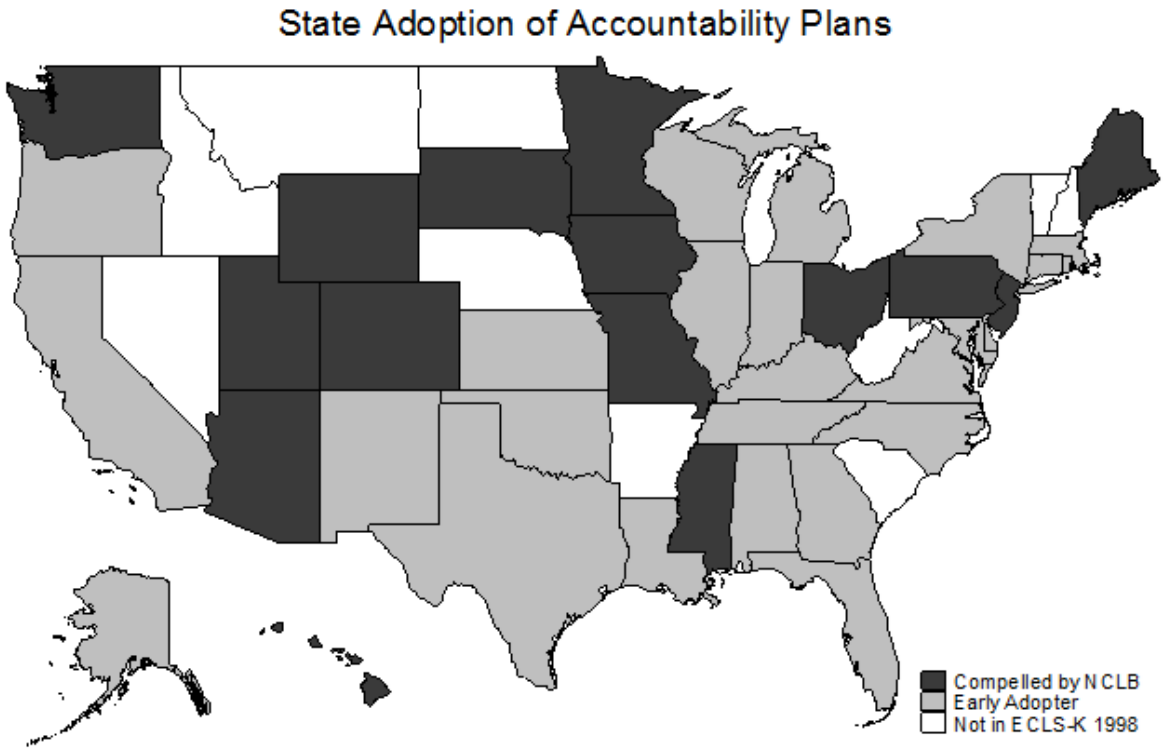
Teacher ratings have the benefit of being drawn from a greater number of survey items that address specific situations rather than the parent ratings which are each from a single question. Moreover, teachers are more likely than parents to have exposure to children along the entire distribution of each skill being measured. The potential impacts of exposure on ratings is

manifest in the drop in parents' ratings of children between the fall of kindergarten and the spring of first grade (see figure Figure 4). No such change is observed on the teacher ratings, suggesting that parents think more highly of their children when they have less exposure to a school setting. A similar dynamic could be happening with the onset of NCLB if the law increased parents' interactions with schools. Table 23 reveals that NCLB increased parental attendance at open houses, parents' informal meetings with teachers, and parent-initiated contact with teachers. The law also marginally increased the number of other parents the survey child's parents spoke with regularly, but the effect is just under the .1 significance threshold ($p = .11$). Ultimately, the data is unable to fully account for the discrepancy of NCLB effects on parents and teacher ratings.

An important limitation of this research is that the structure of this data does not allow the estimation of NCLB effects at numerous ages and stages of child development. The weak and overall null effects of the law on non-cognitive skills might be local to the age group represented here: late elementary school. If teachers are sufficiently focused on non-cognitive skill development at these ages, then a pedagogical shift toward literacy and numeracy may not have reduced this focus appreciably. Alternately, teachers of this age group may specifically target non-cognitive skills as a means to enhancing literacy and numeracy. The impact of NCLB on the non-cognitive skills of older children cannot be ascertained from this analysis, yet is vital to understand.

Finally, it bears repeating that this is a test of the federal accountability mandate on students' non-cognitive skills rather than a test of any particular accountability system and accompanying endogenous implementation choices. Though limited, this focus on the federal mandate provides salient information on the type of policy the federal government is likely to be able to enact. As the federalist tensions in the drafting and implementation of NCLB attest to, the U.S. Department of Education has a limited amount of power to impact local schools (Manna, 2011). It is therefore essential to understand how this major federal policy impacted students overall, without limiting the analysis to only those states whose implementation was particularly faithful to the law's intent.

Figure 1: Early Adopter and Late Adopter States in the ECLS-K 1998



Source: Dee and Jacob (2011) *and the Early Childhood Longitudinal Study of Kindergarten, 1998, base year file*

Table 1: ECLS-K Summary Statistics at Fall Kindergarten Students Attending Public Schools

	Early Adopter	Late Adopter
	Students	
White	**49%**	**62%**
Black	**17%**	**13%**
Hispanic	**23%**	**11%**
Attended Head Start	**12%**	**17%**
Limited English Proficiency	**25%**	**16%**
Free/Reduced Price Lunch	**35%**	**27%**
	Teachers & Schools	
Teacher's with MA	**34%**	**27%**
Title 1 School	**52%**	**67%**

Source: Early Childhood Longitudinal Study of Kindergarten, 1998, base year file

Table 2: Distribution of Parental Assessments of Children's Non-cognitive Skills

Compared to children of the same age, the child can

	Independent	Pay Attention	Solve Problems	Behave/ Relate to others
Better	**33%**	**25%**	**34%**	**27%**
As well	**60%**	**61%**	**57%**	**66%**
Slightly less well	**5.6%**	**12%**	**7.9%**	**5.7%**
Much less well	**0.9%**	**1.8%**	**1%**	**0.8%**

Source: Early Childhood Longitudinal Study of Kindergarten, 1998, base year file

Table 3: Distributions of the Social Rating Scale (Teachers) and the Self-Description Questionnaire (Children)

Teacher SRS	Mean	Median	Standard Deviation
Self-Control	**3.18**	**3.25**	**.614**
Approaches to Learning	**3.07**	**3.14**	**.684**
Interpersonal Skills	**3.08**	**3**	**.64**
Externalizing Behaviors	**1.64**	**1.5**	**.612**
(Reversed)	**(3.36)**	**(3.5)**	**(.612)**
Internalizing Behaviors	**1.58**	**1.5**	**.522**
(Reversed)	**(3.42)**	**(3.5)**	**(.522)**
Child SDQ			
Interpersonal Skills	**3.01**	**3**	**.621**
Externalizing Behaviors	**1.83**	**1.67**	**.65**
(Reversed)	**(3.17)**	**(3.33)**	**(.65**
Internalizing Behaviors	**2.1**	**2**	**.649**
(Reversed)	**(2.9)**	**(3)**	**(.649)**

Source: Early Childhood Longitudinal Study of Kindergarten, 1998, base year file

Table 4: Data Sources in the ECLS-K

Outcome	**Questionnaire**	**Pre-NCLB**			**Post-NCLB**	
Outcomes for Primary Models		**Kindergarten**	**1st Grade**	**3rd Grade**	**5th Grade**	**8th Grade**
Independence	**Parent**	**X**	**X**	**X**	**X**	**X**
Pays Attention	**Parent**	**X**	**X**	**X**	**X**	**X**
Solve Problems	**Parent**	**X**	**X**	**X**	**X**	**X**
Behaves Well	**Parent**	**X**	**X**	**X**	**X**	**X**
Self-Control	**Teacher**	**X**	**X**	**X**	**X**	
Externalizing	**Teacher**	**X**	**X**	**X**	**X**	
Internalizing	**Teacher**	**X**	**X**	**X**	**X**	
Learning Approach	**Teacher**	**X**	**X**	**X**	**X**	
Interpersonal Skill	**Teacher**	**X**	**X**	**X**	**X**	
Outcomes for Secondary Models						
Interpersonal Skill	**Student**			**X**	**X**	
Externalizing	**Student**			**X**	**X**	
Internalizing	**Student**			**X**	**X**	**X**

Source: Early Childhood Longitudinal Study of Kindergarten, 1998

Table 5: Constructs from ECLS-K 1998

Construct	Label	Items per Survey	Examples of Behavior
Self-Control	**CONTRO**	**4**	**controlling temper, accepting peer ideas**
Approaches to Learning	**LEARN**	**6**	**attentiveness, task persistence, eagerness to learn**
Peer Relations	**PEER**	**5**	**expressing feelings/ideas/opinions in positive ways**
Externalizing Problems	**EXTERN**	**5**	**argues, gets angry, acts impulsively, disruptive**
Internalizing Problems	**INTERN**	**4**	**anxiety, loneliness, low self-esteem, and sadness**
Independence	**INDEP**	**1**	**independent and takes care of herself**
Pay Attention	**ATTN**	**1**	**pays attentions**
Solve Problems	**SOLVE**	**1**	**learn, think, and solve problems**
Overall Behavior	**BEHAV**	**1**	**behaves /relates to children and adults**

Source: Early Childhood Longitudinal Study of Kindergarten, 1998

Table 6: Correlations amongst Non-cognitive Skills in the Fall Kindergarten Survey (1998)

	INDEP	**ATTN**	**SOLVE**	**BEHAV**	**EXTERN**	**INTERN**	**PEER**	**CONTRO**	**LEARN**
INDEP	1								
ATTN	0.231	1							
SOLVE	0.254	0.280	1						
BEHAV	0.234	0.264	0.217	1					
EXTERN	0.059	0.195	0.051	0.117	1				
INTERN	0.061	0.060	0.093	0.065	0.244	1			
PEER	0.074	0.178	0.073	0.121	0.690	0.268	1		
CONTRO	0.096	0.175	0.103	0.141	0.553	0.341	0.779	1	
LEARN	0.134	0.241	0.192	0.140	0.507	0.352	0.662	0.696	1

Source: Early Childhood Longitudinal Study of Kindergarten, 1998

Table 7: Correlation amongst Social Rating Scale and Academic Assessments in the Fall Kindergarten Survey (1998)

	EXTERN	INTERN	PEER	CONTRO	LEARN	READ	MATH	GEN
EXTERN	**1.00**							
INTERN	**0.28**	**1.00**						
PEER	**0.57**	**0.35**	**1.00**					
CONTRO	**0.70**	**0.29**	**0.79**	**1.00**				
LEARN	**0.51**	**0.37**	**0.70**	**0.67**	**1.00**			
READ	**0.13**	**0.13**	**0.20**	**0.17**	**0.34**	**1.00**		
MATH	**0.15**	**0.18**	**0.24**	**0.20**	**0.41**	**0.72**	**1.00**	
GEN	**0.14**	**0.15**	**0.25**	**0.20**	**0.34**	**0.50**	**0.61**	**1.00**

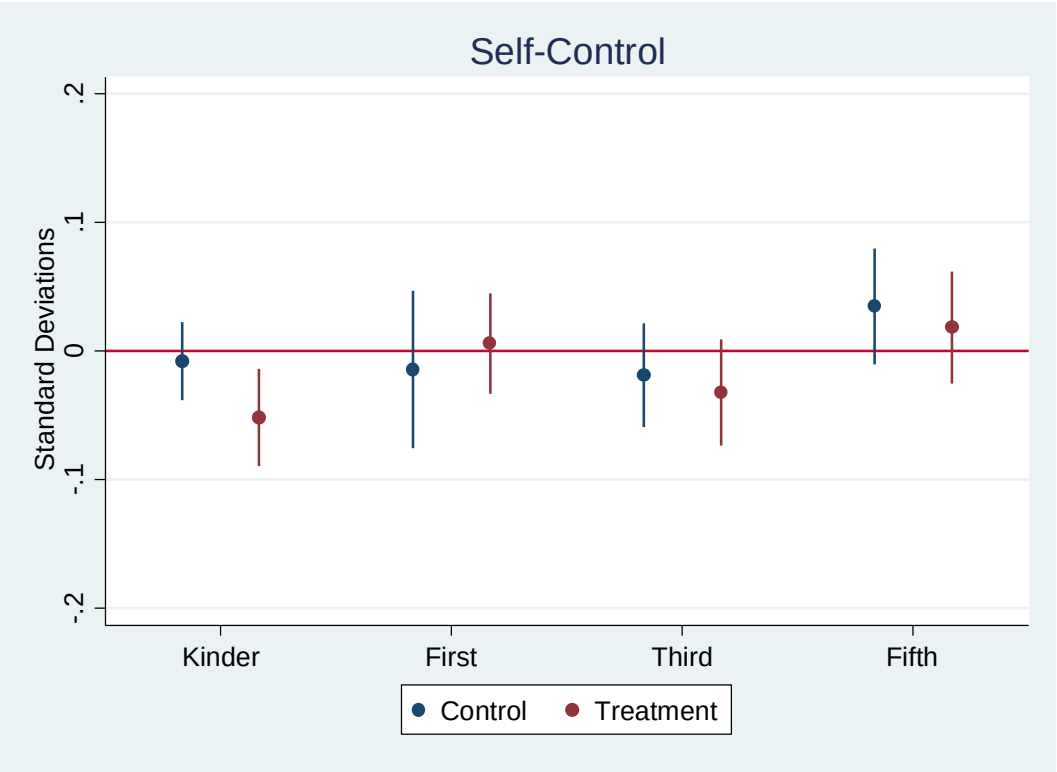
Source: Early Childhood Longitudinal Study of Kindergarten, 1998

Table 8: Within Child Variance and Skills Gaps

	% of Variance within Child	White-Black Gap	White-Hisp Gap	Treat-Control Gap
READ	0.82	0.56	0.55	-0.05
CONTRO	0.45	0.38	0.16	-0.04
EXTERN	0.36	0.32	0	-0.03
INTERN	0.55	0.1	0.12	0.02
PEER	0.46	0.34	0.2	-0.04
LEARN	0.37	0.38	0.26	0.01
BEHAV	0.52	0.04	0	-0.01
ATTN	0.5	0.08	-0.02	-0.02
SOLVE	0.48	0.08	0.06	0
INDEP	0.51	0.02	0	-0.01

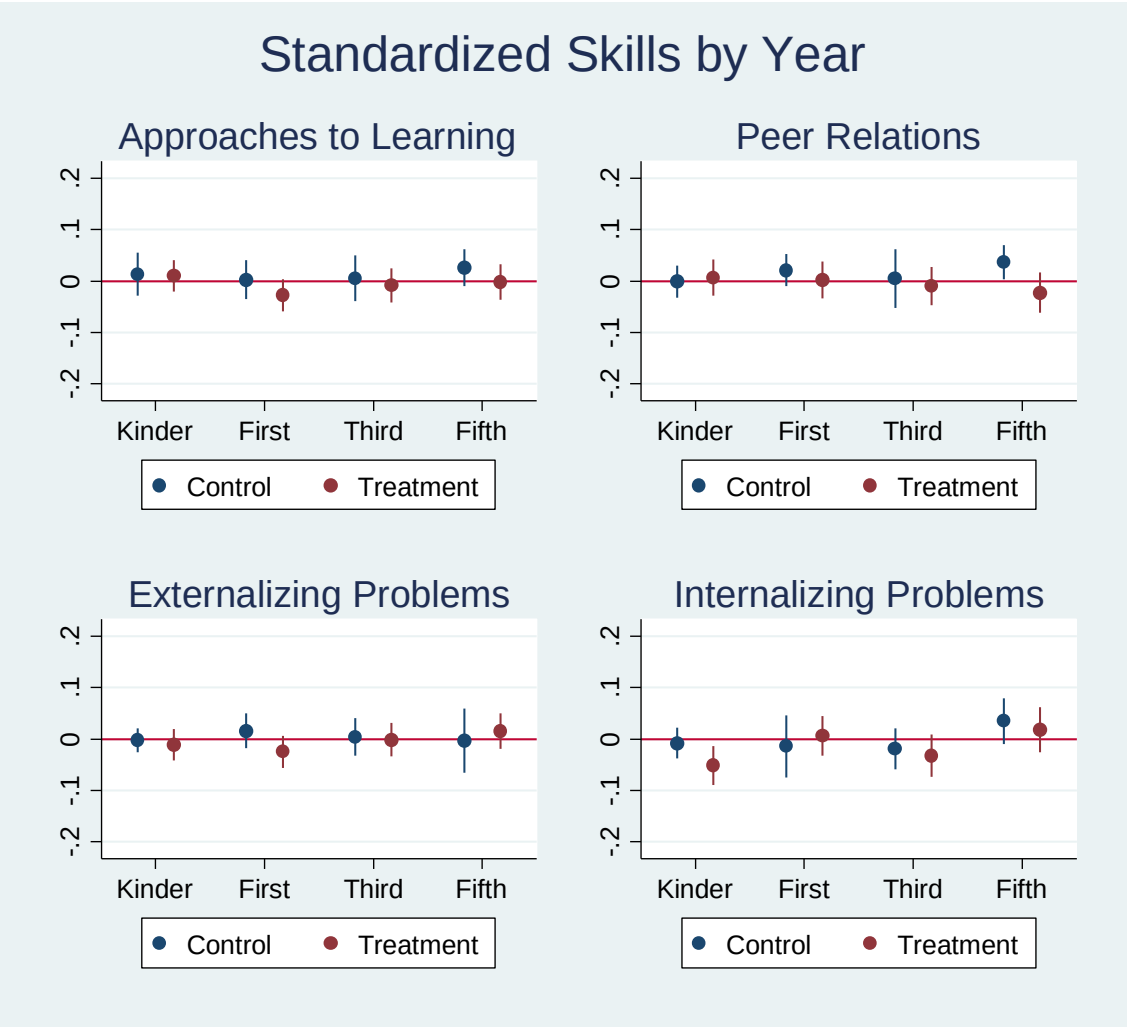
Source: Early Childhood Longitudinal Study of Kindergarten, 1998

Figure 2: Standardized Self-Control from the Social Rating Scale



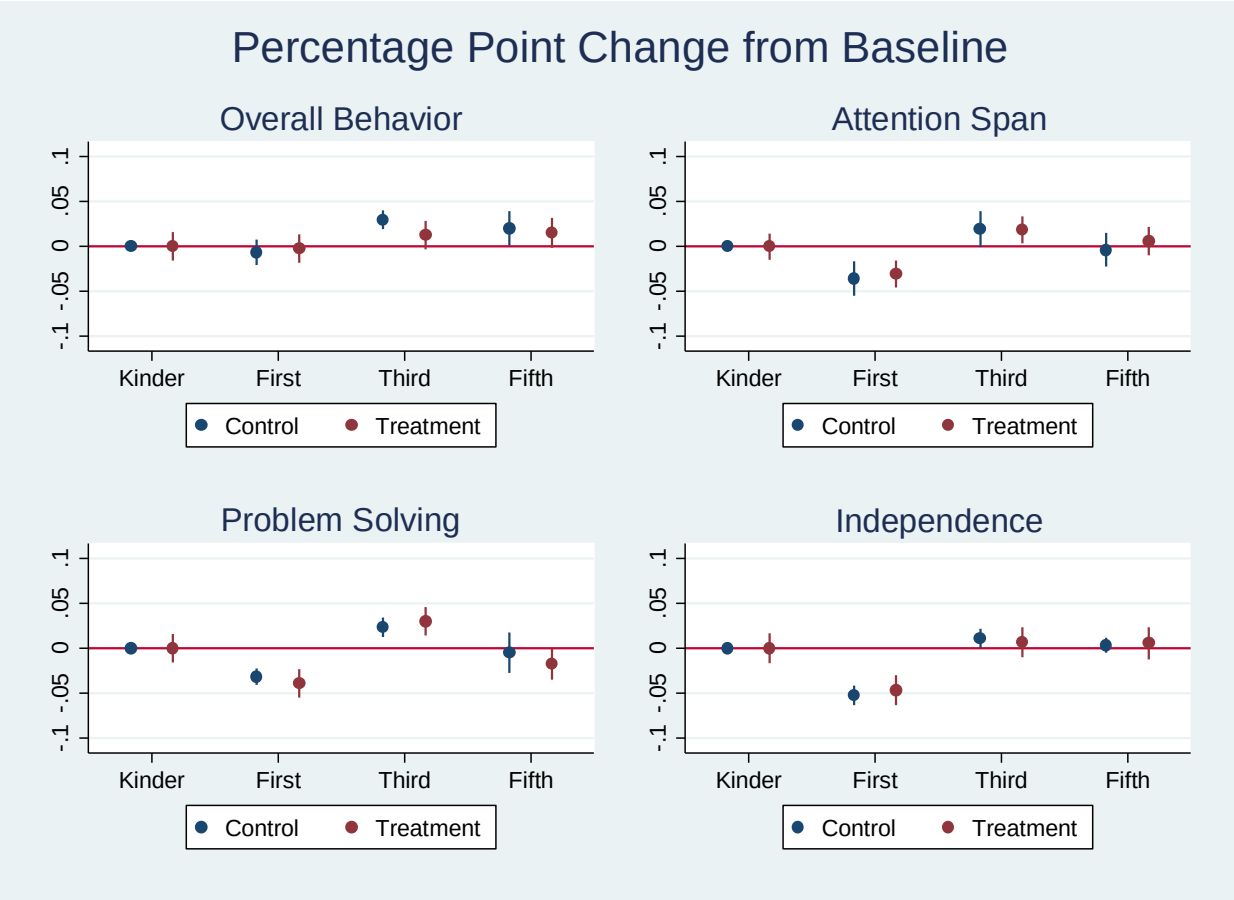
Source: Early Childhood Longitudinal Study of Kindergarten, 1998

Figure 3: Standardized Skills from the Social Rating Scale



Source: Early Childhood Longitudinal Study of Kindergarten, 1998

Figure 4: Parent Ratings of Students' Skills by Grade and Treatment Group



Source: Early Childhood Longitudinal Study of Kindergarten, 1998

Table 9: The Impact of NCLB on Teacher ratings of students' Non-cognitive Skills

	Self-Control	Externalizing Problems	Internalizing Problems	Peer Relations	Approaches to Learning
Year	0.009	0.005	-0.011	0.007	0.005
	(0.008)	(0.007)	(0.008)	(0.007)	(0.005)
NCLB	-0.001	-0.015	0.057	0.020	0.013
	(0.036)	(0.041)	(0.030)*	(0.028)	(0.026)
Year*Treat	-0.007	-0.015	-0.007	-0.009	-0.014
	(0.008)	(0.007)*	(0.008)	(0.007)	(0.005)**
NCLB*Treat	-0.037	0.057	-0.001	-0.044	0.003
	(0.037)	(0.041)	(0.030)	(0.029)	(0.026)
SES (Continuous)	-0.015	-0.031	0.036	-0.031	-0.036
	(0.023)	(0.017)*	(0.018)*	(0.025)	(0.024)
Father Present	0.014	0.018	0.111	0.045	0.062
	(0.021)	(0.015)	(0.027)***	(0.027)	(0.035)
Father Employed Fulltime	0.044	0.018	0.029	0.038	0.015
	(0.022)*	(0.015)	(0.030)	(0.019)*	(0.019)
Constant	-0.095	-0.069	-0.121	-0.123	-0.111
	(0.012)***	(0.009)***	(0.034)***	(0.015)***	(0.025)***
R^2	0.00	0.00	0.00	0.00	0.00
N	39,539	39,828	39,497	39,218	40,102

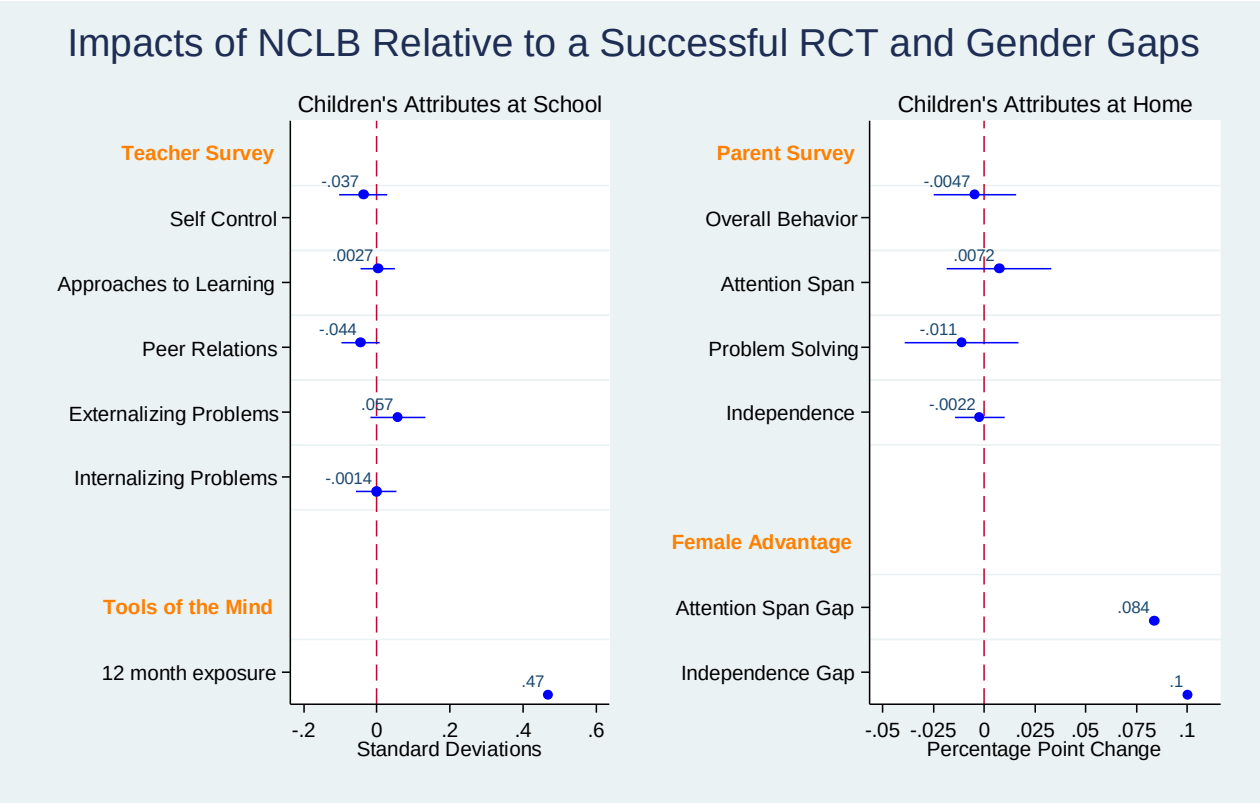
* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Table 10: The Impact of NCLB on Parent ratings of students' Non-cognitive Skills

	Overall Behavior	Attention Span	Problem Solving	Independence
Year	0.003	-0.003	-0.001	-0.009
	(0.001)*	(0.003)	(0.002)	(0.001)***
NCLB	0.018	0.013	0.009	0.035
	(0.011)	(0.014)	(0.015)	(0.007)***
Year*Treat	-0.001	0.001	0.001	0.002
	(0.001)	(0.003)	(0.002)	(0.001)
NCLB*Treat	-0.005	0.007	-0.011	-0.002
	(0.011)	(0.014)	(0.015)	(0.007)
SES (Continuous)	0.019	0.016	0.007	0.005
	(0.006)**	(0.011)	(0.004)*	(0.006)
Father Present	0.002	-0.003	0.009	0.011
	(0.007)	(0.009)	(0.010)	(0.018)
Father Employed Fulltime	-0.024	0.007	-0.005	0.009
	(0.010)**	(0.009)	(0.007)	(0.014)
Constant	0.261	0.238	0.320	0.333
	(0.007)***	(0.011)***	(0.007)***	(0.009)***
R^2	0.00	0.00	0.00	0.00
N	47,878	47,884	47,817	47,904

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Figure 5: Null Results of NCLB in Comparison to a successful RCT and Female Advantage



Source: Early Childhood Longitudinal Study of Kindergarten, 1998

Table 11: Effects of NCLB on SRS Ratings

Sub-Group	Self-Control	Peer Relations	Approaches to Learning	Internalizing Problems	Externalizing Problems	Mean Effect
White	-0.071*	-0.080**	-0.030	-0.039	0.058	-.033
	(0.037)	(0.034)	(0.026)	(0.035)	(0.043)	(.032)
African-American	.092	.116	.200	.271**	.085	.153
	(.060)	(.103)	(.125)	(.100)	(.100)	(.098)
Hispanic	.035	.115**	.096	-.024	.081	.061
	(.060)	(.045)	(.060)	(.081)	(.096)	(.088)
Asian	.026	-.138	-.125	-.053	-.123**	-.083
	(.100)	(.090)	(.080)	(.093)	(.049)	(.103)
Free Lunch	-.031	-.017	.023	.118	.003	.019
	(.041)	(.067)	(.042)	(.075)	(.067)	(.059)
Limited English	-.063	-.091	-.135*	.090	.008	-.038
	(.080)	(.080)	(.073)	(.082)	(.062)	(.065)

*Standard Errors in Parentheses, * p<0.1; ** p<0.05; *** p<0.01, Source: ECLS-K 1998*

Table 12: The Impacts of NCLB on Parents' Ratings of Children's Non-cognitive Skills

Sub-Group	Overall Behavior	Problem Solving	Attention Span	Independence	Mean Effect
White	-0.006 (.014)	0.004 (.017)	0.011 (.012)	0.008 (.013)	.004 (.011)
African-American	-.066* (.030)	-0.080** (0.033)	-0.052 (.053)	-0.112*** (.031)	-.077** (.036)
Hispanic	.006 (.020)	-0.011 (.039)	0.049 (.046)	0.066** (.028)	.027 (.033)
Asian	.072 (.041)	-0.096*** (.022)	-0.003 (.027)	-0.059* (.031)	-.021 (.053)
Free Lunch	.008 (.026)	-0.031 (.036)	-0.018 (.039)	-0.075*** (.020)	-.029 (.022)
Limited English	-.034** (.012)	-0.017 (.013)	-0.074*** (.016)	-0.042** (.018)	-.042* (.026)

*Standard Errors in Parentheses, * p<0.1; ** p<0.05; *** p<0.01, Source: ECLS-K 1998*

Table 13: The Impact of NCLB on Students' Ratings of Non-cognitive Skills (Diff-in-Diff model)

	Peer Relations	Internalizing Problems	Externalizing Problems
NCLB	-0.008 (0.008)	0.007 (0.017)	-0.002 (0.013)
NCLB*Treat	0.009 (0.009)	-0.027 (0.017)	-0.013 (0.013)
SES (Continuous)	0.116 (0.047)**	-0.060 (0.030)*	-0.041 (0.051)
Father Present	0.072 (0.037)*	0.036 (0.039)	0.006 (0.044)
Father Employed Fulltime	-0.091 (0.032)**	-0.046 (0.041)	0.010 (0.022)
Constant	0.014 (0.021)	0.011 (0.028)	0.014 (0.038)
R^2	0.00	0.00	0.00
N	15,348	15,348	15,348

* p<0.1; ** p<0.05; *** p<0.01, *Source: ECLS-K 1998*

Table 14: The Impact of NCLB on Students' Ratings of Non-cognitive Skills by Race

	Peer Relations	Externalizing Problems	Internalizing Problems
NCLB	-0.022	-0.004	0.010
	(0.020)	(0.011)	(0.021)
NCLB*Treat	0.028	-0.022	-0.051
	(0.020)	(0.011)*	(0.021)**
SES (Continuous)	0.111	-0.042	-0.060
	(0.048)**	(0.050)	(0.031)*
Father Present	0.077	0.008	0.037
	(0.036)*	(0.042)	(0.041)
Father Employed Fulltime	-0.091	0.009	-0.045
	(0.032)**	(0.022)	(0.041)
African-American*NCLB	0.112	0.006	0.103
	(0.098)	(0.067)	(0.046)**
African-American*NCLB*Treat	-0.204	0.016	0.058
	(0.097)*	(0.067)	(0.046)
Hispanic*NCLB	0.099	-0.029	-0.037
	(0.052)*	(0.031)	(0.052)
Hispanic*NCLB*Treat	-0.014	0.126	0.162
	(0.051)	(0.031)***	(0.052)**
Asian*NCLB	-0.083	0.050	-0.090
	(0.063)	(0.085)	(0.046)*
Asian*NCLB*Treat	-0.013	-0.069	-0.074
	(0.063)	(0.086)	(0.046)
Constant	0.010	0.013	0.010
	(0.022)	(0.036)	(0.027)
R^2	0.00	0.00	0.00
N	15,348	15,348	15,348

* $p<0.1$; ** $p<0.05$; *** $p<0.01$, *Source: ECLS-K 1998*

Table 15: The Impact of NCLB on Students' Ratings of Non-cognitive Skills by Free/Reduced Price Lunch Status

	Peer Relations	Externalizing Problems	Internalizing Problems
NCLB	-0.040	-0.010	-0.016
	(0.016)**	(0.018)	(0.019)
NCLB*Treat	0.007	0.017	-0.005
	(0.017)	(0.018)	(0.019)
SES (Continuous)	0.112	-0.044	-0.064
	(0.049)**	(0.051)	(0.032)*
Father Present	0.072	0.004	0.035
	(0.035)*	(0.044)	(0.039)
Father Employed Fulltime	-0.090	0.012	-0.044
	(0.032)**	(0.022)	(0.041)
Free/Reduced Lunch*NCLB	0.126	0.032	0.091
	(0.035)***	(0.039)	(0.042)*
Free/Reduced Lunch*NCLB*Treat	0.013	-0.124	-0.088
	(0.036)	(0.039)***	(0.042)*
Constant	0.013	0.014	0.010
	(0.021)	(0.038)	(0.028)
R^2	0.00	0.00	0.00
N	15,348	15,348	15,348

* $p<0.1$; ** $p<0.05$; *** $p<0.01$, *Source: ECLS-K 1998*

Table 16: The Impact of NCLB on Students' Ratings of Non-cognitive Skills by LEP Status

	Peer Relations	Externalizing Problems	Internalizing Problems
NCLB	-0.011	-0.012	0.015
	(0.011)	(0.011)	(0.020)
NCLB*Treat	0.006	-0.015	-0.029
	(0.012)	(0.011)	(0.020)
SES (Continuous)	0.116	-0.039	-0.061
	(0.047)**	(0.050)	(0.030)*
Father Present	0.072	0.004	0.038
	(0.037)*	(0.043)	(0.039)
Father Employed Fulltime	-0.091	0.010	-0.046
	(0.032)**	(0.023)	(0.041)
Limited Eng Prof*NCLB	0.011	0.044	-0.035
	(0.025)	(0.038)	(0.042)
Limited Eng Prof*NCLB*Treat	0.022	0.033	-0.004
	(0.025)	(0.038)	(0.042)
Constant	0.015	0.016	0.010
	(0.021)	(0.038)	(0.028)
R^2	0.00	0.00	0.00
N	15,348	15,348	15,348

* $p<0.1$; ** $p<0.05$; *** $p<0.01$, Source: ECLS-K 1998

Table 17: The Impact of NCLB on Standardized Math and Reading Scores

	Std. Math Score	Std. Reading Score
Year	0.029	0.012
	(0.010)**	(0.014)
NCLB	-0.074	-0.045
	(0.036)*	(0.028)
Year*Treat	-0.039	-0.004
	(0.009)***	(0.014)
NCLB*Treat	0.104	0.044
	(0.036)**	(0.028)
SES (Continuous)	0.032	0.018
	(0.020)	(0.018)
Father Present	-0.015	-0.031
	(0.025)	(0.032)
Father Employed Fulltime	-0.007	-0.026
	(0.020)	(0.021)
Constant	-0.036	-0.064
	(0.032)	(0.028)**
R^2	0.00	0.00
N	21,063	19,628

* $p<0.1$; ** $p<0.05$; *** $p<0.01$, Source: ECLS-K 1998

Table 18: The Impact of NCLB on Standardized Tests of Math and Reading by Race

	Std. Math Score	Std. Reading Score
Year	0.034	0.036
	(0.011)**	(0.012)**
NCLB	-0.087	-0.087
	(0.040)*	(0.024)***
Year*Treat	-0.032	-0.002
	(0.011)**	(0.012)
NCLB*Treat	0.098	0.049
	(0.040)**	(0.024)*
SES (Continuous)	0.028	0.015
	(0.019)	(0.018)
Father Present	-0.011	-0.031
	(0.022)	(0.027)
Father Employed Fulltime	-0.000	-0.013
	(0.018)	(0.017)
African-American*Year	-0.077	-0.154
	(0.012)***	(0.016)***
African-American*NCLB	0.041	0.216
	(0.041)	(0.033)***
African-American*Year*Treat	-0.057	-0.035
	(0.012)***	(0.017)*
African-American*NCLB*Treat	0.095	0.065
	(0.041)**	(0.035)*
Hispanic*Year	0.011	-0.062
	(0.013)	(0.023)**
Hispanic*NCLB	0.025	0.128
	(0.039)	(0.085)
Hispanic*Year*Treat	-0.007	0.027
	(0.013)	(0.023)
Hispanic*NCLB*Treat	-0.013	-0.092
	(0.040)	(0.085)
Asian*Year	0.005	-0.089
	(0.018)	(0.037)**
Asian*NCLB	0.099	0.200
	(0.049)*	(0.096)*
Asian*Year*Treat	0.018	-0.002
	(0.018)	(0.037)
Asian*NCLB*Treat	-0.074	-0.078
	(0.050)	(0.096)
Constant	-0.045	-0.074
	(0.028)	(0.025)**
R^2	0.02	0.03
N	21,063	19,628

* $p<0.1$; ** $p<0.05$; *** $p<0.01$, Source: ECLS-K 1998

Table 19: The Impact of NCLB on Standardized Tests of Math and Reading by free/reduced price lunch status

	Std. Math Score	Std. Reading Score
Year	0.040	0.035
	(0.010)***	(0.013)**
NCLB	-0.091	-0.099
	(0.041)**	(0.022)***
Year*Treat	-0.026	0.001
	(0.009)**	(0.013)
NCLB*Treat	0.080	0.056
	(0.040)*	(0.022)**
SES (Continuous)	0.030	0.015
	(0.019)	(0.018)
Father Present	-0.012	-0.024
	(0.024)	(0.032)
Father Employed Fulltime	0.002	-0.013
	(0.018)	(0.018)
Free/Reduced Lunch*Year	-0.048	-0.106
	(0.010)***	(0.021)***
Free/Reduced Lunch*NCLB	0.072	0.252
	(0.039)*	(0.030)***
Free/Reduced Lunch*Year*Treat	-0.058	-0.026
	(0.010)***	(0.022)
Free/Reduced Lunch*NCLB*Treat	0.105	-0.058
	(0.038)**	(0.029)*
Constant	-0.045	-0.080
	(0.033)	(0.026)**
R^2	0.02	0.02
N	21,063	19,628

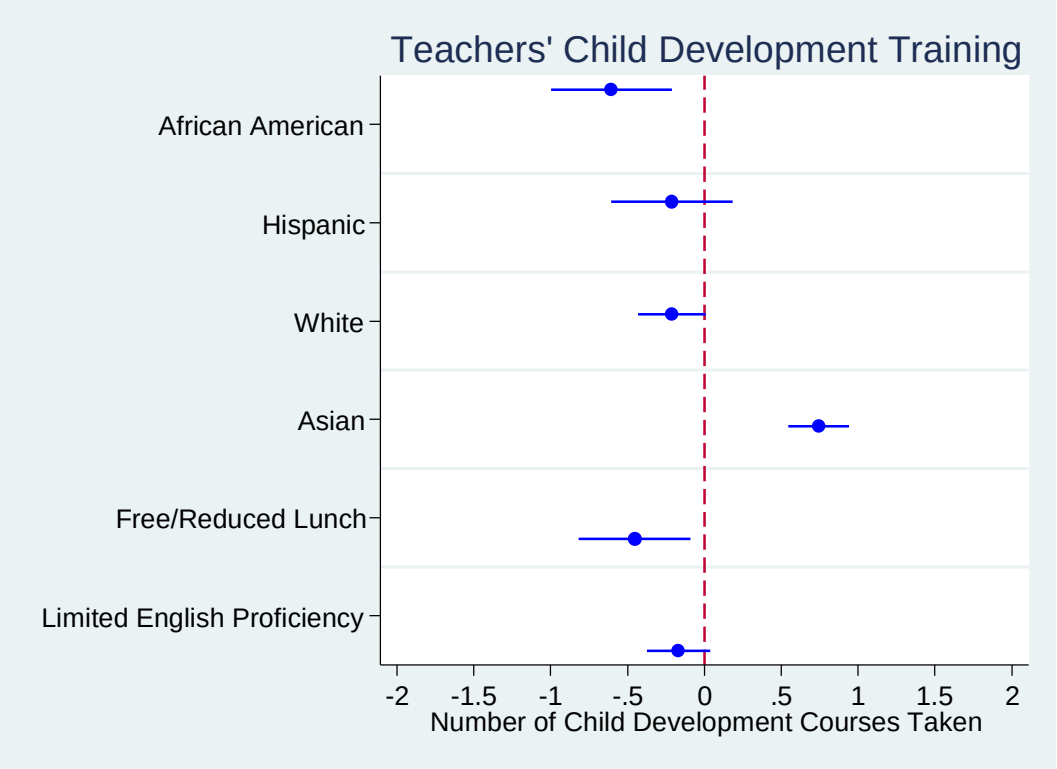
* $p<0.1$; ** $p<0.05$; *** $p<0.01$, Source: ECLS-K 1998

Table 20: The Impact of NCLB on Standardized Tests of Math and Reading by LEP status

	Std. Math Score	Std. Reading Score
Year	0.022	0.012
	(0.010)**	(0.018)
NCLB	-0.067	-0.060
	(0.042)	(0.034)
Year*Treat	-0.033	-0.003
	(0.009)***	(0.017)
NCLB*Treat	0.093	0.066
	(0.042)**	(0.034)*
SES (Continuous)	0.032	0.019
	(0.020)	(0.018)
Father Present	-0.015	-0.032
	(0.024)	(0.031)
Father Employed Fulltime	-0.007	-0.025
	(0.021)	(0.021)
Limited Eng Prof*Year	0.031	-0.001
	(0.010)***	(0.024)
Limited Eng Prof*NCLB	-0.034	0.073
	(0.042)	(0.070)
Limited Eng Prof*Year*Treat	-0.028	-0.005
	(0.010)**	(0.024)
Limited Eng Prof*NCLB*Treat	0.052	-0.118
	(0.042)	(0.070)
Constant	-0.036	-0.064
	(0.031)	(0.027)**
R^2	0.00	0.00
N	21,063	19,628

* $p<0.1$; ** $p<0.05$; *** $p<0.01$, Source: ECLS-K 1998

Figure 6: The Impact of NCLB on Pedagogical Training for Non-cognitive Skills, by Student Characteristic



Source: Early Childhood Longitudinal Study of Kindergarten, 1998

Table 21: The Impact of NCLB on Teacher Emphasis in Student Evaluation

	Effort	Homework	Participate	Behavior	Std. Tests
Year	-0.000	0.046	0.013	-0.031	0.023
	(0.010)	(0.011)***	(0.016)	(0.017)*	(0.012)*
NCLB	-0.008	-0.070	-0.074	-0.008	0.012
	(0.038)	(0.043)	(0.055)	(0.059)	(0.043)
Year*Treat	-0.029	0.007	-0.015	-0.040	-0.011
	(0.010)**	(0.011)	(0.016)	(0.017)**	(0.012)
NCLB*Treat	0.129	0.005	0.067	0.092	0.061
	(0.038)***	(0.043)	(0.055)	(0.059)	(0.043)
African-American*Year	0.025	0.010	-0.031	0.035	-0.019
	(0.023)	(0.022)	(0.019)	(0.024)	(0.017)
African-American*NCLB	-0.044	-0.042	0.148	0.007	0.052
	(0.070)	(0.054)	(0.087)	(0.047)	(0.049)
African-American*Year*Treat	0.048	-0.090	0.079	-0.004	0.005
	(0.023)*	(0.022)***	(0.020)***	(0.024)	(0.018)
African-American*NCLB*Treat	-0.033	0.268	-0.203	-0.019	0.009
	(0.070)	(0.054)***	(0.087)**	(0.047)	(0.049)
Hispanic*Year	-0.022	-0.021	-0.055	-0.037	0.014
	(0.021)	(0.030)	(0.025)*	(0.025)	(0.022)
Hispanic*NCLB	0.014	-0.020	0.104	0.106	-0.104
	(0.098)	(0.124)	(0.073)	(0.092)	(0.049)*
Hispanic*Year*Treat	0.040	-0.013	0.023	0.059	-0.036
	(0.021)*	(0.030)	(0.024)	(0.025)**	(0.023)
Hispanic*NCLB*Treat	-0.061	0.118	-0.019	-0.170	0.104
	(0.098)	(0.124)	(0.073)	(0.093)*	(0.050)*
Asian*Year	0.010	0.018	-0.016	0.008	0.001
	(0.023)	(0.023)	(0.011)	(0.019)	(0.013)
Asian*NCLB	-0.072	-0.142	0.043	-0.034	-0.012
	(0.051)	(0.057)**	(0.053)	(0.052)	(0.069)
Asian*Year*Treat	0.001	-0.037	-0.017	-0.011	0.057
	(0.023)	(0.023)	(0.011)	(0.019)	(0.013)***
Asian*NCLB*Treat	0.009	0.323	0.042	0.113	-0.180
	(0.051)	(0.058)***	(0.053)	(0.052)*	(0.068)**
Constant	0.697	0.251	0.417	0.717	0.148
	(0.032)***	(0.022)***	(0.030)***	(0.025)***	(0.012)***
R^2	0.01	0.02	0.00	0.03	0.02
N	21,208	20,892	21,196	21,150	20,987

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Table 22: Adjusted R^2 from Regressions of 5th grade Tests and Academic Rating Scales on Non-cognitive Skills

Student Skills	**Standardized Tests**		**Teachers' Academic Rating Scale**	
Measured by	Reading	Math	Reading	Math
Teacher	.153	.143	.368	.23
Parent	.154	.148	.147	.120
Child	.163	.154	.138	.112

Source: Early Childhood Longitudinal Study of Kindergarten, 1998, 5th grade file

Table 23: The Impact of NCLB on Parents' Interactions with School (Teachers and other Parents)

	# of Parents Contacted	Informal Meetings	Initiate Contact w/Teacher	Attend Open House
Year	0.279	-0.012	0.052	0.029
	(0.067)***	(0.005)**	(0.005)***	(0.002)***
NCLB	-0.791	-0.009	-0.137	-0.082
	(0.192)***	(0.027)	(0.011)***	(0.008)***
Year*Treat	-0.040	-0.010	-0.006	-0.006
	(0.068)	(0.005)*	(0.005)	(0.002)**
NCLB*Treat	0.338	0.056	0.021	0.019
	(0.195)	(0.027)*	(0.011)*	(0.008)**
SES (Continuous)	-0.009	-0.003	0.023	0.018
	(0.065)	(0.007)	(0.017)	(0.005)***
Father Present	0.313	0.003	0.006	0.005
	(0.082)***	(0.017)	(0.009)	(0.010)
Father Employed Fulltime	-0.107	-0.019	0.003	0.008
	(0.106)	(0.016)	(0.011)	(0.008)
Constant	1.676	0.695	0.493	0.719
	(0.116)***	(0.008)***	(0.008)***	(0.006)***
R^2	0.01	0.01	0.02	0.01
N	33,470	29,804	33,656	33,619

* $p<0.1$; ** $p<0.05$; *** $p<0.01$, *Source: Early Childhood Longitudinal Study of Kindergarten, 1998*

Money and Freedom: The Impact of California's School Finance Reform on Academic Achievement and the Composition of District Spending

Co-authored with Rucker C. Johnson
Associate Professor of Public Policy
Goldman School of Public Policy
University of California, Berkeley

Section I- Introduction

While reallocating school revenues so that districts serving disadvantaged children can offer equal (or superior) educational quality to that of districts serving more advantaged children, the economic and policy literature is divided as to whether increases in per-pupil spending will meaningfully bolster student achievement. Some high-profile studies have found weak relationships between per-pupil expenditures and student achievement (Coleman et al., 1966; Hanushek, 2003), yet these have failed to identify causal relationships because they do not sufficiently account for selection bias, the demographic composition of schools, and family background factors that influence both school spending and student achievement.

Several studies have more plausibly isolated the causal effect of increased school spending on student success by analyzing major school finance reforms, both legislative and judicially mandated. These reforms increased expenditures per pupil and, consequently, narrowed gaps in performance on standardized tests, increased high school graduation rates, and bolstered adult success in the labor market (Candelaria & Shores, 2015; Card & Payne, 2002; Jackson, Johnson, & Persico, 2016; Johnson and Jackson, 2017; Lafortune, Rothstein, & Schanzenbach, 2015). While this body of research reveals a clear link between per-pupil spending and student outcomes on average, there is some evidence that historically disadvantaged students are not the prime beneficiaries of many school finance reforms. This is primarily because reforms have typically sought to equalize spending across levels of district property wealth, which is only partially correlated with student-level disadvantage (Hoxby, 2001; Hyman, 2013; Lafortune, Rothstein, Schanzenbach, 2015).

Moreover, many reforms transferred considerable fiscal power to state governments, that could place restrictions on how district revenues could be spent. It is unclear how efficacious this strategy is in enhancing student achievement. While several studies have addressed the impact of particular categories of funding, such as capital improvements (Cellini, Ferreira, & Rothstein, 2010) or pedagogical technology (Leuven, Lindahl, Oosterbeek, & Webbink, 2007), there remains a dearth of well-identified studies of the effects of restricted funding generally.

California's recent major school finance reform, the Local Control Funding Formula (LCFF) signed into law in 2013, provides an opportunity to separately test for the effects of 1) a substantial change in the levels of funding, and 2) the extent of restrictions on school financial resources, within a policy directed specifically at disadvantaged students rather than district property wealth. The LCFF, detailed in the next section, is both a massive investment in districts serving disadvantaged students and a modest relaxation of restrictions on district expenditures. It is California's attempt to overcome decades of legal and economic turmoil that had placed the

state's average district revenues, just prior to the policy change, among the nation's lowest (California Budget Report, 2017). The policy reallocates district revenues based almost entirely on the proportion of unduplicated disadvantaged students in each district -- those who qualify for free- or reduced-price lunch, have limited English proficiency, are in foster care, or are homeless. Moreover, the state relinquished many of the restrictions on how districts could spend their revenues, creating a great deal more flexibility for some districts but not others. Given the magnitude and heterogeneous nature of changes to school finance that resulted from LCFF, the policy provides a test of how financial resources and flexibility can each shape student achievement.

For cohorts born between 1990 and 2000, we constructed a school-by-cohort-level panel data set of per-pupil spending, high school graduation rates, and student achievement in high school in math and reading, for all public schools in California. Through an analysis of the first four years of LCFF (2013 through 2016-17), we show LCFF significantly increased per-pupil spending and the state now has among the most progressive funding formulas in the country. This study is among the first to provide evidence of LCFF's impacts on student outcomes. Our research also contributes to the school finance reform literature by providing the first estimates of how simultaneous changes in the levels and restricted nature of state funding impact academic achievement, the composition of district spending, and markers of teacher quality.

The context of our study enables us to provide fresh evidence of the impacts of a school finance reform explicitly targeted toward disadvantaged students as opposed to district property wealth, and to explore how greater financial flexibility impacts district spending and student achievement. Using school and district fixed effects as well as dynamic changes in the decade leading up to the policy, we leverage the heterogeneous, abrupt changes in funding induced by LCFF in a two-stage least squares event-study framework. We present event-study figures that show no evidence of pre-existing time trends in student outcomes (conditional on controls), which supports the validity of the research design.

Our empirical framework hypothesizes that the effects of LCFF on student outcomes are a function of 1) the number of school-age years of "exposure" to the policy (which takes into account the staggered rollout of LCFF wherein it did not first became near-fully funded until the 2015-16 school year); and 2) the district-specific "dosage" students are exposed to (when fully funded), which is captured primarily by the LCFF-induced increase in district per-pupil revenues (based on the funding formula parameters) and, secondarily, by the corresponding reduction in the proportion of funding subject to restrictions on how revenues can be spent. If California's new school finance policy has causal beneficial impacts on student outcomes, we expect to find a dose-response relationship with outcomes improving more for students who experienced greater school-age years of "exposure" and larger spending increases ("dosage"), respectively. Our research design employs an instrumental variables approach, using the LCFF funding formula as instruments, to isolate the effects of increases in district per-pupil spending on student outcomes. The empirical strategy compares changes in average student outcomes across cohorts from the same school before and after LCFF-induced changes in district per-pupil revenue (over and beyond statewide, cohort-specific time trends). We simultaneously account for potential impacts of releasing funding from restrictions on how it is spent, using a district's pre-LCFF reliance on restricted funding as an instrument for the proportion of district revenue that is subject to restrictions (interacted with post-LCFF years). In this way, we are able to jointly test the impact of increases in per-pupil spending and impact of greater district discretion in how it is spent, independently of one another, in the same model.

To preview the results, we find that LCFF-induced increases in school spending led to significant increases in high school graduation rates and academic achievement, particularly among poor and minority students. A $1,000 increase in district per-pupil spending experienced in grades 10-12 leads to a 5.9 percentage-point increase in high school graduation rates on average among all children, with similar effects by race and poverty. On average among poor children, a $1,000 increase in district per-pupil spending experienced in 8th through 11th grades leads to a 0.19 standard-deviation increase in math test scores, and a 0.08 standard-deviation increase in reading test scores in 11th grade. These improvements in high school academic achievement closely track the timing of LCFF implementation, school-age years of exposure and the amount of district-specific LCFF-induced spending increase (and are independent of the effects of changes in the proportion of funding that is subject to restrictions).

The remainder of the chapter is organized as follows. Section II provides greater detail about the Local Control Funding Formula, followed by a section highlighting prior related studies. Sections IV and V describe the data and detail our identification strategy. Section VI presents the descriptive patterns and regression results. Section VII concludes with a summary discussion of the findings, policy implications, and directions for future research.

Section II – The Local Control Funding Formula

Beginning in 2013, the Local Control Funding Formula (LCFF) reallocated significant state funding to disadvantaged districts while also releasing a great deal of that funding from restrictions on how it could be spent. The policy was California's attempt to replace a highly centralized and complex school finance system, laden with myriad categorical funding programs directed to specific purposes, with a system that is simple, transparent, and (proponents claim) more equitable (Wolf & Sands, 2016).

Several landmark events help explain the state's school finance landscape at the dawn of the LCFF era. California is notable for being the first state in which reform advocates, pressing for more equity in school finance, prevailed in a state supreme court and for subsequently enacting one of the most stringent cross-district equalization plans (Sonstelie, Brunner, & Ardon, 2000, pp. 33–65). This momentous state Supreme Court decision of 1971 was followed seven years later by a major tax revolt, with voters overwhelmingly approving severe limits on property tax increases (known as Proposition 13). Accelerating what was already a national trend (Corcoran & Evans, 2015), Proposition 13 dramatically increased the state's role in funding California's schools. A decade later, voters also approved a proposition requiring the state to spend a particular percentage of the state budget on K-12 schools. None of these events helped to shield school district revenues from the impacts of the Great Recession, during which district budgets dropped by 20 percent over two years, a fall from which they had not meaningfully recovered in 2012, just prior to LCFF.

Allocation in the pre-LCFF system was achieved by the state supplementing local property taxes in order to bring each district up to a "revenue limit," a mostly uniform per pupil funding allotment that depended on state economic conditions, with some differentiation allowed for certain purposes but little explicit weighting for student demographic characteristics. For districts whose property tax wealth was insufficient to meet the revenue limit, the state complemented local property taxes until the limit was reached, so that funding was equalized across such districts despite changes to local tax revenues.

Concurrently, concern that the state's stringent finance equalization policy would harm districts facing higher operation costs led to the expansion of "categorical" aid programs that

directed resources to particular expenditure categories such as transportation and special education (Weston, 2011, pp. 7–14). From the early 1980s to 2008, the state created 90 such categories through which it allocated roughly a quarter of its school revenues. Timar (1994) attributes the steady growth in these restricted categories to political patronage rather than district need, while others have criticized restricted revenue as orthogonal to student achievement (Kirst, Goertz, & Odden, 2007) and potentially stifling to innovation (Grubb, 2009). A 2009 policy suspended the lion's share of these restrictions in lieu of increased revenue, but many remained through 2012. These categorical programs did not count toward districts' revenue limits.

California's major shift in school finance reform, which first took effect in the 2013-2014 school year, replaced revenue limits with LCFF base funding differentiated by grade span, and it requires the student-to-teacher ratio in the early elementary grades (K-3) to not exceed 24 to 1 once LCFF is fully implemented. The three core components of the LCFF are (1) base grant, (2) supplemental grant, and (3) concentration grant. There is a guaranteed minimum equal to the amount received in 2012-13, adjusted for changes in average daily attendance (ADA) and local revenue. Roughly 100 "Basic Aid" districts have local revenue per pupil in excess of LCFF targets and receive nothing. For a few districts, there is also "economic recovery target" funding to restore pre-recession funding levels. Roughly 10 percent of state funding is outside the LCFF in the form of special education, Home-to-School Transportation and Targeted Instructional Improvement block grants, and school lunches.

The LCFF Base Grants establish a uniform grant that is based on average daily attendance (ADA) and varies by grade level. These grade-specific grants are adjusted for meeting K-3 class-size requirements (10.4%) and to support 9-12 college/career standards (2.6%). These 2015-16 base grants per ADA (including adjustments) were as follows: $7,820 for grades K-3, $7,189 for grades 4-6, $7,403 for grades 7-8, and $8,801 for grades 9-12.[12]

Under LCFF, in addition to the uniform per-pupil base grant that depends only on grade-level enrollment proportions (K-3, 4-6, 7-8, 9-12), school districts receive a per-pupil supplemental grant that is a weighted function of student demographics in the district, and a concentration grant for districts with a high proportion of disadvantaged students. The official formula for district *d* is given in equation 1 below:

$$F_d = 0.2 \times G_d \times H_d + 0.5 \times G_d \times max[H_d - 0.55, 0] \qquad (1)$$

where G_d is the per-pupil base grant given by the state and H_d is the unduplicated proportion of disadvantaged students: those eligible for free or reduced-price lunch, with limited English proficiency, in foster care, or homeless.

The state's allocation of Supplemental and Concentration Grants is the focal point of our use of the funding formula to isolate exogenous changes in district-level revenue caused by the state policy changes. As noted, LCFF defines high-need ("unduplicated") pupils as free-or reduced-price lunch eligible, English learners, and foster youth. The Supplemental Grant is 20% of the base grant X the high-need share of enrollment. The Concentration Grant is 50% of the base grant X the high-need share of enrollment above 55%. Concentration grants begin when a district has 55% or more high-need students.

This creates a nonlinear formula with a kink at 55% disadvantaged, which we will exploit as an alternative identification strategy (2SLS-IV regression kink design) to tease out the effects of LCFF-induced increases in per-pupil spending among students in high-poverty schools (as

[12] Source: http://www.cde.ca.gov/fg/aa/pa/pa1516rates.asp

discussed in detail in Section VI). Figure 1a shows the impact of student disadvantage on total LCFF per-pupil funding and figure 1b shows the same for the supplemental and concentration grants per pupil. For the latter, the slope is $1600.6 when the proportion of disadvantaged students in the district is between zero and fifty-five percent, then jumps to $5323.5 when that proportion climbs above 55%. This should be considered the intended increase, as the state initially lacked the resources to fully fund districts at their target level. The discrepancy between target and realized funding dropped precipitously over the first three years of the policy, by which time most of the target funding had been secured. Figure 2 (top left-hand panel) contains the distribution of disadvantage across districts in the 2012-2013 school year, wherein 62% of the students in the median district were classified as disadvantage (the corresponding student-enrollment weighted distribution is shown in the bottom left-hand panel). The distribution of LCFF funding four years later, when the policy was nearly fully funded, is shown in figure 2 (top right-hand panel shows distribution across districts, and bottom right-hand panel shows corresponding student enrollment-weighted distribution). The median district received $9,192 per pupil in unrestricted LCFF revenue from the state, representing roughly 75% of total funding.

State regulations require concentration grants be used to "increase or improve" services for high-need pupils "as compared to services provided to all pupils". Districts with 55% or more high-need students may spend these resources districtwide. If a school serves 40% or more high-need students, resources can be expended school wide. However, the district's Local Control Accountability Plan (LCAP) must identify these services and how they are principally directed to high-need students.

The new, dramatically overhauled system of school finance mandates that each district devise a Local Control Accountability Plan (LCAP), which is akin to the recipe and ingredients they will use to prepare a nutritious, equitable learning meal for every student beginning in preschool through 12th grade (high school graduation). But, there are minimal reporting requirements in Local Control Accountability Plans (LCAP). Some have expressed concerns that LCFF's granting greater autonomy over what services and programs the new funding supports will result in a set of unintended consequences. This alternative view posits that, without sufficient accountability, the "no-strings-attached" provision will result in money not reaching the students in greatest need; and, according to this view, the new funding will be allocated inequitably toward more affluent students and schools within districts. This is a long-standing debate between the advantages and disadvantages of a fiscally-centralized funding system that accounts for heterogeneous local schools and student needs. This study informs that debate.

It has been argued that a "one-size-fits-all" approach to school funding constrains local innovation and hampers the efficient use of resources to maximize student performance. K12 school leaders have long advocated for fiscal sovereignty, rather than categorical restrictions, that allows the tailored use of resources that best meets local needs and improves student outcomes. High-needs students are the fastest growing group of children in California and across the country: more than 60% of the state's public school students are low-income; more than one-quarter are English learners; concentrated poverty and segregation are widespread; and there is a large achievement gap by race and class.

The architects of LCFF reasoned that the answer to this rapid growth in pupil needs requires the interdependence of having more money *and* having greater autonomy over how funding is distributed to meet those needs. To address the deleterious effects of concentrated poverty, the funding formula allocates concentration grants to school districts with more than 55

percent of high-needs students (as defined by low-income, foster-youth, or English learners). The LCFF, which went into effect in the 2013-14 school year with a multi-year phase-in period, replaces the complex web of regulations and rules with a more transparent and progressive school funding system.

The second major component of the LCFF policy is the removal of restrictions from nearly all sources of state funding. Unlike the state's pre-LCFF basic funding allotment, an increasing list of categorical aid programs was not equalized across districts. Just prior to a 2009 reform, the state distributed over 20% of district revenues through approximately 60 categories. Many of these supported highly specific, sometimes voluntary school programs, such as counseling for grades 7 through 12, class size reduction for grade 9, incentives for physical education teachers, oral health assessments, and school library improvements (Weston, 2011). In response to severe financial strain due to the recession, the state enacted "Categorical Flexibility", a policy that suspended a great deal of restrictions starting in the 2009-2010 school year. Between that year and the year just prior to LCFF (2012-2013), approximately 12% of state funding was subject to categorical restrictions. That figure was cut to 9% by the third year of LCFF (2015-2016), as can be seen in figure 4. Though LCFF nominally removed a large number of categorical programs, in reality the bulk of these programs had been suspended in 2009 and were no longer relevant. Consequently, LCFF's impact on restricted funding was far more modest than the prior policy.

These general trends toward fewer restrictions and increased funding had different implications for districts across and within levels of student disadvantage. For example, Tamalpais Union High serves an extraordinarily affluent suburban area north of San Francisco. Because of its comparatively small population of disadvantaged students and low reliance on restricted revenue sources, Tamalpais Union High witnessed few changes in the levels of or restrictions on its funding over the LCFF period. Conversely, Compton Unified, which serves a lower-income city south of Los Angeles, witnessed large increases in state funding as well as substantial reductions in restrictions on how the funding could be spent. Other districts, such as Fremont Unified in the San Francisco Bay Area and Kerman Unified in the Central Valley, saw one aspect of its funding change markedly but not the other. As figure 5 illustrates, these four districts are not extreme outliers and their experience over the LCFF period exemplifies the variation in changes brought about by the policy's two main components.

California's new accountability context is also an important, potentially relevant element of the LCFF policy. The Local Control and Accountability Plan (LCAP) replaces the state's centralized accountability system with one that relies on individual district accountability plans, written to address specific goals in promoting student achievement.

The way a school reform rolls out is an important facet of the policy design. Traditionally, reforms roll out incrementally over time, which allows for the manifold adjustments (from personnel to curriculum) to be made at the local and district levels. However, while more immediate dispersal of funding is attractive, it often precludes a district's ability to enact bold, transformative curricular reform that can span a decade amid the constant uncertainty of available funding from year-to-year. This circumstance is typical for many districts, but particularly common for urban and low-income districts. Such fiscal uncertainty in a district is similar to the instability families that live paycheck-to-paycheck experience, which leads to suboptimal investments, rather than sustained, high-quality investments that lead to continual improvement. LCFF aims to change all of this with a $18 billion commitment in increased state support over 8 years.

Section III- School Finance Reforms

Despite the seemingly simple proposition that increasing funding to school districts will enhance the educational achievement of students, a long history of education finance scholarship suggests otherwise. Starting with the foundational Coleman Report (Coleman et al., 1966), observational work has routinely failed to find a meaningful correlation between school expenditures and student achievement. This massive study of school resources and student performance, undertaken in response to the Civil Rights Act of 1964 to assess racial disparities in schools, surveyed over 639,000 students, teachers, and principals in a representative sample of schools in the United States in 1965. While the report found substantial achievement disparities across racial groups and within racial groups across schools, very little of the variance in these measures of achievement could be accounted for by school resources. These early cross-sectional results have been reflected in aggregate time series comparisons, where National Assessment of Educational Progress (NAEP) scores only slightly increased since the 1970s despite substantial concurrent increases in school resources (Hanushek, 2003). Debates over this basic finding, that student achievement varies considerably across schools and teachers but not because of identifiable resources, have been a mainstay of education policy research for the past 50 years (Burtless, 1996; Goldhaber, 2015; Hanushek, 2011; Hanushek, Rivkin, & Taylor, 1996). A common criticism of this literature is that observational studies on the link between school resources and student achievement lack causal warrant due to the endogenous selection of students into schools and the potentially compensatory nature of school finance; compelling evidence on the impacts of school resources should come from field or natural experiments (Murnane & Willett, 2011, pp. 5–7). This challenge of identification has motivated a focus on the impacts of school finance reforms, where sharp changes in funding have been imposed on districts in a more plausibly arbitrary manner (Jackson, Johnson, & Persico, 2015).

Proponents of state-level school finance reforms seek to redress the disadvantage in school resources many children face due to the historic reliance of school districts on local revenues (P. L. Howell & Miller, 1997, p. 42; C. M. Hoxby, 1996, p. 69). Because of vast differences in local jurisdictions' wealth and preferences for education spending, children face a substantial amount of inequality in school resources across states and districts. For example, the ratio of the 95th percentile of district-level per-pupil spending to the 5th percentile was 2.73 in the early 1970s, the dawn of the first major school finance reform era. This relationship fell for decades, hitting a nadir of 1.98 in 2000, then climbed back to 2.55 by 2011 (Corcoran & Evans, 2015, p. 358). Though residential segregation across racial and socioeconomic lines has increased since the early 1970s (Clotfelter, 2004, Chapter 3), between-district inequality is largely a function of household incomes and property wealth, so that the largest inequalities across student demographic groups are based on district-level household income averages rather than individual student poverty status or ethnicity. In the early 1970s, expenditures per pupil were 1.4 times higher for pupils in wealthy districts than those in poorer districts, but only 1.08 times higher for non-poor vs. poor students and only 1.02 for whites vs. non-whites (Corcoran, Evans, Godwin, Murray, & Schwab, 2004, p. 440). Successive waves of school finance reforms attenuated these figures but did not change their ordinal relationship- average wealth in the district is still a better predictor of between-district disparities than individual student demographics.

Court-mandated and legislated school finance reforms have sought to either equalize or ensure an adequate level of school resources (Koski & Hahnel, 2015). The "equity era" began

with the California Supreme Court's 1971 decision in *Serrano v. Priest* that the current system of local school finance ran afoul of constitutional guarantees of equal protection, a judicial victory for what reform advocates labeled "Proposition One": the quality of a child's schooling should not be a function of wealth within a state (Springer et al., 2015, p. 10). Despite a subsequent setback in the United States Supreme Court in 1973, the ensuing equity era witnessed successful challenges to unequal district funding formulas in 10 states between 1971 and 1988 (Jackson et al., 2016). Recognizing that equality could be achieved through a mere reduction in expenditures at the top of the distribution, thus undermining their underlying goal of educational enhancement, reform advocates began litigating on the basis of educational adequacy (Clune, 1994). A major shift in strategy, this "adequacy era" was heralded by a 1989 Kentucky Supreme Court decision that the state's constitution guaranteed an adequate level of educational resources rather than merely equal resources across students. Comparable court rulings proliferated in the following decades in conjunction with similarly themed legislative decisions, with 27 states witnessing at least one school finance reform event through 2013 (Lafortune et al., 2015, pp. 66–67).

A growing body of literature has tried to assess the impact of school finance reforms on the levels and distributions of school finance and student achievement. In general, the causal identification in these studies has either leveraged reform-induced variation in funding within individual states or assumed that the onset of reforms are conditionally exogenous events in cross-state analyses, with non-reform states creating the counterfactual trends in student achievement, educational attainment, and labor market participation. The former group includes analyses of school finance reforms from the 1990s in Kentucky, Maryland, Massachusetts, Michigan, and Vermont. Guryan (2001) uses nonlinearities in a district funding formula brought about through the Massachusetts Education Reform Act of 1993 to identify the impact of state aid on district revenues and student achievement. Conditioning on a smooth function of district property wealth, Guryan uses sharp discontinuities in the state aid formula as exogenous instruments for district revenues. Within three years of reform, districts just below the state aid cutoff thresholds had spent roughly 65% of their new revenues, with concurrent increases in their 4th grade students' performance on standardized tests of math, science, and social studies. The new revenue does not appear to have affected performance on 4th grade literacy tests or any 8th grade tests in any of the four subjects. Similary, Papke (2005, 2008) uses discontinuities in Michigan's funding formula brought about by the state's 1994 reform (Proposal A) to instrument for school expenditures. Reform-induced increases in expenditures led to meaningful increases in the percentage of 4th grade students who successfully passed the state's standardized tests of numeracy, which a subsequent study reveals were concentrated in initially low-spending districts (Roy, 2011). This effect on pass rates for numeracy tests in initially low-spending districts is supported by Sherlock's (2011) analysis of Vermont's Equal Education Opportunity Act of 1997, yet that state's reform did not affect pass rates for literacy or writing. In the wake of Kentucky's 1990 reform, African-American students performed better on both the literacy and numeracy portions of the ACT, though the effect on all students is not distinguishable from zero. African-American performance on the 8th grade NAEP test appears to increase substantially (.12 standard deviations), yet the study lacks the power to distinguish this effect from zero (Clark, 2003). Moreover, reform-induced increases in expenditures in Maryland did little to increase student attainment, despite substantially reducing district spending inequality (Chung, 2015).

While these mixed results from individual state school finance reforms point to the heterogeneity in school finance systems, a concurrent literature has found mostly impressive

impacts from school finance reforms generally. In this group of studies, the causal identification comes from variation in the presence of successful school finance reforms across time and states, with controls for endogenous state characteristics that are either fixed or vary in the pre-school finance reform periods. Murray, Evans, and Schwab (1998) analyze the impact of court-mandated school finance reforms in 16 states from 1971 to 1992, finding that reforms induced an 8 to 11 percent increase in spending in the bottom half of the distribution, leading to a 19 to 34 percent reduction in inequality across districts within states. Card and Payne (2002) find that successful court challenges led to a narrowing of district inequality within states, that the share of this state revenue translated into expenditures is between 30% and 65%, and that the reforms reduced the SAT test score gap across students divided by parental education, and possibly raised the SAT participation rate in the lower education group. More recently, Candelaria and Shores (2015) find that even seven years after court-ordered reforms, per-pupil revenues and graduation rates were higher for high poverty students. Expanding the treatment to both court-ordered and legislative reforms, LaFortune, Rothstein, and Schanzenbach (2015) find a gradual reduction of the income-dependence of state-level National Assessment of Educational Progress scores. While the immediate impact of reforms is insignificant, the ten-year impact of school reforms is a tenth of a standard deviation closure of the test score gap across district income levels. The immediate impact of school finance reforms on district financial equity persists over this period as well and is not attenuated by recapture of revenues by local taxpayers. Looking at longer-run outcomes, Jackson, Johnson, and Persico (2016) and Johnson and Jackson (2017) find that court-ordered school finance reforms increase educational attainment and wages, and lead to significant reductions in both the annual incidence of poverty and incarceration in adulthood, particularly for children from low-income families.

While the Jackson, Johnson, Persico (2016) study provides important evidence of the long-run beneficial impacts of earlier-era court-ordered school finance reforms, the recent LaFortune, Rothstein, Schazenbach study focuses on more recent school finance reforms during the adequacy era and documents significant impacts on test scores. If more recent school finance reforms have different effects on student learning, it is valuable to learn about the effects of the LCFF, even given the compelling results in Jackson, Johnson and Persico. As a whole, the school-finance-reform literature suggests that, *on average*, redistributive school finance reforms result in increased resources to less wealthy districts and enhanced achievement for students residing in those districts in both their academic careers and subsequent labor market outcomes.

A subsequent analysis of Michigan's reform illustrates factors that might mediate the impact of school finance reforms. Using a modification of the strategy in Papke (2008), Hyman finds that district revenues increased by only 58 cents of each dollar of increased state aid, pupil-administrator ratios decreased while pupil-teacher ratios did not, and the extra revenue was targeted to comparatively affluent schools. Though the reform ultimately led to increases in post-secondary attendance and degree attainment, these effects were also larger for comparatively affluent students. The relatively weak flypaper effect, potentially inefficient use of new revenue, and within-district allocation favoring affluent schools present policy challenges but also frustrate the ability to draw strong inferences from school finance reforms regarding the causal link between school spending and student achievement.

The inefficient use of new funding is a particular salient issue. Skeptics claim that school administrators are not sufficiently incentivized to spend money in an efficient manner (Hanushek & Raymond, 2001, p. 381), which suggests that funding restrictions might increase student achievement if the public can identify and enforce more productive uses of resources. Many

scholars and policy analysts that have supported increased school spending have noted that student achievement might be enhanced by systems that help ensure spending is allocated toward the most productive uses. However, in the case of California, increased restrictions on state funding have not been particularly effective. In the aftermath of severe finance reform in the 1970s, the state began making categorical funding available for specific purposes to acknowledge that schools face different costs to produce the same good (Weston, 2011). Over time more funding was directed to categorical programs, even as the state general revenues declined in periods of recession (Sonstelie et al., 2000, pp. 59–62). The growth in these categorical funds has been attributed to political patronage (Sonstelie et al., 2000, pp. 63–64) and the desire to circumvent collective bargaining agreements (Kirst et al., 2007, pp. 7–8), rather than student need (Timar, 1994). There is a relative dearth of studies on the causal impact of restricted funding on student achievement. However, two recent, well-identified studies of restricted spending increases are instructive. Cellini, Ferreira, and Rothstein (2010) assess the impacts of increased capital outlays by exploiting the sharp discontinuity in available capital improvement bonds due to local election results. The impacts of capital outlays on third grade standardized test scores are small and statistically indistinguishable from zero for most years in the 15-year period following a successful bond measure.[13] That test scores do not respond to capital improvement may not come as a great surprise, yet evidence from the Netherlands suggests that increased funding restricted to personnel and technology fares even worse. Leuven, Lindahl, Oosterbeek, and Webbink (2007) analyze the impact of two Dutch subsidies, one for personnel and one for computers, by exploiting the fact that only schools with over 70% disadvantaged students qualified for the increased funding. The results are dispiritingly negative across multiple specifications and achievement tests for both subsidies.

Three details of California's LCFF, together with the state's policy context, help provide an informative test of the impact of a school finance reform on district expenditures and student achievement. First, the state's school finance and property tax systems place severe constraints on the ability of local tax payers to influence their districts' revenue (Timar, 2006). Second, rather than re-allocating revenues across district property wealth, LCFF distributes funds on the basis of student disadvantage. Hoxby (2001) cautions that state responses to school finance reforms vary widely, with dramatically different consequences of state policy choices on education funding. For example, state aid formulas that distribute revenue to districts based on districts' local property tax wealth are endogenous to school finance and student achievement. These formulae can encourage an overall reduction in district revenues, potentially making the least-advantaged students worse off. Moreover, because district-level property wealth and student-level disadvantage are imperfectly correlated, a redistribution of revenue based on district wealth will imperfectly target disadvantage students. A particularly instructive example comes from the LaFortune, Rothstein, and Schanzenbach (2015) study, which reveals that the equalizing effect of school finance reforms on NAEP scores is not present across student-level racial or income gaps.

A third feature of LCFF affords an evaluation of the efficacy of state restrictions via categorical aid. California's recent reform gives districts a modest increase in discretion over expenditures by transferring revenue out of categorical aid programs and into the basic funding that districts receive from the state. This increase in fiscal freedom was felt heterogeneously

[13] The authors find that housing prices increase substantially and persistently after successful bond measures, suggesting that homebuyers value the new capital funds despite their weak relationship to student literacy and numeracy.

across districts and thus allows for a joint test of the independent effects of increasing both per-pupil revenues and budgetary discretion. While the recent economics literature provides a fuller picture of the impacts of school finance generally, comparatively little is known about the impact of restrictions on how money can be spent.

Section IV- Empirical Strategy

The primary empirical challenge in estimating the effects of school spending on student outcomes is that spending and the quality of schools tend to be highly correlated with child family and neighborhood socioeconomic factors, due to the combination of parental choices and residential location constraints (e.g., zoning policies and availability of affordable housing) that sort more advantaged children into better quality schools. Compensatory spending reforms may understate the effects of increased funding on student outcomes if the pre-existing student disadvantage that funding is targeted toward is not fully taken into account.[14]

Our research design employs an event study in combination with a simulated instrumental variables approach to circumvent this challenge using the LCFF funding formula and timing of implementation to isolate exogenous changes in district per-pupil revenue and promised availability of this funding from the state in future years as well. Our simulated instrumental variables (IV) approach for supplemental/concentration grants uses the following three funding formula parameters that determine funding: the baseline percentage of high-need students in the district (H_d); the district's base grant (G_d); and the formula that allocates additional funding based on pupil needs in a given district. These three funding formula parameters are used to construct our instrument ($\boldsymbol{Z_d}$):

$$Z_d = 0.2 \times G_d \times H_d^{2013} + 0.5 \times G_d \times max\left[H_d^{2013} - 0.55, 0\right]$$

Importantly, these reform-induced changes in district spending, which are credibly identified from the funding formula (and which serve as the instrumental variables), are unrelated to changes in child family and neighborhood characteristics conditional on the baseline level of disadvantage in each district. We refer to this reform-induced change in district per-pupil spending from the state as the "dosage" (in the parlance of the medical and treatment effects literature), which is district-specific. The "dosage" amount here refers to the LCFF fully funded amount. High-poverty districts are high-dosage and those with small proportions of disadvantaged students are low-dosage in accord with the funding formula.

We refer to "exposure" as the number of school-age years a child was exposed to the LCFF policy, which is birth cohort-specific, recognizing there is a phase-in period of implementation toward the formula being fully funded. For example, in models of high school graduation rates, cohorts born before 1996 are "unexposed" cohorts, as they had already reached age 18 prior to LCFF's enactment.

LCFF established a multiyear phase-in timeline to incrementally close the gap between actual funding and new target levels of funding. The research design explicitly accounts for this through the estimation of fully non-parametric event-study models that show the evolution of school inputs and student outcomes in both the years before and after the law's implementation separately for "high- and low-dosage" districts.

---

[14] This point has been illustrated Johnson, R. C. & Jackson, C. K. (2017). "Reducing Inequality Through Dynamic Complementarity: Evidence from Head Start and Public School Spending". NBER working paper #23489 and Jackson, C. K., Johnson, R. C., & Persico, C. (2016). The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms. *Quarterly Journal of Economics*, 131(1).

For this purpose, for cohorts born between 1990 and 2000, we constructed a school-by-cohort-level panel data set of school-age years of per-pupil revenue, high school graduation rates, and student achievement in high school in math and reading, for all public schools in California. These data are matched with LCFF school-reform variables. This chapter focuses on high school graduation rates (the four-year cohort rate, which is consistently measured since 2009 for public schools in California), as well as high school achievement using 11th grade mathematics and reading standardized test scores (that are NAEP-norm adjusted as discussed in Section V).[15] Our analysis excludes charter schools.

Our inclusion of school fixed effects accounts for all time-invariant school-level factors, and the inclusion of birth-year fixed effects accounts for statewide trends in outcomes. Thus, factors such as persistent differences in teacher quality across schools, and statewide changes in economic conditions, are not a potential source of bias. The empirical strategy effectively compares changes in average student outcomes across cohorts from the same school before and after LCFF-induced changes in district per-pupil revenue (that exist over and beyond year/cohort-specific average changes over time).

If school spending has causal effects on student outcomes, we expect to find patterns of results that increase in both "dosage" (i.e., the amount of spending change) and the number of school-age years of exposure.[16] This dose-response relationship is indeed the pattern of results we find and document in this chapter (Section VI). Importantly, we find no corresponding evidence of pre-existing time trends, which supports the validity of the research design to detect causal impacts. By 2016 (the most recent year for which data is presently available), the maximum number of school-age years of exposure is four, since the first year of enactment is during the 2013-14 school year. Target levels approached fully-funded status in the 2015-16 school year. So, for example, the high school graduating class of 2016 would have been potentially exposed to LCFF throughout their high school years; and similarly, student achievement in 11th grade during the 2016-17 school year corresponds with cohorts that had been potentially exposed to LCFF since the time they entered 8th grade (albeit at nearly fully-funded levels in only the last two of those years).

We identify the impact of the levels of and restrictions on state financing to school districts by leveraging the heterogeneous, conditionally exogenous changes in funding induced by the LCFF policy. The preferred estimation strategy, a two-stage least squares event study with school and district fixed-effects,[17] is robust to both fixed and dynamic endogenous selection of students into districts and districts into financing regimes. One of the key substantive innovations in this analysis is the inclusion of two exogenous treatment variables in the second stage equation: the predicted levels of per-pupil spending (as instrumented by the funding formula) and the predicted unrestricted proportion of that funding (as instrumented by the 2012 (pre-LCFF) proportion reliance on restricted funding). The ability to separate the effects of a

[15] NAEP adjustments follow procedures outlined in Reardon, S.F., Kalogrides, D., & Ho, A. (2017). Linking U.S. School District Test Score Distributions to a Common Scale (CEPA Working Paper No.16-09). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp16-09.

[16] A similar research design and empirical setup to identify the causal effects of K12 spending is used in Johnson, R. C. & Jackson, C. K. (2017). "Reducing Inequality Through Dynamic Complementarity: Evidence from Head Start and Public School Spending". NBER working paper #23489 and Jackson, C. K., Johnson, R. C., & Persico, C. (2016). The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms. *Quarterly Journal of Economics*, 131(1), which examine earlier era court-ordered school finance reforms.

[17] District fixed-effects are used where district finances are the outcome of interest; school fixed-effects are used where student achievement and high school graduation rates are the outcomes of interest.

per-pupil spending increase from the effects of a decrease in restrictions within the same model and policy environment is unique, as most prior studies have focused on either the impact of per-pupil spending increases (and equalizations) or the impact of increases in funding restricted to a particular purpose. This separation is achieved with the use of two separate sets of instruments: the formula weights (derived from a district's proportion of disadvantaged students) for per-pupil spending; and the proportion of each district's funding subject to restrictions in the year before the policy (2012) for the restricted proportion of funding. We however, show that, the main patterns of effects of per-pupil spending on student outcomes is similar with and without accounting for the proportion of revenue that is unrestricted.

As aforementioned, the key challenge of causal inference is to isolate the impact of the SFR policy changes as distinct from pre-existing trends and other coincident changes that may also affect graduation rates or any of the other outcomes of interest. Figure 6 provides visual evidence that the formula, the basis of state funding allocation, does not predict changes in funding levels in the four years leading up to the policy change. The linear fit is flat in the left panel, indicating that changes in funding levels from 2009 to 2012 are not predicted by the precise demographic weights used in the LCFF. The right panel is the same graph for the actual LCFF period, in which a distinct linear increase is visually detectable, which confirms that, at least in a simple multi-year change analysis, LCFF has increased district revenues via the formula. Figure 7 contains only the linear fits from each panel of Figure 6 together in the same graph for clarity of comparison. Table 1 contains the simple linear regressions that undergird the panels in Figures 6 and 7. As can be seen, district formula does not predict linear changes in district finances in the years leading up to LCFF. The coefficient on formula should be interpreted as the predicted impact of moving from a formula of zero, corresponding to having no disadvantaged students, to a formula value of one, which would be more than double the actual highest formula possible: .425. Accordingly, the statistically insignificant coefficient in the first column should be read as a precise zero, which the observed R^2 value of .0004 reinforces. Conversely, the second column reveals a strong, statistically significant relationship between district formula and post-LCFF changes in funding, providing simple evidence of the conditional exogeneity of the formula.

Though the policy treatments contained in LCFF were outside of districts' control and appear at least upon graphical inspection to be exogenous, it might still be the case that graduation rates would have changed during the LCFF era due to continued economic growth of the state (or other temporally correlated events), irrespective of the changes in district finances. Periods of recession and economic growth can have different impacts on district revenues and graduation rates across the spectrum of district-level disadvantage. California's economy continued to expand in the LCFF era as the state recovered from the Great Recession. If graduation rates in less-advantaged (higher formula) districts are more responsive to the economy than are rates in advantaged (lower formula) districts, then a positive correlation between district revenues and graduation rates would be partially due to this structural economic relationship and not the new LCFF revenues. An analogous story could apply regarding categorical restrictions on district revenue.

This analysis addresses such structural economic relationships by controlling for the association between the policy treatments and graduation rates (and all other outcomes) that can be predicted by the decade of California's economic performance leading up to LCFF, 2004 to 2012. This period covers the pre-recession housing bubble, the housing crash and ensuing recession, and the recovery, providing sufficient variation in economic performance with which

to predict district finances and graduation rates. A simple time-trend analysis of state funding reveals dramatic, non-linear changes over these years. Figure 3 shows a steady increase in funding during the housing bubble years (2004-2006), followed by a reduction in the recession years (2007-2009) that continues into the years prior to LCFF (2010-2012).

To model this role of business cycle fluctuations as it most closely relates to school finance, we use non-K-12 expenditures by the state of California, both overall and to all local sources. Statewide expenditures, rather than statewide revenues, are used because school districts' revenues are a function of what the state spends, which, because of smoothing over time, is not perfectly correlated with state GDP, tax receipts, or other revenue sources. We interact this state expenditure variable with a complete set of district fixed effects in order to obtain a district-specific relationship between statewide expenditures and district per-pupil revenue, and likewise for state local assistance provided, excluding education[18]; and include linear time trends and interact them with the funding formula. That is, for the pre-LCFF period (1995-2012), we regress district per-pupil revenue on the full set of interaction terms and district fixed effects, as in equation (2):

$$C_{db} = \alpha + \beta_d CAExp_b + \beta_d CALocalAssist_{db} + \theta timetrend + \gamma Formula_d * timetrend + \mu_d + \varepsilon_{db} \qquad (2)$$

where C_{db} is the district per-pupil revenue from the state for district d for birth cohort b; $CAExp_b$ is the total non-K-12 state expenditures per pupil for birth cohort b; $CALocalAssist_{db}$ is state local assistance provided (excluding education); $Formula_d$ is the LCFF funding formula parameter for district d; and μ_d is a vector of district fixed effects; ε_{db} is a stochastic error term for district d for birth cohort b. These models are run for the years 1995 through 2012, just prior to LCFF, and then used to predict the level of district per-pupil revenues from state sources for all years in the data, including the post-LCFF era (i.e., 2013 through 2016-17). We then take the predicted average of C_{db} during ages 15-17 to include in the regression models as controls.

Thus, in our models we account for these other potential district-level changes that are not driven by LCFF, with the inclusion as an additional control variable, the predicted district per-pupil revenue from the state ($C_{db}^{age15-17}$), based on prior funding and state-wide California spending on non-K12 expenditures (based on pre-LCFF district-specific relationship between prior funding variables and district revenue from the state). This is an estimate of the counterfactual district revenue from the state if LCFF had not occurred. As shown in Figure 8, the prediction closely matches the actual average level of revenue in all years prior to LCFF; the significant departure of actual average revenue from its average prediction in the post-LCFF years (as expected) is plausibly attributable fully to the new LCFF formula. Including this in the primary regressions controls for dynamic, district-specific relationships between changes in economic conditions and district finances.[19]

Figure 9 shows the evolution of district per-pupil revenue from the state before and after LCFF for high-poverty (large spending increase) and low-poverty districts (small spending increase). In these figures, a "high-poverty district" receives $2,500 per-pupil revenue from the state when LCFF is fully funded, whereas a "low-poverty district" receives $500 per-pupil

[18] Total state expenditures, excluding public K-12 spending, covers categories such as health and human services, transportation, and the department of corrections. Total local assistance, excluding public K12 spending, covers categories such as medical assistance programs and social services. Both variables are adjusted to real 2015 dollars, and divided by the total state K-12 enrollment in each year.

[19] The results are similar with and without this additional control.

revenue (in accordance with the funding formula). This evidence that finds no pre-existing time trend also further supports the research design's ability to uncover causal effects.

The full first stage models are presented below in equations (3) and (4).

$$\widehat{ppe}_{db}^{15-17} = \pi_1\left(SFRExp_{db} \times \widehat{dose}_d\right) + \pi_2(SFRExp_{db} \times \widehat{unrs12}_d) + \gamma_1 \cdot C_{db} + \theta_{d1} + \tau_{b,1} \quad (3)$$

$$\widehat{unrs}_{db}^{15-17} = \pi_3\left(SFRExp_{db} \times \widehat{unrs12}_d\right) + \pi_4\left(SFRExp_{db} \times \widehat{dose}_d\right) + \gamma_2 \cdot C_{db} + \theta_{d2} + \tau_{b,2} \quad (4)$$

$\widehat{ppe}_{db}^{15-17}$ is average per-pupil revenue from state (in real 2015 dollars) during expected school-age years (ages 15 through 17) in an individual's childhood school district, $\widehat{unrs}_{db}^{15-17}$ is average proportion of revenue from state that is unrestricted during expected school-age years (ages 15 through 17) in an individual's childhood school district, $SFRExp_{db}$ is the number of school-age years that occurred after LCFF first implemented (0 = 17 years old, 4 = 13 years old, etc.), with each year entered as dummy indicator, $\widehat{dose}_d$ is the decile of the LCFF concentration/supplement grant*spline (based on funding formula), $\widehat{unrs12}_d$ is the 2012 (pre-LCFF) proportion of revenue from state that was unrestricted, *s* indexes school, *d* indexes district , *b* indexes birth year, *g* indexes group (all kids; poor kids; or racial/ethnic group). Outside of the interactions with $SFRExp_{db}$, both $\widehat{dose}_d$ and $\widehat{unrs12}_d$ are subsumed by the district or school fixed effects. Each first-stage regression provides information on how the policy levers actually altered district finances.

Overall there is a large first-stage effect of LCFF on district per-pupil spending (using only the funding formula parameter instruments), and there is a strong first-stage relationship between the 2012 (pre-LCFF) proportion of revenue that was unrestricted on subsequent changes in the proportion of revenue unrestricted in the post-LCFF period (independent of the funding formula). Table 2 contains the *F*-statistics from the first-stage regressions of district per-pupil spending and proportion unrestricted on the respective set of instruments, which both exceed 30. The table shows that the LCFF-related instrumental variables have a strong, statistically significant relationship to the endogenous financial variables, and have sufficient independent variation to identify their respective effects.[20] As expected, the 2012 (pre-LCFF) district proportion of revenue that was unrestricted is not predictive of per-pupil spending independent of the funding formula instruments.

The second stage is represented in equation (5):

$$Y_{gsdb} = \beta_1 \cdot \widehat{ppe}_{db}^{15-17} + \beta_2 \cdot \widehat{unrs}_{db}^{15-17} + \gamma \cdot C_{db} + \theta_{sd} + \tau_b + \varepsilon_{gsdb} \quad (5)$$

where Y_{gsdb} is the outcome of interest for group *g* in school *s* in district *d* for birth year *b*, θ_{sd} is a vector of school fixed effects and τ_b are birth year fixed effects. Because we have interest in estimating potential spending effects on average student achievement at the school-level, as well as impacts on achievement gaps, we estimate models for all children and separately for poor children ("poor" is defined in this chapter as eligibility for free/reduced-price lunch), non-poor children, and by race/ethnicity. We also conduct a series of placebo (falsification) tests to ensure that the estimated effects are indeed due to the impacts of LCFF and not other coincident policy changes.

[20] We force the identifying variation in the proportion of revenue that is unrestricted to operate only through its prediction based on the 2012 pre-LCFF proportion unrestricted interacted with the post-LCFF years; and independent of and not through the funding formula parameters.

Average district per-pupil spending during ages 15-17 is inflation-adjusted using the CPI-U deflator (in real 2015 dollars) and then expressed in thousands[21], and the average proportion of district revenue that is unrestricted during ages 15-17 has been standardized, so that a one standard deviation increase is roughly 4 percentage points; in both cases this is done in order to facilitate interpretation of marginal effects and so the estimated effects are in the range we observe LCFF-induced variation in our key explanatory variables. The 2SLS-IV regressions are weighted by 2013 school enrollment. Standard errors are robust to heteroscedasticity and clustered at the district level.

Section V- Data

Our analysis relies on publicly available teacher-, school-, and district-level data from the California Education Department. Charter schools are excluded from the analysis, as are virtual and other non-traditional schools. Districts with insufficient years of data have also been removed. The final data set and analysis thus reflects traditional schools in elementary, high school, and unified school districts that have been in continuous operation in California from 1995 through 2017. Annual district financial records are available in aggregate from the standardized account code structure (SACS) unaudited actual data files from 2003 forward, prior to which the files reflect a previous accounting structure that is similar with respect to coarse revenue and expenditure categories but not fine-grained expenditures.[22] This analysis primarily uses data going back to the 1995-1996 school year up to the most current year for which all necessary data is available (2016-2017), which include the first four post-LCFF school years: 2013-2014, 2014-2015, 2015-2016, 2016-2017. Each SACS file contains data on all general ledger financial records (both expenditures and revenues) for public school districts (Local Educational Agencies) in a given year. Each entry in the data is a particular financial record for a district aggregated for each relevant combination of "account" (revenue vs. expenditure), "fund" (general fund vs. a variety of special categories), "resource" (unrestricted vs. restriction categories), "goal" (Pre-K, K-12, Adult Education), "function" (Instruction, Special Education, etc.), and "object" (detailed source and purpose information). The previous accounting structure is less detailed in some of the finance categories, so analyses of certain expenditure categories can only go back as far as 2003.

This detailed financial data is transformed into real 2015 dollars per pupil using the consumer price index and enrollment data from the district. Figure 10 presents district expenditures per pupil for four categories over the period 2004 through 2016-17: teacher salaries, administrator salaries, buildings, and employee benefits (both health and retirement). The relative portions spent on these four categories do not change dramatically over this time period, with the exception of employee benefits. The LCFF era also witnessed a sharp rise in the amount of money districts spent on employee benefits. Rather than reflecting more generous compensation packages, this increase was due to districts taking over a greater share of payments into the state teachers' retirement system. Over the same time period, the proportion of expenditures going toward instruction and teacher salaries have decreased, as can be seen in Figures 10 and 11.

---

[21] For the analyses of 11th grade math and reading test scores, we examine the impacts of average per-pupil spending during ages 13-16 (i.e., 8th through 11th grades) and the corresponding impacts of the average proportion of funding that is unrestricted during ages 13-16 for cohort *b* in district *d*.

[22] Available here http://www.cde.ca.gov/ds/fd/fd/

The state of California's overall and local expenditures data, together comprising the underlying economic variables used in this analysis, come from the monthly statements of general fund cash receipts and disbursements made available through the State Controller's Office.[23] The June monthly statement in each fiscal year contains data for the preceding fiscal year. Expenditures by the state are broken down into two major categories: State Operations and Local Assistance. State operations cover categories such as health and human services, transportation, and the department of corrections. Local assistance covers categories such as public K-12, medical assistance programs, and social services. The two variables used in this analysis are the total state expenditures and the total local expenditures, each without the K-12 spending, adjusted to real 2015 dollars, and divided by the total state K-12 enrollment in each year. The school-level enrollment (average daily attendance) data comes from the financial records.

The main component of redistribution in the LCFF period is the proportion of (unduplicated) students who receive free- or reduced-price lunch, are of limited English proficiency, in foster care, or homeless. Unduplicated counts along those demographic lines are not available prior to 2012. Including both poverty and English language learner counts, which are available in each year, would severely overstate the proportion disadvantaged in many districts. Moreover, since 2013 there is an incentive for a district to endogenously classify its students as disadvantaged, such as through an increased effort to collect a student's socioeconomic status or retaining students in the limited English proficiency category. To circumvent these data limitations, the treatment is constructed from each district's proportion of unduplicated disadvantaged pupils in the first year of the policy, the 2013-2014 school year. This is used as the district's stable proportion of disadvantaged students across all years.[24]

The two endogenous regressors for which the policy changes serve as instruments, district per-pupil spending and the proportion of that funding not subject to restriction, are constructed from the SACS data. Total per-pupil spending is defined as the total expenditures divided by enrollment, and, likewise, total per-pupil revenue from the state is defined as the total revenue from all state sources (according to the "object" codes), divided by enrollment; both adjusted for inflation to represent 2015 dollars. The proportion not subject to restriction is defined by the "resource" codes and is simply the total per-pupil district revenue from all state sources under all unrestricted codes divided by the total district per-pupil revenue from the state in each district.

The data that include markers of teacher quality come from the California Department of Education as well. The state maintains an annual file of all teaching staff in each public school containing the staff members' education level, years of experience, and years working in the district, among other variables.[25] Each staff member is given a unique code that is not consistent across years, so that the staff records can only be merged with other records (schools, class assignment, etc.) within each year. We aggregate the staff records to create the following school by year variables: mean years of experience, mean years in the district, number of teachers in the school, and proportion of teachers with a master's degree or higher.

---

[23] Available here http://www.sco.ca.gov/ard_state_cash.html

[24] The formula can vary from zero to 0.425. Note that this is the formula weight, not the raw percent of disadvantaged students. A district with 50% disadvantaged students would receive 20% * 50% = 10% additional funding.

[25] Available here http://www.cde.ca.gov/ds/sd/df/filesstaffdemo.asp

The high school graduation-rate analysis sample includes data on over 400,000 students per year in the in the 384 unified and high school districts with sufficient data across the years 2009 through 2016. We use the state's adjusted four-year cohort graduation rate, which has been available only since the 2009-2010 school year. While this figure more accurately measures high schools' performance, the lack of commensurate measures prior to 2009 means that our model is truncated for the graduation rate analysis. The veracity of certain schools' and districts' record keeping has recently been called into question (e.g., see OIG report), raising concerns that high-poverty schools are still not properly calculating the graduation rate. Though the problem was found in only a handful of schools in a single district and did not provide direct evidence of inaccurate graduation rate data, we run our models with and without schools in the highest percentiles of poverty; the main pattern of results are unchanged.

The annual school files provide aggregate data on four-year cohort graduation rates – both the number in the cohort and the number of graduates from the cohort for each year. The district-level, four-year cohort graduation rate was created by dividing the aggregate number of graduates across all traditional schools in the district by the aggregate number of students in the cohort across all traditional schools. The yearly graduate rate figures are higher than the state totals in each year because charter and non-traditional schools, which typically have lower graduation rates, have been filtered out.

In the first year of the LCFF period, the state of California suspended its STAR testing program and began using the new "Smarter Balanced" tests in the following year (Cardine, 2013), complicating longitudinal analysis of student achievement over this time period. The new testing regime's computer-based administration and content focus are sufficiently different from the material and paper-and-pencil nature of STAR testing that the superintendent of public instruction cautioned against any comparison across the two tests after the first wave of results revealed significantly lower student performance on the new test (Noguchi, 2015). To overcome this challenge, we norm both the STAR and Smarter Balanced tests to the National Assessment of Educational Progress (NAEP), which, over this time period, has not changed and has been given to a representative sample of California students biennially. We follow the procedure in Reardon et al. (2017) but extended the norming to the school-subgroup level. Each school-subgroup score in each year thus reflects standardized performance on the NAEP scale. Because this scale does not change across the analysis time period, test scores in this normed metric can be compared both before and after the onset of LCFF. This norming enables comparable measurement over time to analyze student performance and comparisons of that performance before and after LCFF changes in spending. Changes in the testing procedures that could otherwise lead to biases are also accounted for through our inclusion of year fixed effects, which pick up average year-to-year trends in student performance that may be attributable to the changes in standardized test measurement.

Section VI- Results

We focus our discussion first on the results of the impacts of per-pupil spending on high school achievement. Table 3 and figures 12 and 13 present the main results from the analysis of high school graduation rates. The first row of the table shows that a $1,000 increase in the average per-pupil spending experienced during ages 15-17 (i.e., 10th through 12th grades) increases the high school graduation rate for students overall by 5.89 percentage points, with comparable effect sizes for low-income (5.1 percentage-point increase) and Hispanic students

(5.68 percentage-point increase). This effect is strongest for African-American students, at 7.71 percentage points.

The second row of Table 3 reveals that a one standard deviation increase in budgetary flexibility[26] experienced during ages 15-17 leads to a 1.41 percentage point gain in the high school graduation rate for students overall. Similar to the impact of expenditures, this effect is most pronounced among African-American students, for whom the graduation rate increase is 2.88 percentage points. The effects for all other groups are smaller and statistically insignificant.

Figure 14 and Table 4 present the results for 11th grade math and reading standardized test scores by child poverty status. We find that a $1,000 increase in the average per-pupil spending during ages 13-16 (i.e., 8th through 11th grades) leads to a 0.19 standard deviation increase in math and a 0.08 standard deviation increase in reading for poor children. The same increase leads to a 0.08 standard deviations in reading for non-poor children, for whom no impact on math achievement is detectable.

Table 5 presents similar results by ethnicity. Hispanics comprise 54% of California's public school children, and 24% of schoolchildren are non-Hispanic whites. We present results for Hispanics and non-Hispanic whites in Figure 17. Only 5.8% of California's public school children are black, so we are not able to break out the results on school-level test scores separately for black students due to missing reported information in public data when small numbers of blacks are in a school. We find, among Hispanic children, that a $1,000 increase in per-pupil spending during ages 13-16 leads to an increase of 0.19 standard deviations in math, and 0.11 standard deviations in reading. No statistically significant effects are detectable for white children.

We next explore an alternative complimentary research design that exploits the fact that the funding formula involved concentration grants for districts that have more than 55 percent of their enrollment comprised of disadvantaged students (limited English proficiency, foster child, free lunch). This funding rule creates a kink in the LCFF funding received as a function of the district proportion of disadvantaged students, and can be leveraged within a two-stage least squares regression kink design (2SLS-RKD-IV). We first present graphical depictions of the kink at 55 percent and its direct effects on per-pupil revenues from the state and per-pupil spending for large (vs small) SFR-induced spending increases for successive post-LCFF cohorts (Figure 15). In contrast, and as a falsification check, we show that there is no positive kink relationship in per-pupil revenues (at 55%) for pre-LCFF cohorts—it is indeed flat and statistically insignificant (Figure 16). The identification assumption of the research kink design is that, absent the additional LCFF revenue, there would be no associated kink in outcomes beyond a district's 55-percent threshold of disadvantage; and thus, any kink in outcomes beyond that point can be interpreted appropriately as consistent with being attributable to the causal effects of per-pupil funding on student outcomes. We find this is indeed the case, as our graphical results show for post-LCFF cohorts that the kink and resultant improvements in both high school graduation rates and high school math achievement is more pronounced for cohorts that have been exposed to the increased resources for more of their school-age years and for whom the dosage was higher (i.e., as represented by the steeper upward-sloping kink beyond 55% shown in Figure 17). As a placebo test, we show in contrast that no such positive kink relationship is found for pre-LCFF cohorts' high school graduation rates; in fact, outcomes are decreasing in district proportion of disadvantaged students through the 55-percent thresholds for unexposed LCFF cohorts, while for exposed cohorts the trajectory turns upward (Figures 18).

[26] A one standard deviation increase in the proportion of revenue that is unrestricted is roughly 4 percentage points.

Table 6 presents these 2SLS-IV-RKD estimates and the previous 2SLS-IV estimates side-by-side for comparison for high school graduation and high school math achievement. With regard to interpretation, it is important to note the local average treatment associated with the 2SLS-IV-RKD estimates are more in line with the average effects of spending increases among students in high poverty schools, while we compare them with the average effects of spending we find among poor children across all schools on average using the 2SLS-IV estimates. We find that the 2SLS-IV-RKD estimates are larger though, as expected, with significantly less precision; but we find significant effects for both high school graduation rates and 11th grade math test scores using the regression kink design (insignificant 2SLS-RKD estimated effects in the case of reading). For example, the 2SLS-IV-RKD results indicate that a $1,000 increase in per-pupil spending experienced throughout high school years leads to an 8.77 percentage-point increase in high school graduation rates (Figure 20).

Given these results from both the 2SLS-IV and 2SLS-RKD estimates, it is natural to ask how the schools and districts achieved such improvements; yet doing such an analysis requires successfully choosing the correct subset of expenditures from an immense data set. We focus here on teacher salaries and administrator salaries, employee benefits, buildings, instruction, special education, preschool spending per 4-year old, and teacher professional development. The "buildings" category includes construction of new buildings and improvements and repairs to existing structures. The "instruction" category includes expenditures on regular K-12 education, as opposed to special, bilingual, or adult education, alternative schools, and a host of non-regular educational goals.

Table 7 presents the impact of LCFF on school inputs and the composition of district spending. Column 1 reveals that the increase in revenues caused the average school-level student-to-teacher ratio to fall by 0.2368 overall, whereas the increase in budgetary flexibility leads to a slight increase of 0.0722. The increase in flexibility also leads to a 3.8 percent increase in the likelihood that a teacher has limited experience, as can be seen in column 3. In Table 8, column 4 shows the impacts of revenues and flexibility on district per-pupil spending. Row 1 shows that 83 cents of every dollar is passed through as expenditures.[27] This is a relatively strong flypaper effect given the range found in the school finance reform literature generally, and is expected as Proposition 13 allows very limited scope for increases in state funding to lead to local property tax savings. Greater budgetary freedom causes a slight drop in per-pupil annual expenditures, but this may arise from a shift in accounting system requirements (e.g., reporting between general fund vs deferred maintenance fund).[28] Row 2 column 5 shows that the average teacher salary increases by 2.7 percent for every 10 percent increase in per-pupil revenues.

Row 1, columns 6 through 13 contain the proportion of the increased revenue that is spent on various expenditure categories. We find 11 percent of the increase went toward teacher salaries, 24 percent went toward instructional expenditures (including teacher salaries), 3 percent went toward administrator salaries, 12 percent went toward employee benefits, 5 percent was spent on capital improvements, and 6 percent was spent on special education.

---

[27] The estimated effect is larger over a two-year period as some district revenues in a given year are applied to a future school year's expenditures in an accounting sense that this district finance data may not fully capture due to the reporting requirements with the California Department of Education.

[28] In an accounting sense, districts can shift funding from one fund to another and this could lead to what looks like crowd-out when in fact it is not. For example, districts could take some state aid which is deposited into the general fund and then transfer it to the deferred maintenance fund, which would appear like crowd-out (e.g., see https://www.cde.ca.gov/fg/ac/ac/sacsminutes050614a.asp).

In our final set of analyses, we attempt to provide suggestive evidence of potential mechanisms. For the regression models that explore potential mechanisms, we instrumented for "teacher salaries per pupil", "administrative salaries per pupil", "capital expenditures per pupil" and "employee benefits per pupil" in the same model (and controlling for instrumented proportion of district revenue from state that is unrestricted).[29] The results show that LCFF-induced increases in teacher salaries per pupil (which include both increases in the number of teachers hired and increases in teacher salary) are significantly related to student achievement--for children from low-income families and Hispanic students (figures 19-20). On the other hand, administrative salaries, capital expenditures, and employee benefits are not found to be significantly related to student achievement. We acknowledge these exploratory patterns are far from definitive and are meant only to be suggestive. But they are supportive of the overall pattern of results; one interpretation may be that when increased resources make it to the classroom, they may more directly influence learning outcomes.

Section VII- Discussion

Overall, LCFF achieved its immediate purpose of increasing funding to districts with disadvantaged students. Though the policy is nearly fully funded, after three years of increases, district revenues were substantially higher than they would have been in the absence of LCFF. This was mostly due to the mechanical increase in funding to disadvantaged districts, but also to the reasonably strong flypaper effect. The inability of property owners to respond by lowering their contribution to public schools is likely a key contextual factor accounting for this result, but it is possible that some crowd out of instructional expenditures will be seen in the future due to mounting pension debt obligations (Koedel and Gassman, 2018). The policy also achieved its second immediate goal of reducing restrictions on state funding. A vast majority of state revenue is no longer subject to restrictions, though roughly 9 percent of the median district's budget is still tied up in categorical revenue streams. Spending patterns were not altered dramatically by the policy, with notable increases in spending on employee benefits being the exception to that generalization.

Increases in per-pupil spending caused by LCFF led to significant increases in high school graduation rates and student achievement. We find the effects increase in both the amount of spending increases and the number of school-age years of exposure. We find no evidence of differential pre-reform trending. Furthermore, we find a similar pattern of results across all three empirical approaches ((1) event-study difference-in-difference; (2) 2SLS-IV; (3) 2SLS-RKD-IV models), wherein the improvements in high school academic achievement closely track the timing of LCFF implementation, school-age years of exposure and the amount of district-specific LCFF-induced spending increase.

In particular, the increases in per-pupil spending led to significant increases in high school graduation rates overall by nearly six percentage points (associated with a $1,000 spending increase throughout high school), while the increase in expenditure flexibility increased

[29] For these analyses, we put the key explanatory variables in standard deviation units to facilitate a more straightforward comparison of effect sizes, and so the estimated effects are in the range we observe LCFF-induced variation in our key explanatory variables. A one-standard deviation increase in the proportion of district revenue from state that is unrestricted is roughly 0.04; a one-standard deviation increase in teacher salaries per pupil is roughly $500; a one-standard deviation increase in administrative salaries per pupil is roughly $100; a one-standard deviation increase in capital expenditures per pupil is also about $100; a one-standard deviation increase in employee benefits per pupil is roughly $500.

graduation rates by 1.4 a percentage points for each standard deviation increase in budgetary freedom. The effects were heterogeneous across student demographic groups, being strongest for African-American students, but positive and statistically significant for all student subgroups. The increases in per-pupil spending improved test scores as well, with the additional expenditures significantly boosting literacy and numeracy for Hispanic and poor children. This more targeted effect is somewhat expected, as the policy was meant to deliver greater resources for low-income students and students with limited English proficiency.

We find, for low-income students, that a $1,000 increase in district per-pupil spending during ages 13-16 led to a 0.19 standard deviation increase in 11th grade mathematics test scores. To put this magnitude in perspective, the 0.19 standard deviation increase in high school math achievement is equivalent to 37% of the average mathematics achievement gap between poor and non-poor students in 11th grade; is equivalent to 24% of the average mathematics black-white achievement gap in 11th grade; and is equivalent to 34% of the average mathematics Hispanic-white achievement gap in 11thgrade (based on data from all CA public schools, 2003-16). On average, students gain about 0.25 standard deviations each 10 months of high school (one year), so the 0.19 standard deviation increase in high school math achievement (resultant from a 1,000 increase in district per-pupil revenue during ages 13-16) is equivalent to approximately 7 months of learning (i.e., 0.19/.25).

We find, for low-income students, that a $1,000 increase in district revenue per-pupil during ages 13-16 led to a 0.08 standard deviation increase in 11th grade reading test scores. To put this magnitude in perspective, the 0.08 standard deviation increase in high school reading achievement is equivalent to 13% of the average reading achievement gap between poor and non-poor students in 11th grade; is equivalent to 10% of the average reading black-white achievement gap in 11th grade; and is equivalent to 12% of the average reading Hispanic-white achievement gap in 11th grade (based on data from all CA public schools, 2003-16). This 0.08 standard deviation increase in high school reading achievement (resultant from a 1,000 increase in district per-pupil revenue during ages 13-16) is equivalent to approximately 3 months of learning (i.e., 0.08/.25). These are meant as rough back-of-the-envelope calculations to facilitate putting the magnitudes in perspective. In sum, the evidence suggests that money targeted to students' needs can make a significant difference in student outcomes and can narrow achievement gaps.

The magnitudes of these effects are large and broadly similar to those found in recent studies that use quasi-experimental methods. Candelaria and Shores (2015) find that, seven years after a reform event, per-pupil revenues increase by an average of 11.9% and graduation rates increase by an average of 8.4 percentage points in the poorest quartile of districts. In California, $1,000 was 11.8% of average per pupil expenditures on the eve of LCFF in 2012.[30] With three successive years of exposure to such an increase in per-pupil spending, this 11.8% increase led to a 5.89 percentage point increase in the graduation rate, a smaller effect than is present in Candelaria and Shores. However, the event-study graph contained in Figure 13 shows that this effect is increasing with duration of exposure, suggesting that a seven-year effect may be substantially larger and is in line with the previous study.

As for test scores, an increase of $1,000 in per pupil expenditures over four years raises numeracy scores for poor children by 0.19 standard deviations (event study graph shown in Figure 14). This is the precise magnitude found in a national study by Lafortune, Rothstein, and

[30] The average expense per student (in terms of average daily attendance) was $8,448 in the 2012-2013 school year. For details, see https://www.cde.ca.gov/ds/fd/ec/currentexpense.asp

Schanzenbach (2015, p. 6), wherein this effect took ten years to manifest rather than four. Several differences in the two studies may explain the accelerated appearance of the effect in California. First, our estimates come from NAEP-normed tests given in 11th grade as compared to 4th and 8th grade NAEP tests in the national study. It may be that schools serving high school students are better able to either capture new revenues or translate them into student achievement. Second, the nature of LCFF, with its explicit focus on disadvantaged students and localized accountability structure, may have encouraged more efficient uses of the new expenditures than did the "adequacy" reforms studied previously. Third, the effect in California may exhibit diminishing returns over time, so that the 10-year effect may be similar to the 4-year effect found here. This would be the case if districts are able to reach their long-run level of efficiency within several years rather than a decade. In any event, both the test-score and graduation-rate effects are within plausible ranges that one would expect, given recent studies.

Our estimated significant effects of per-pupil spending are robust across student outcomes and identification strategies and robust to a variety of falsification checks. On the other hand, the estimated effects of reductions in the proportion of funding with categorical restrictions exhibit a far less consistent pattern of results across outcomes and subgroups. This may simply be an artifact that the reduction in the proportion of funding with restrictions was a much more modest change, but this requires further investigation and may require more years of data before definitive conclusions can be reached on the latter.

The context of California's legal and policy environment is important to consider when making sense of these results. It bears repeating that it is extraordinarily difficult for tax payers to capture the new state revenue with lower property tax rates. This condition may not hold in other states. Second, the LCFF era followed a period of deprivation for district resources. States in which schools are well-funded may see diminishing marginal returns to budget increases of LCFF's scale. Third, the changes in budgetary restrictions are of a limited scale. These results say little about the magnitude of impacts one should expect from much larger changes or changes at a different baseline level of restrictions.

The impacts of the new policy are still reverberating, and the verdict is still out; but, given the magnitude of redistribution in the LCFF, the policy provides a test of how state policy and school resources can shape student achievement and reduce inequality. Notwithstanding those limitations, this study is among the first to document impacts of LCFF on student outcomes, and jointly test the impact of a simultaneous change in school district revenues, directed toward disadvantaged students, and budgetary restrictions on how such revenues can be spent. The findings suggest that both revenue and flexibility can be productive in enhancing the academic achievement and educational attainment of disadvantaged students. These findings are particularly noteworthy in light of the fact that LCFF is a recent reform and has been gradually rolled out to become fully funded and implemented in the past year. The country is watching as it is anticipated that, if successful, the new school finance measure may lead other states to adopt similar legislation. Time will tell—in the interim, this new research evidence suggests that money targeted to the needs of students, and allocated by local districts to meet those needs, can make a difference in student outcomes.

Figure 1a: LCFF Funding and % Disadvantage

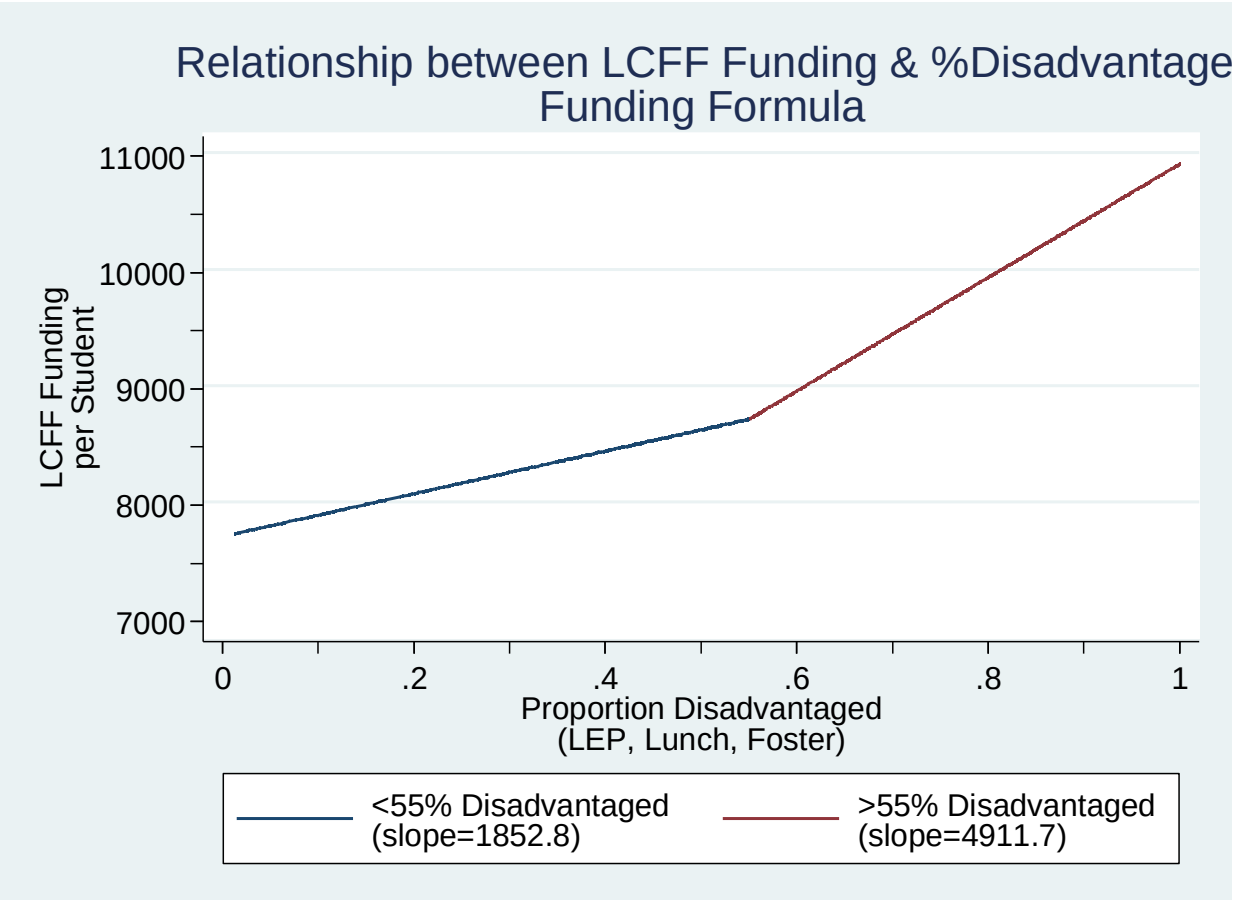


Figure 1b: Supplemental and Concentration Funding per % Disadvantage

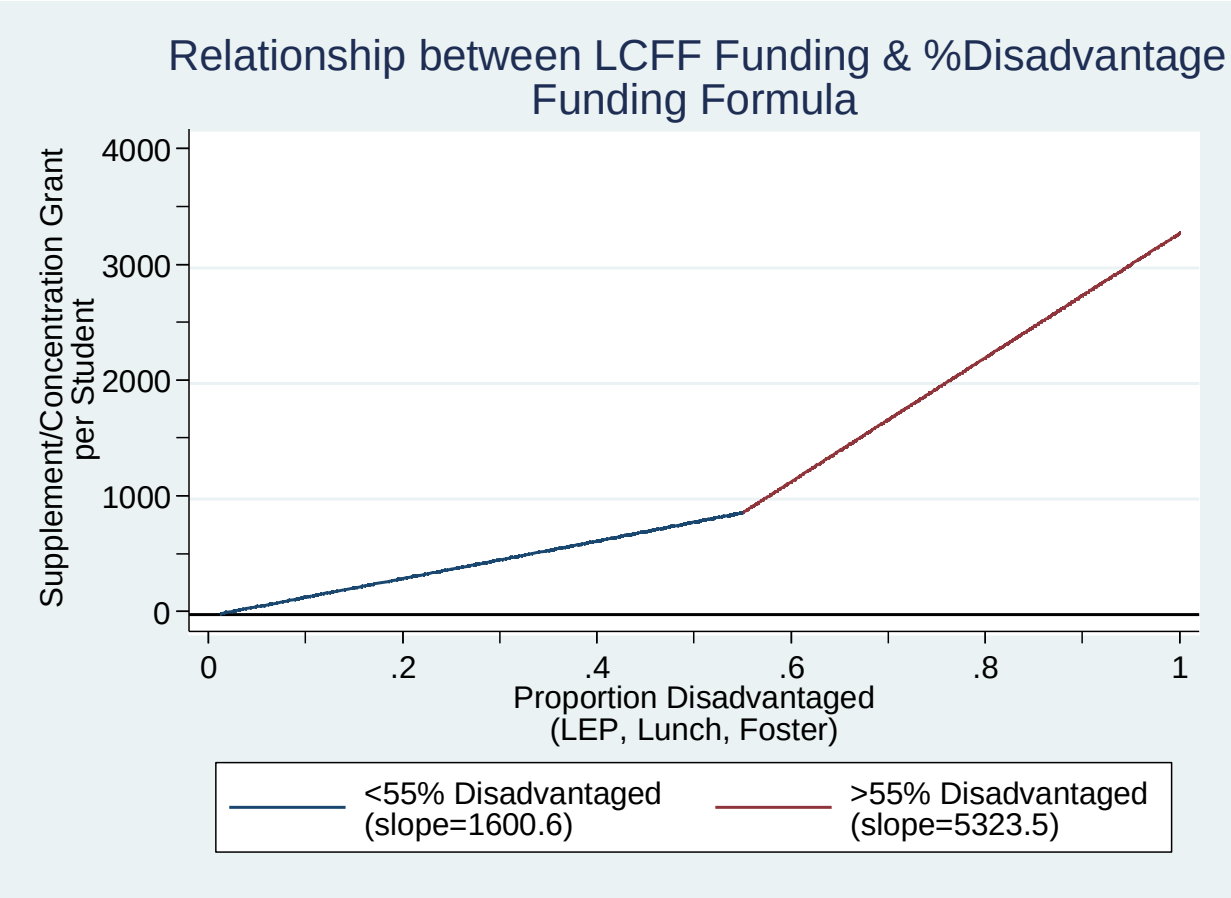


Figure 2: The Distribution of Funding and Disadvantage

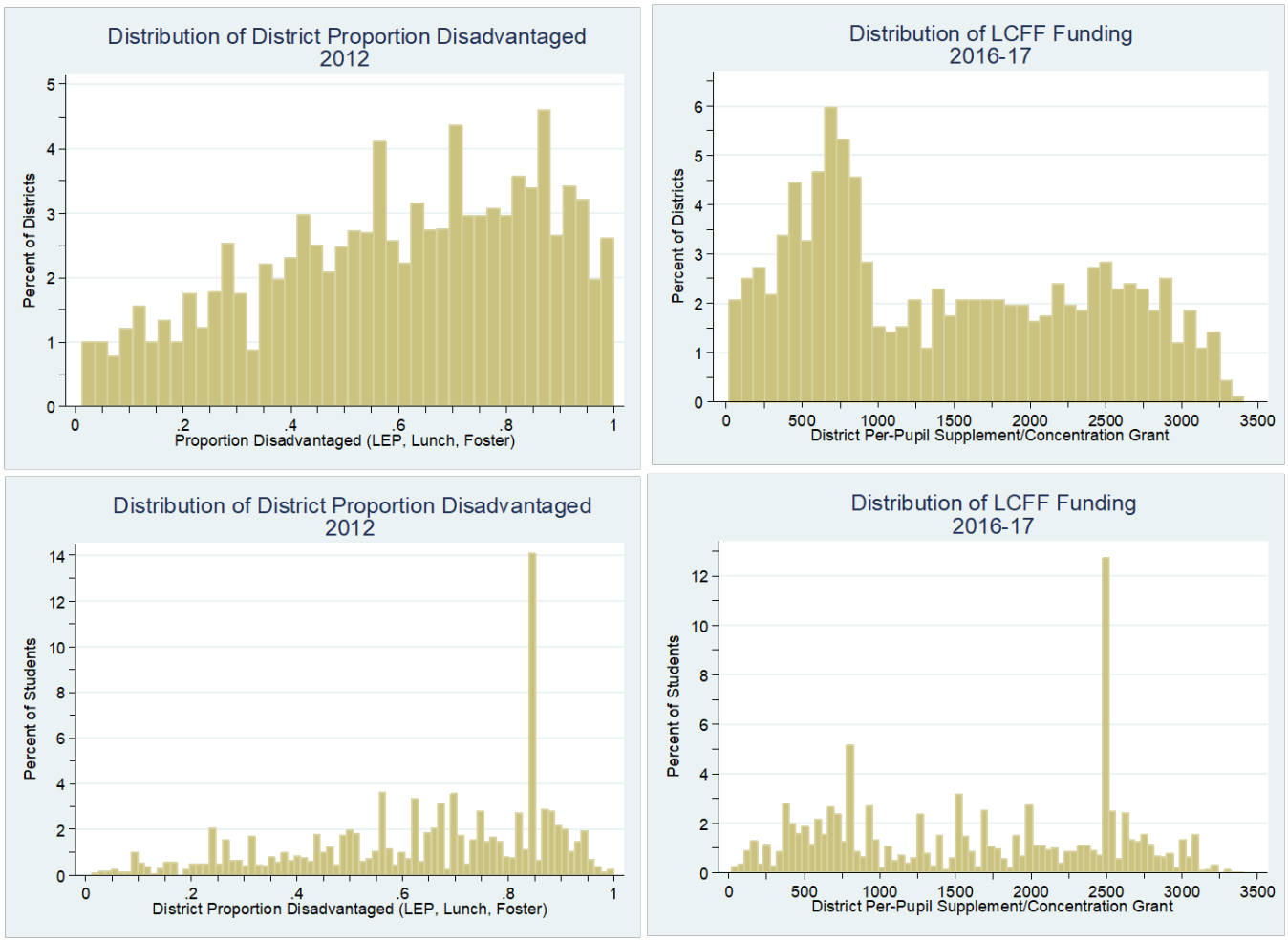


Figure 3: Revenue by Source, 1995 - 2016

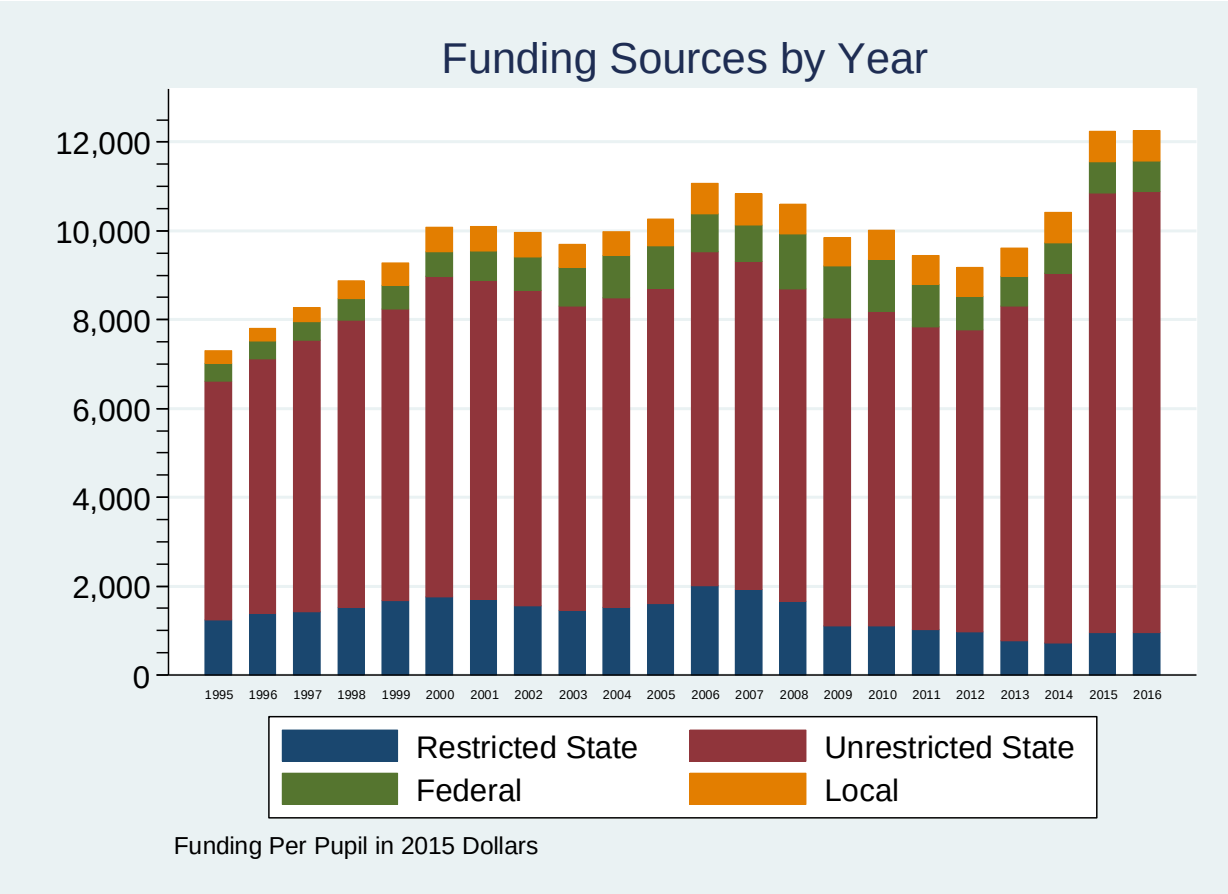


Figure 4: Categorical Restrictions as a Percent of Revenue, 1995-2016

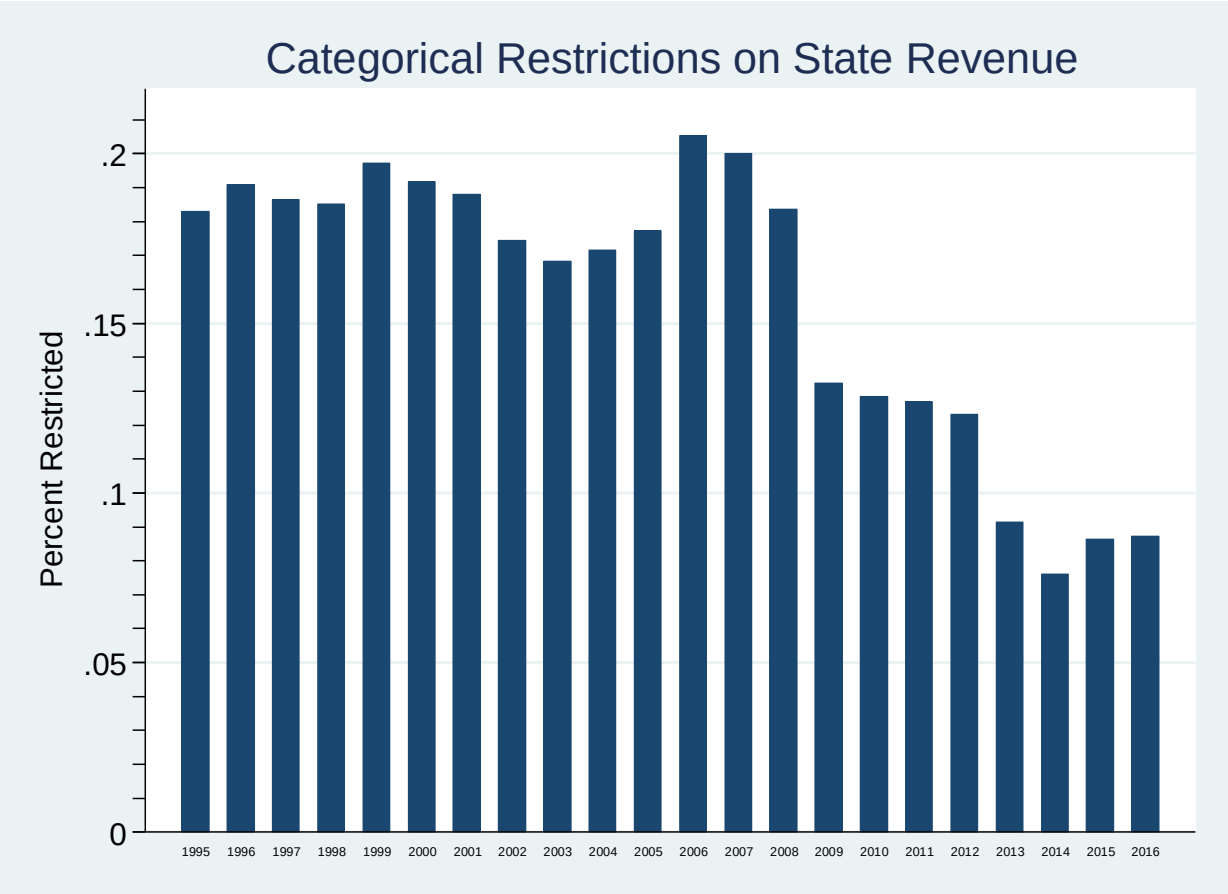


Figure 5: Change in Revenue and Restrictions, 2012-2016

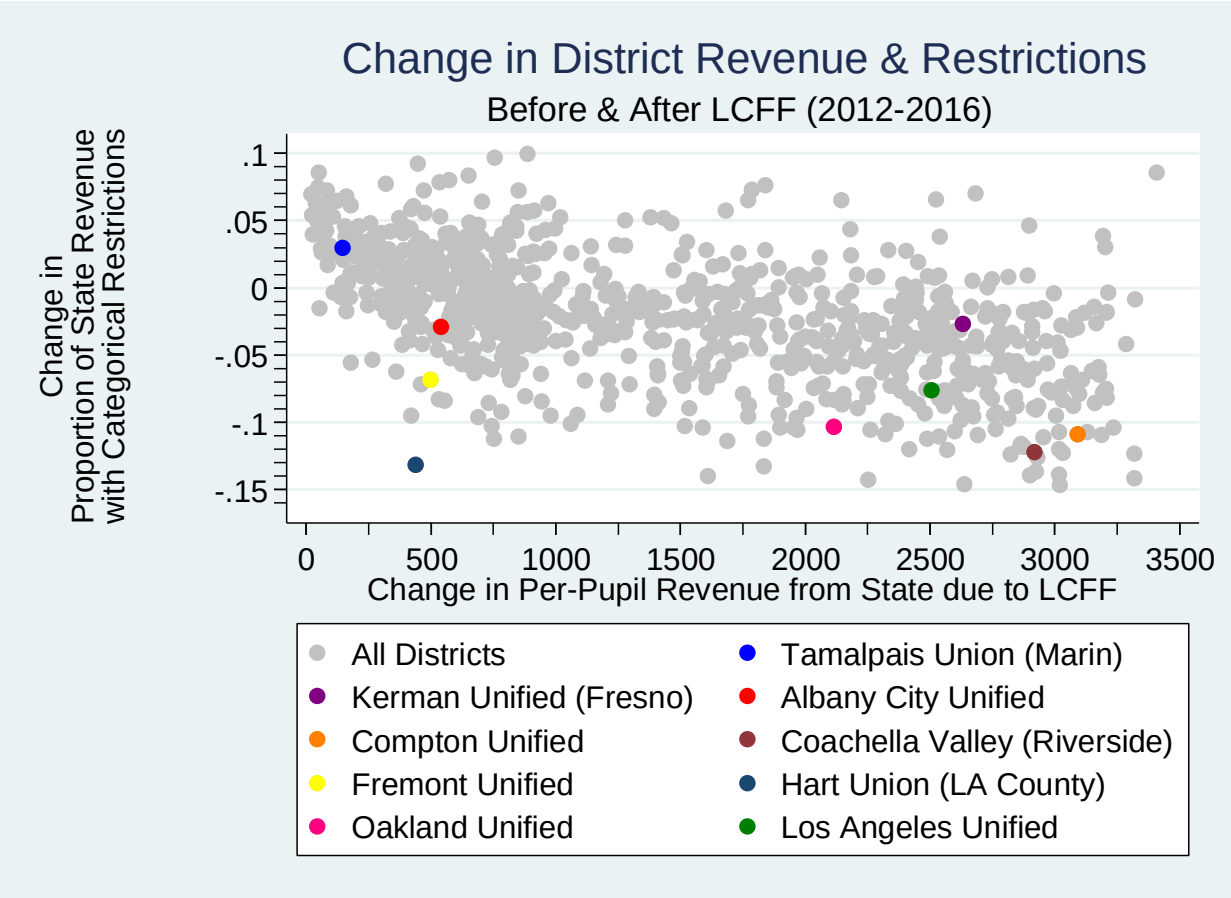


Figure 6: Relationship between Formula and Funding Change before and after LCFF

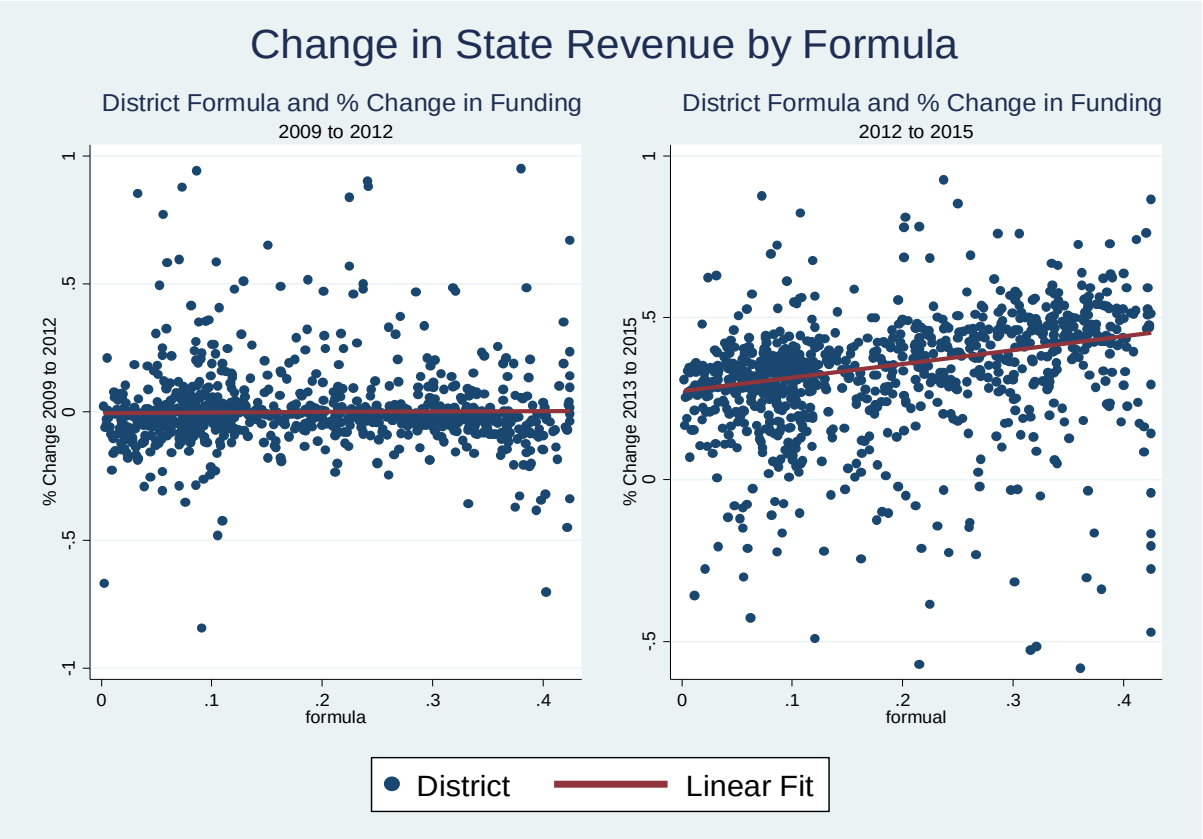


Figure 7: Funding and Revenue Change, Linear Fit

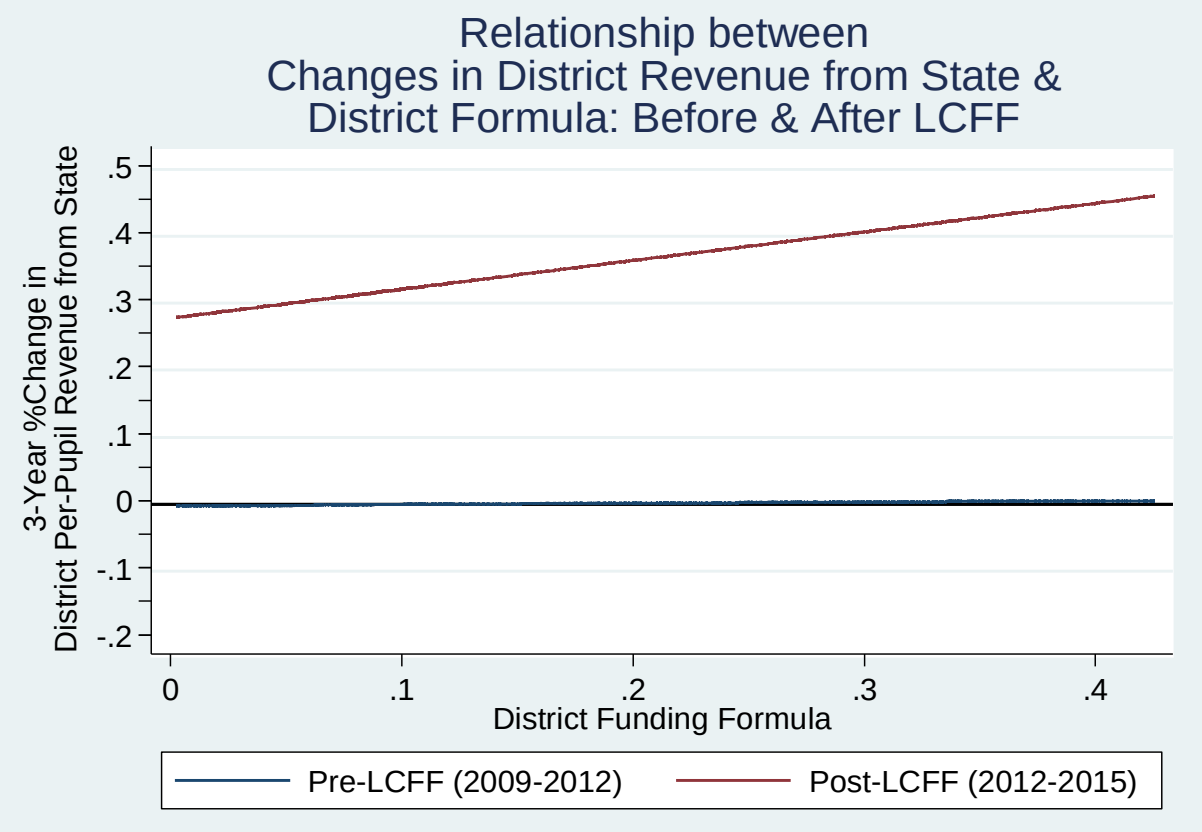


Table 1: Regressions of Revenue Change on District Disadvantage

	%Change in $$ (2009-2019)	%Change in $$ (2012-2015)
Formula	0.019	0.445
	(0.048)	(0.066)**
_cons	-0.003	0.269
	(0.009)	(0.010)**
R^2	0.0004	0.1817
N	889	905

*p<0.05

**p<0.01

Figure 8: Predicted vs. Actual Revenue, 2004 - 2016

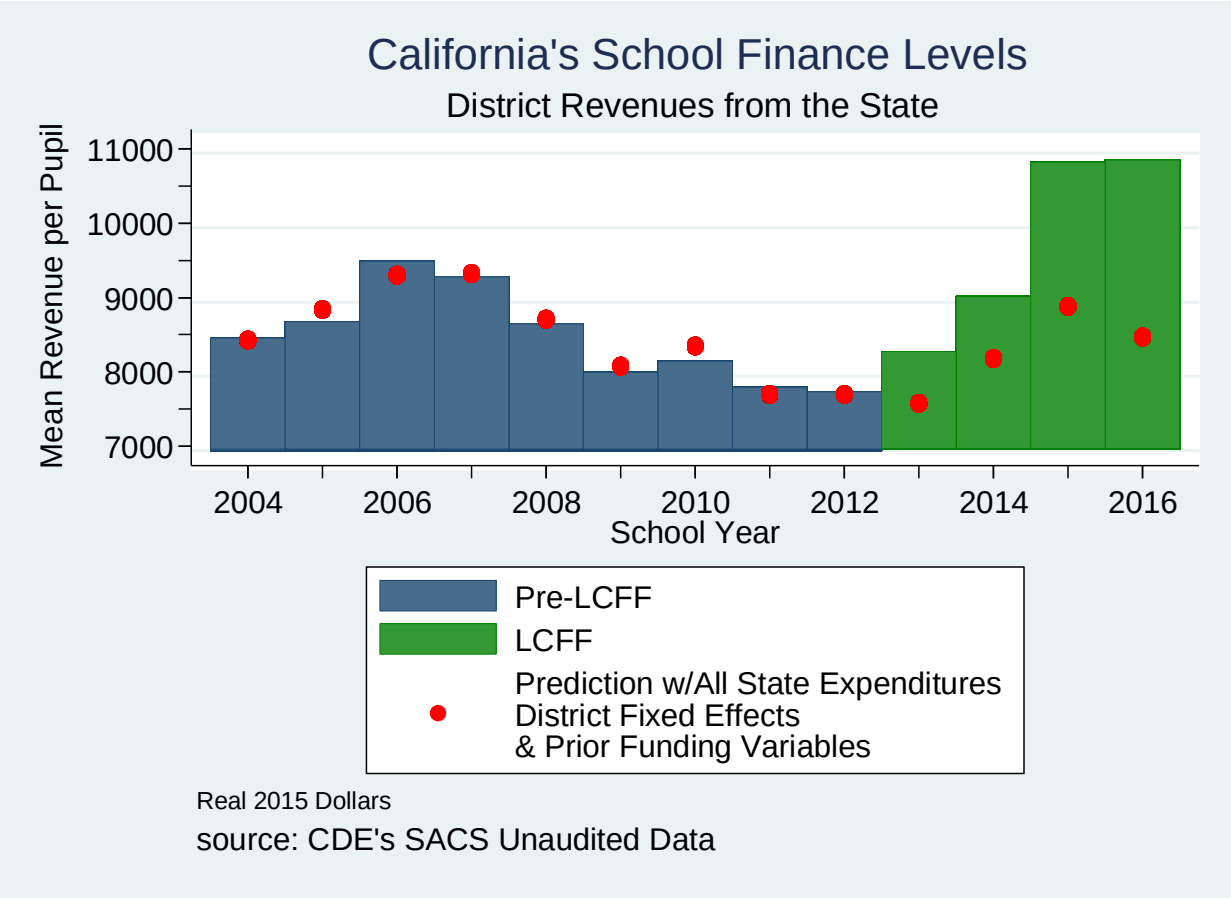


Figure 9: Event Study Graphs by District-Level Disadvantage

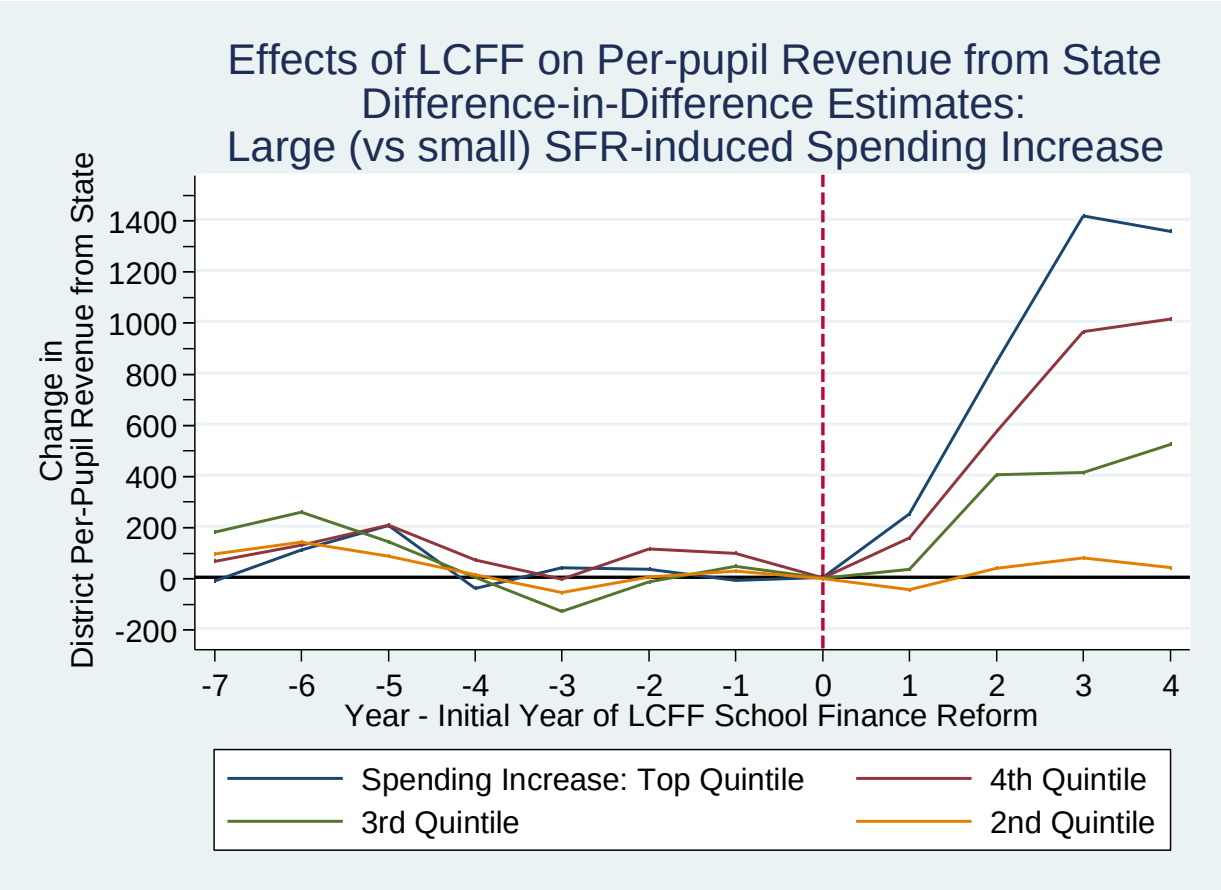


Table 2: F-Statistics on Excluded LCFF Instruments from 1st-Stage Models of District Per-Pupil Spending & Proportion of Funding with Restrictions

	2SLS-IV			
	First-Stage Models, Dependent variable:			
	(1)	(2)	(3)	(4)
	District Per-pupil Spending		Proportion of Revenue from State with no Categorical Restrictions	
Model	without 2012 %Restricted	with 2012 %Restricted	without LCFF dosage*LCFF exposure	with LCFF dosage*LCFF exposure
Excluded Instruments	LCFF dosage* LCFF exposure	LCFF dosage* LCFF exposure	2012 %Restricted* LCFF exposure	2012 %Restricted* LCFF exposure
First-Stage, F-Statistic on excluded instruments	33.38	36.20	175.47	62.86
Number of Districts	920	920	920	920

Figure 10: Expenditures for Salaries, Buildings, and Benefits

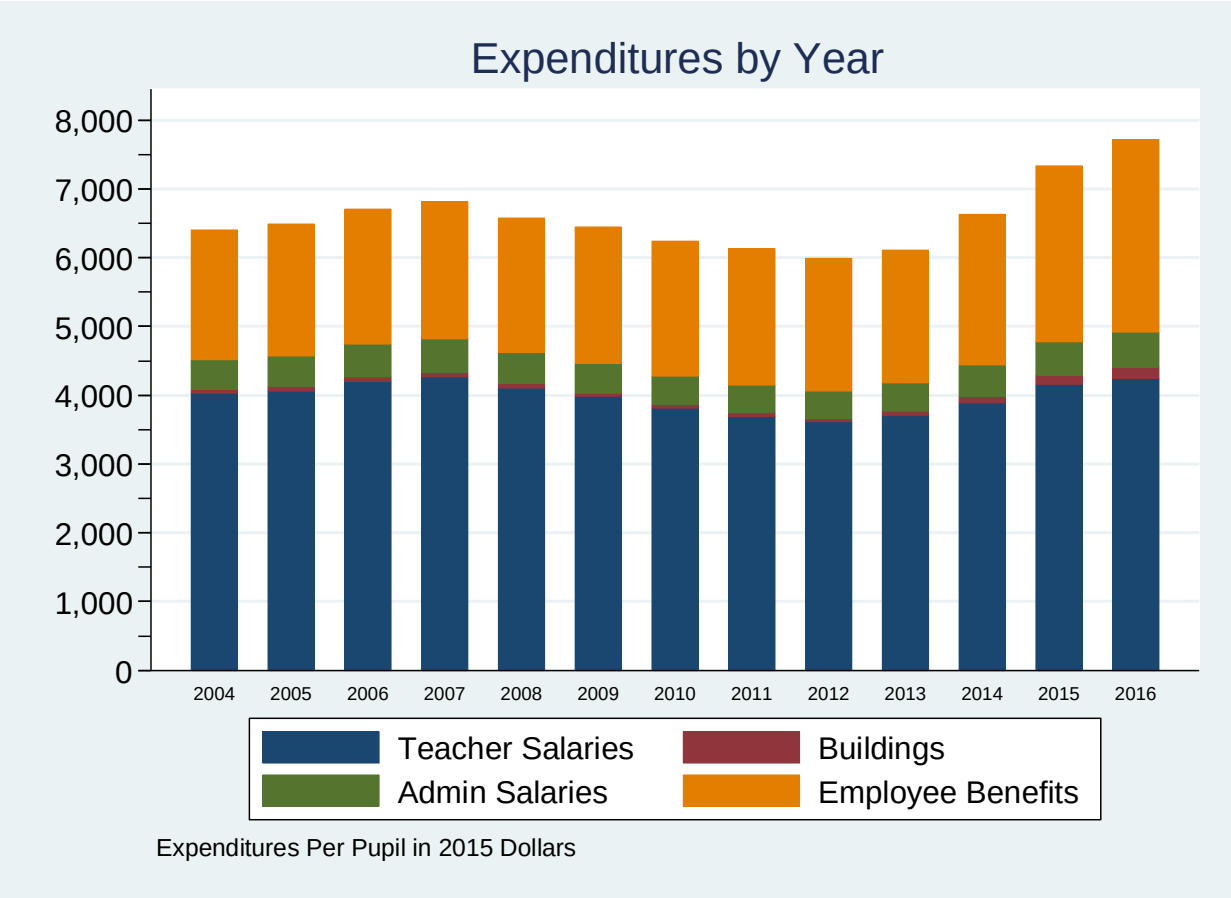


Figure 11: Instructional Expenditures by Year

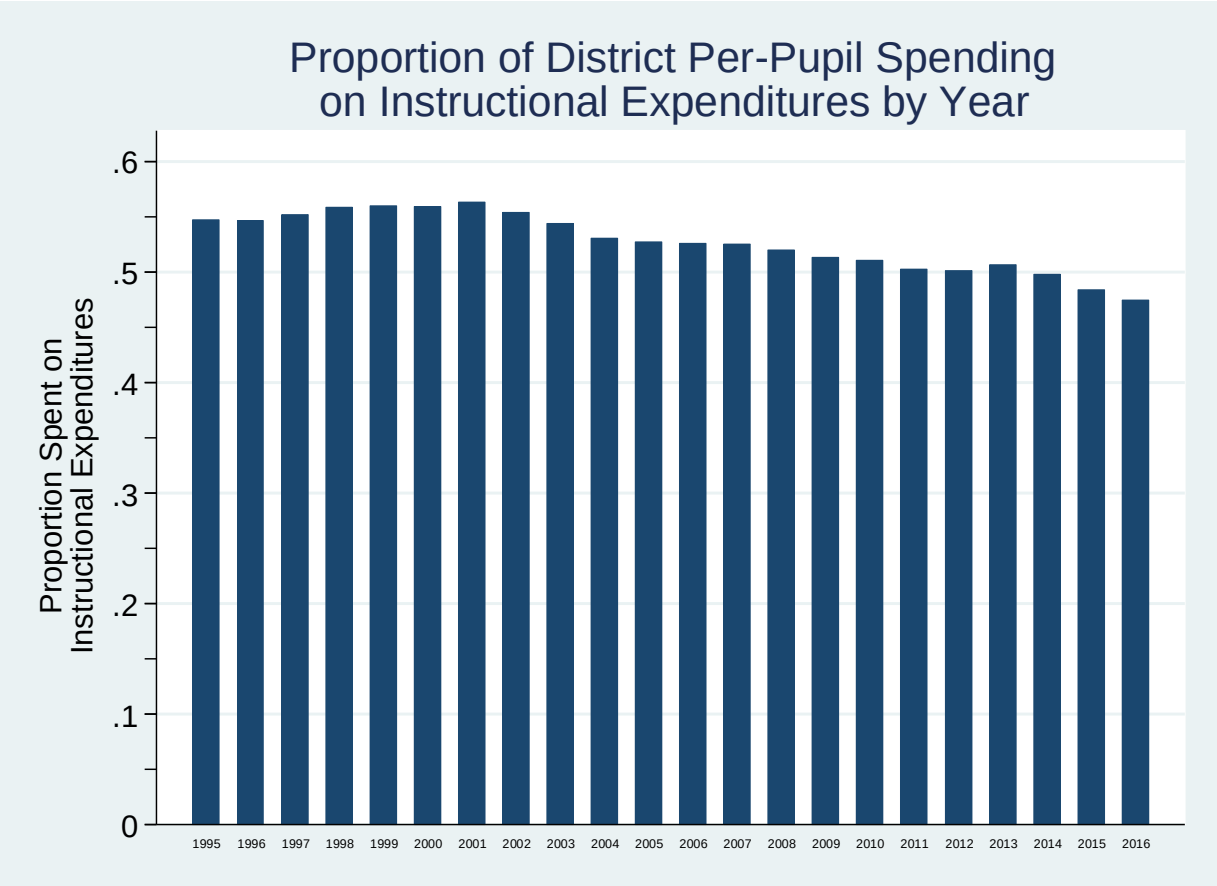


Table 3: 2SLS-IV Estimates of Effects of LCFF SFR-Induced Effects of Per-Pupil Spending on High School Graduation Rates
by Child Poverty & Race/Ethnicity

	(1)	(2)	(3)	(4)	(5)
	High School Graduation Rate (4-Year Cohort)				
KEY EXPLANATORY VARIABLE:	2SLS-IV				
	All Kids	Poor Kids	Whites	Blacks	Hispanics
(SFR) Instrumented avg District Per-pupil Spending, ages 15-17 (in 000s)	0.0589***	0.0510***	0.0412**	0.0771**	0.0568***
	(0.0146)	(0.0166)	(0.0197)	(0.0323)	(0.0172)
(Instrumented) Proportion of Revenue from State Unrestricted, ages 15-17 (in std units)	0.0141**	0.0044	0.0075	0.0288**	0.0094
	(0.0059)	(0.0074)	(0.0057)	(0.0140)	(0.0080)
Number of School-Year Observations	9,882	9,438	6,829	3,325	8,527
Number of Schools	1,585	1,535	1,121	621	1,394
Number of School Districts	399	395	361	182	373

Figure 12: The Impact of Expenditures on Graduation Rates

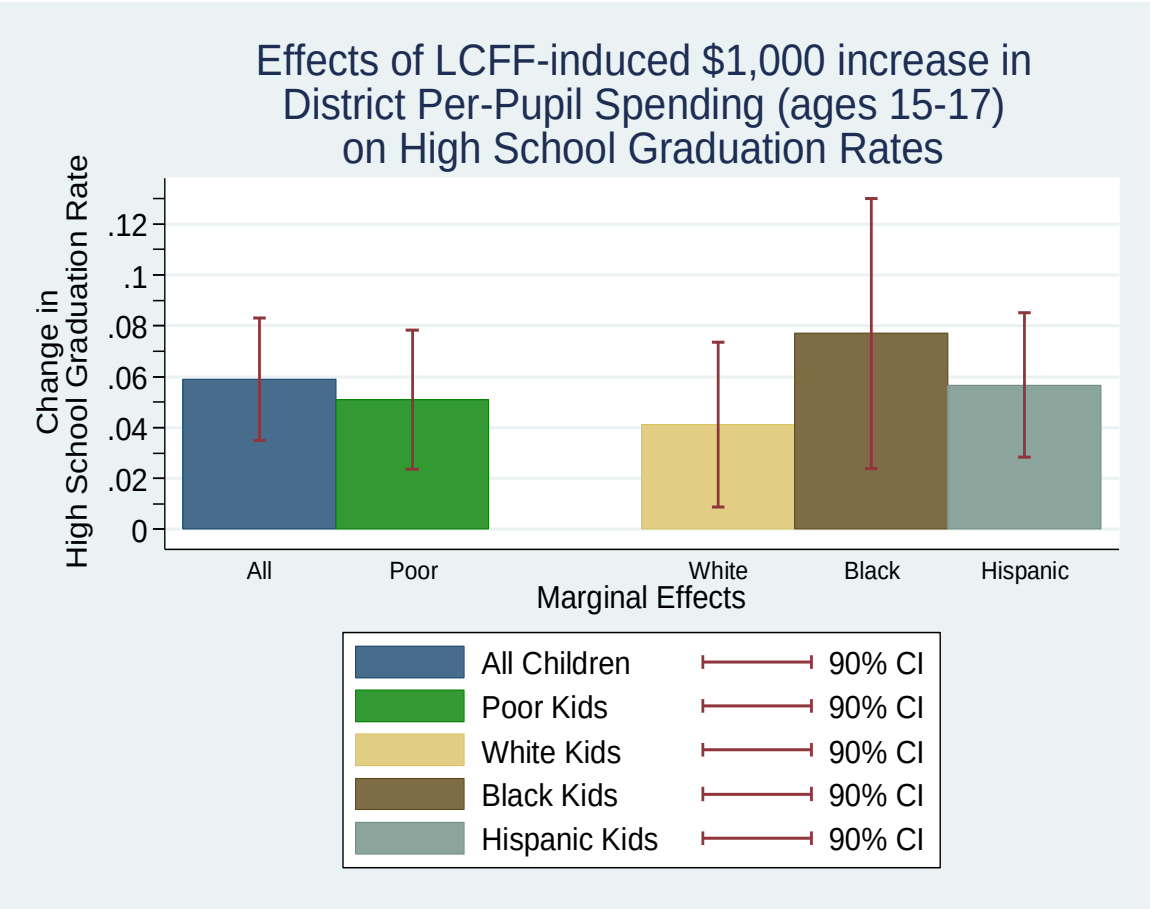


Figure 13: The Impact of Budgetary Freedom on Graduation Rates

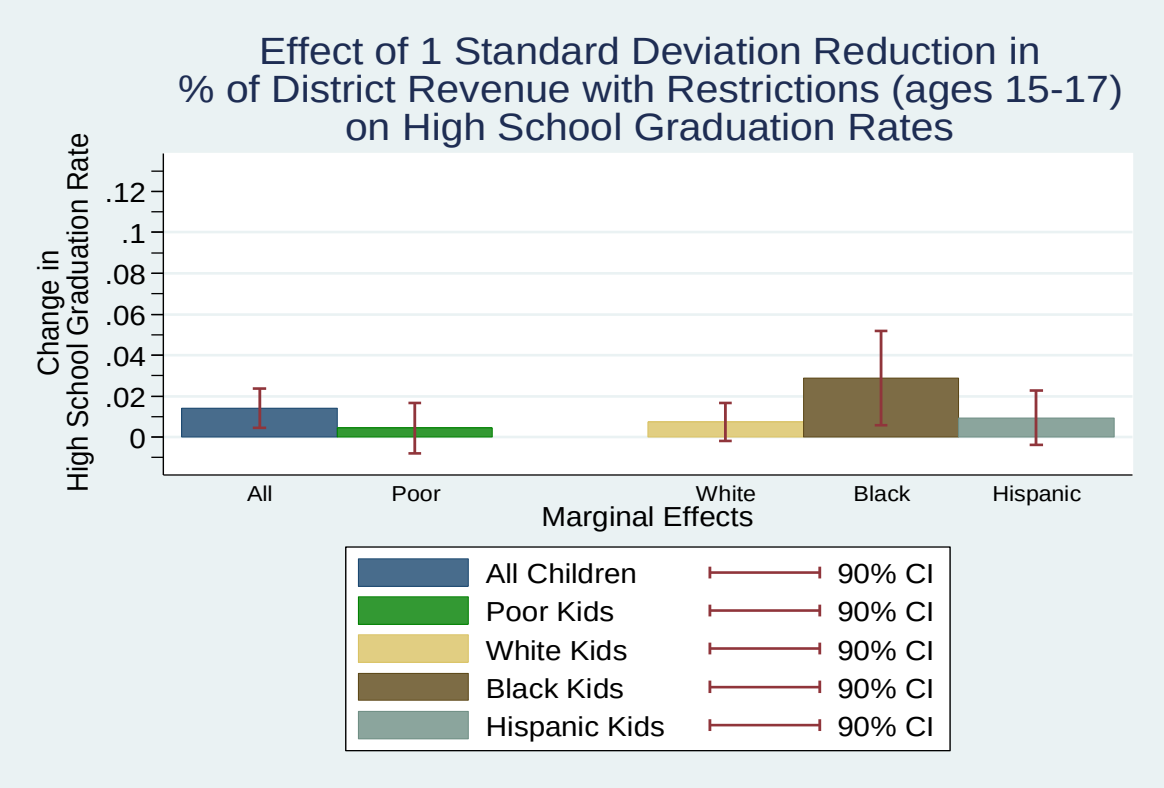


Figure 14: The Impact of Expenditures on Academic Achievement

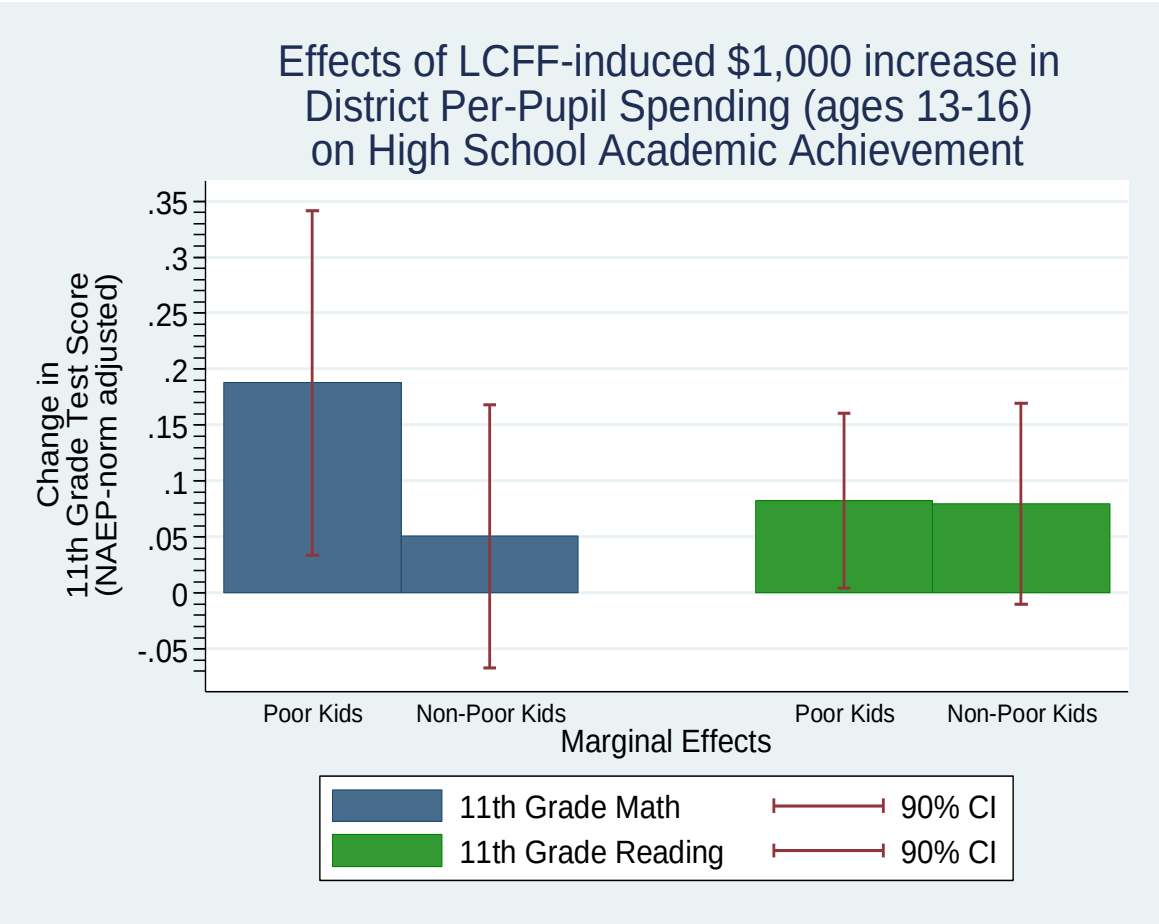


Table 4: The Impact of LCFF on Academic Achievement

	(1)	(2)	(3)	(4)
	Standardized Test Score (NAEP-scale normed)			
	2SLS-IV			
KEY EXPLANATORY VARIABLE	Math, 11th grade		Reading, 11th grade	
	Poor Kids	Non-Poor Kids	Poor Kids	Non-Poor Kids
(SFR) Instrumented avg District Per-pupil Spending, ages 13-16 (in 000s)	0.1876**	0.0506	0.0825*	0.0793+
	(0.0937)	(0.0713)	(0.0475)	(0.0546)
(Instrumented) Proportion of Revenue from State Unrestricted, ages 13-16 (in std units)	0.1752**	-0.2609	-0.0765**	0.2006
	(0.0751)	(0.2366)	(0.0372)	(0.1840)
Number of School-Year Observations	8,816	8,801	13,414	12,088
Number of Schools	1,426	1,228	1,540	1,421
Number of School Districts	387	367	394	384

Table 5: Academic Impacts of LCFF by Ethnicity

	(1)	(2)	(3)	(4)
	Standardized Test Score (NAEP-scale normed)			
	2SLS-IV			
KEY EXPLANATORY VARIABLE	Math, 11th grade		Reading, 11th grade	
	Hispanics	Whites	Hispanics	Whites
(SFR) Instrumented avg District Per-pupil Spending, ages 13-16 (in 000s)	0.1937*	0.0093	0.1136*	-0.0106
	(0.0991)	(0.0786)	(0.0582)	(0.0531)
(Instrumented) Proportion of Revenue from State Unrestricted, ages 13-16 (in std units)	0.2539**	-0.5707**	-0.0962	0.0911
	(0.1184)	(0.2790)	(0.0615)	(0.2231)
Number of School-Year Observations	8,455	7,101	12,859	10,330
Number of Schools	1,353	1,002	1,463	1,204
Number of School Districts	365	351	374	373

Figure 15: Regression Kink Estimates of LCFF's Impact on Revenue

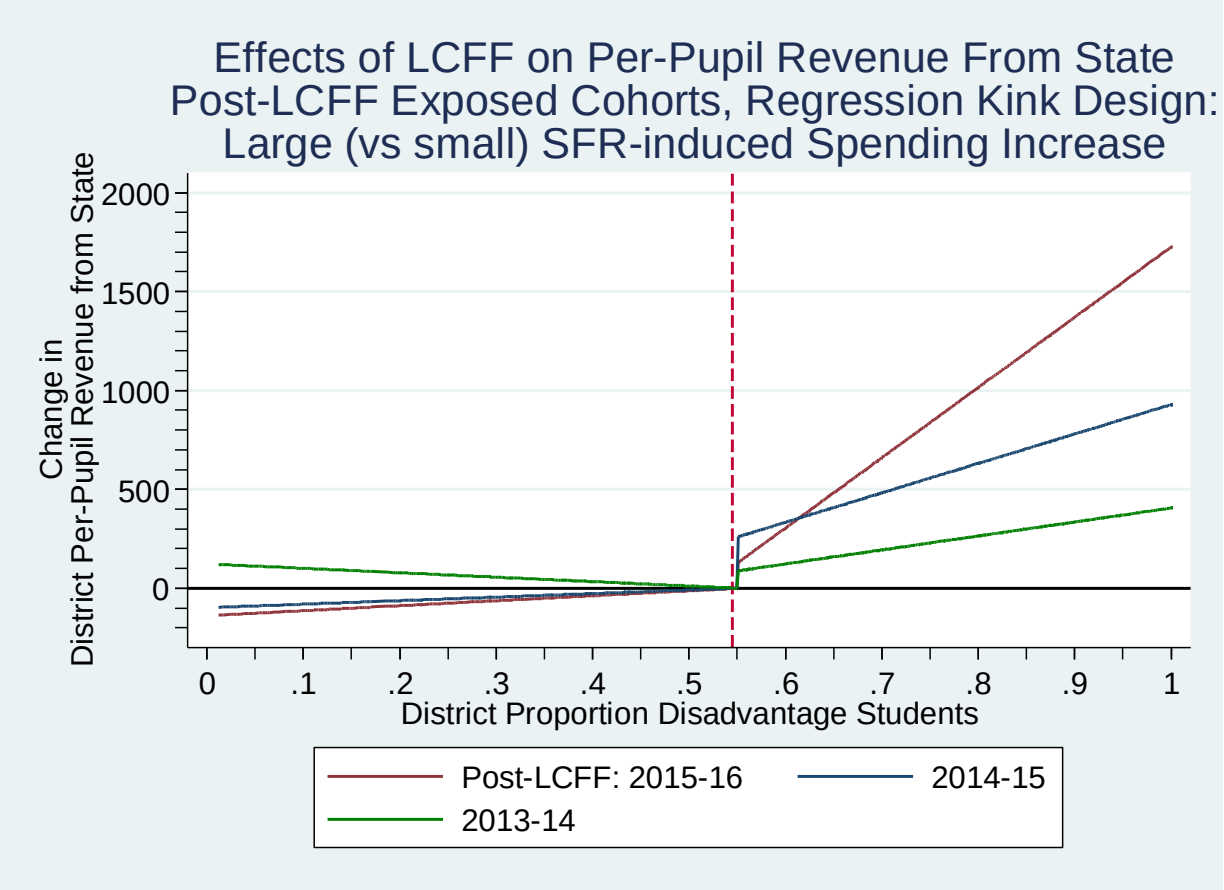


Figure 16: Regression Kink Placebo Test of Impact of LCFF on Revenue

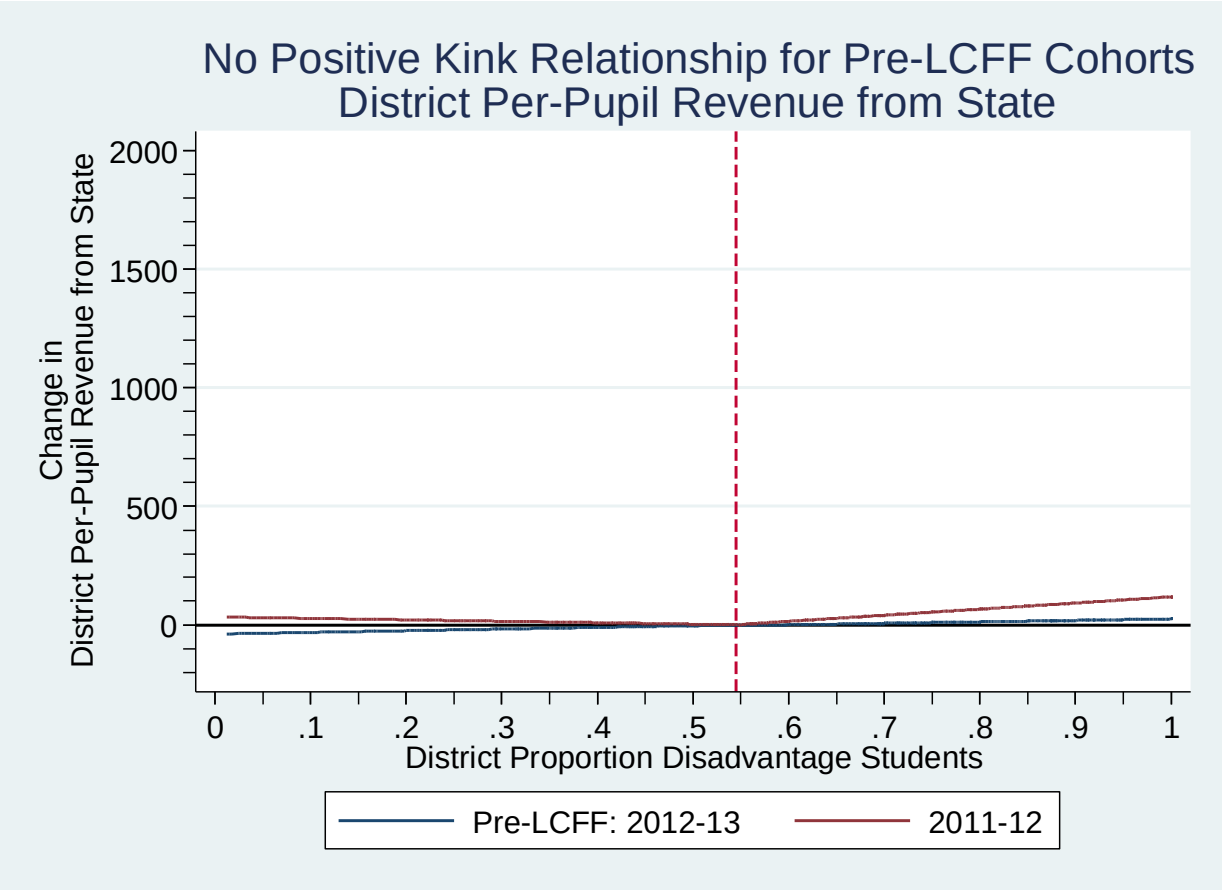


Figure 17: Regression Kink Evidence of Impact of LCFF on Graduation Rates

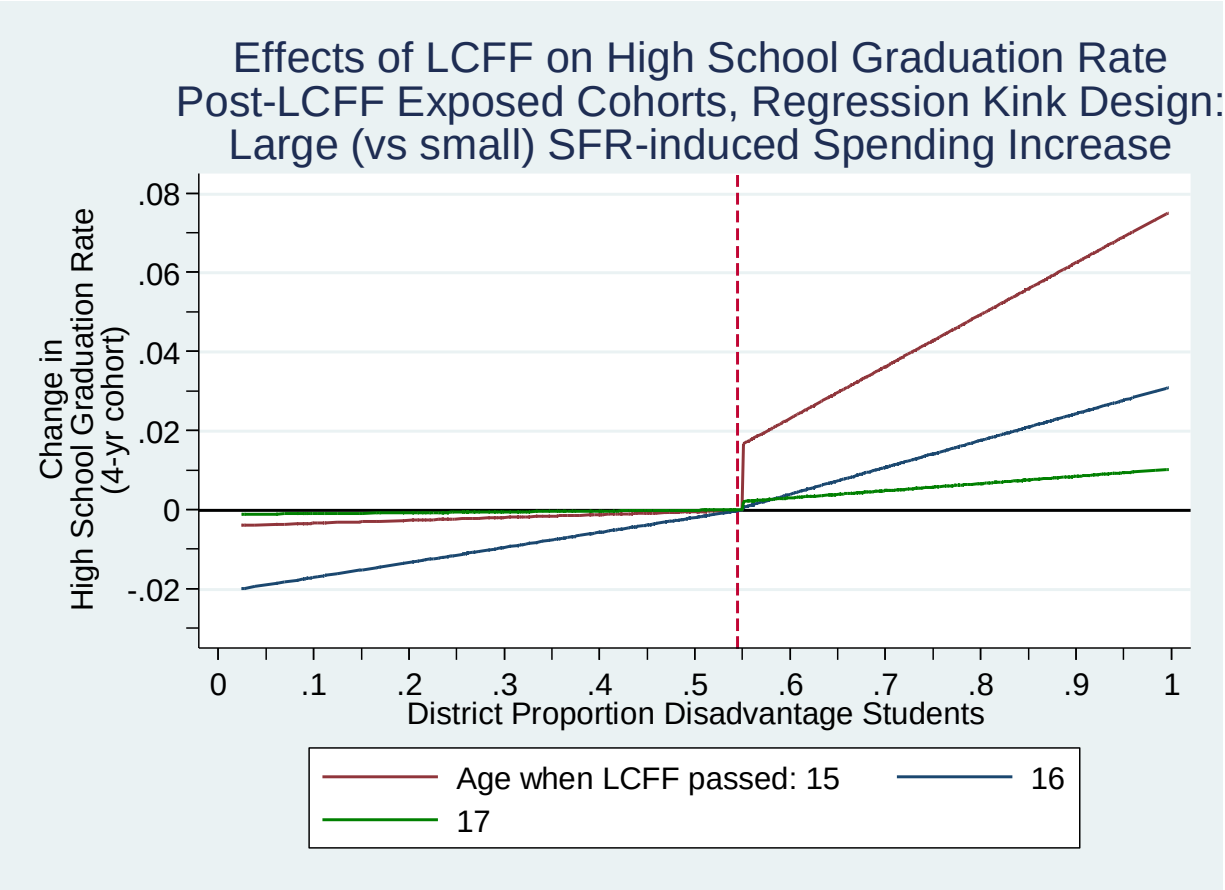


Figure 18: Placebo Test of LCFF's Impact on Graduation Rates

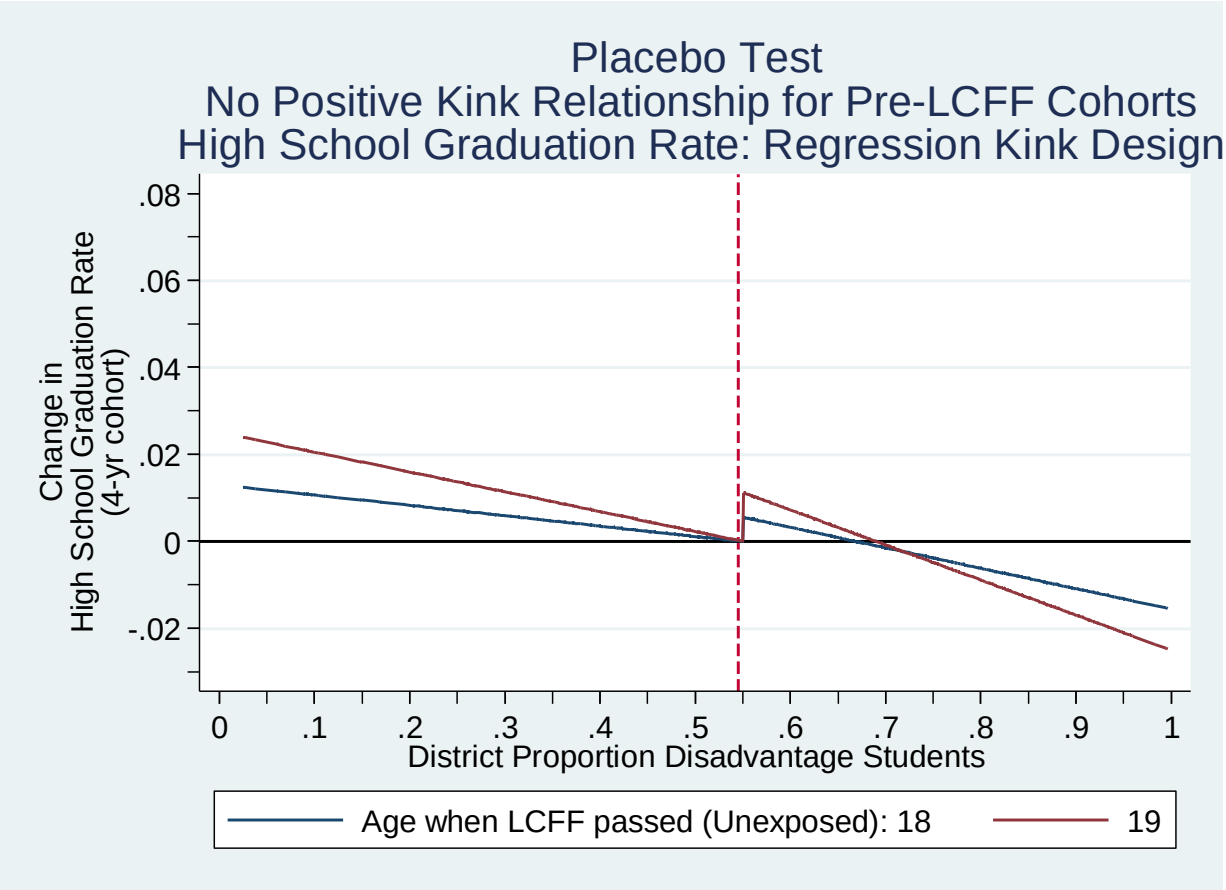


Table 6: Regression Kink Estimates of LCFF's Impact on Graduation Rates and Numeracy

	(1)	(2)	(3)	(4)
KEY EXPLANATORY VARIABLE	High School Graduation Rate (4-Year Cohort)		Math, 11th grade Standardized Test Score (NAEP-scale normed)	
	Poor Kids, 2SLS-IV	All Kids, 2SLS-RKD	Poor Kids, 2SLS-IV	All Kids, 2SLS-RKD
(SFR) Instrumented avg District Per-pupil Spending, ages 13-16 (in 000s)	0.0510***	0.0877*	0.1876**	0.4848***
	(0.0166)	(0.0435)	(0.0937)	(0.1834)
Number of School-Year Observations	9,438	9,882	8,816	10,790
Number of Schools	1,535	1,585	1,426	1,534
Number of School Districts	395	399	387	395

Table 7: Impact of LCFF on School-level Inputs

KEY EXPLANATORY VARIABLE:	(1)	(2)	(3)
	School-level Inputs		
	Avg Student-to-Teacher Ratio	Proportion of Teachers w/ Masters	Prob(avg Teacher has Limited Experience)
(Instrumented) avg District Per-pupil Revenue from State	**-0.2368*** **	**0.0023**	**0.0062**
	(0.0549)	**(0.0028)**	**(0.0067)**
(Instrumented) Percent of Revenue from State Unrestricted	**0.07221*** **	**0.0002**	**0.0038** **
	(0.0148)	**(0.0007)**	**(0.0017)**
Number of Schools	9,231	9,231	9,231
Number of School Districts	920	920	920

{Insert Table 8 here}

Figure 19: Effect of Expenditures by Category on low-income children

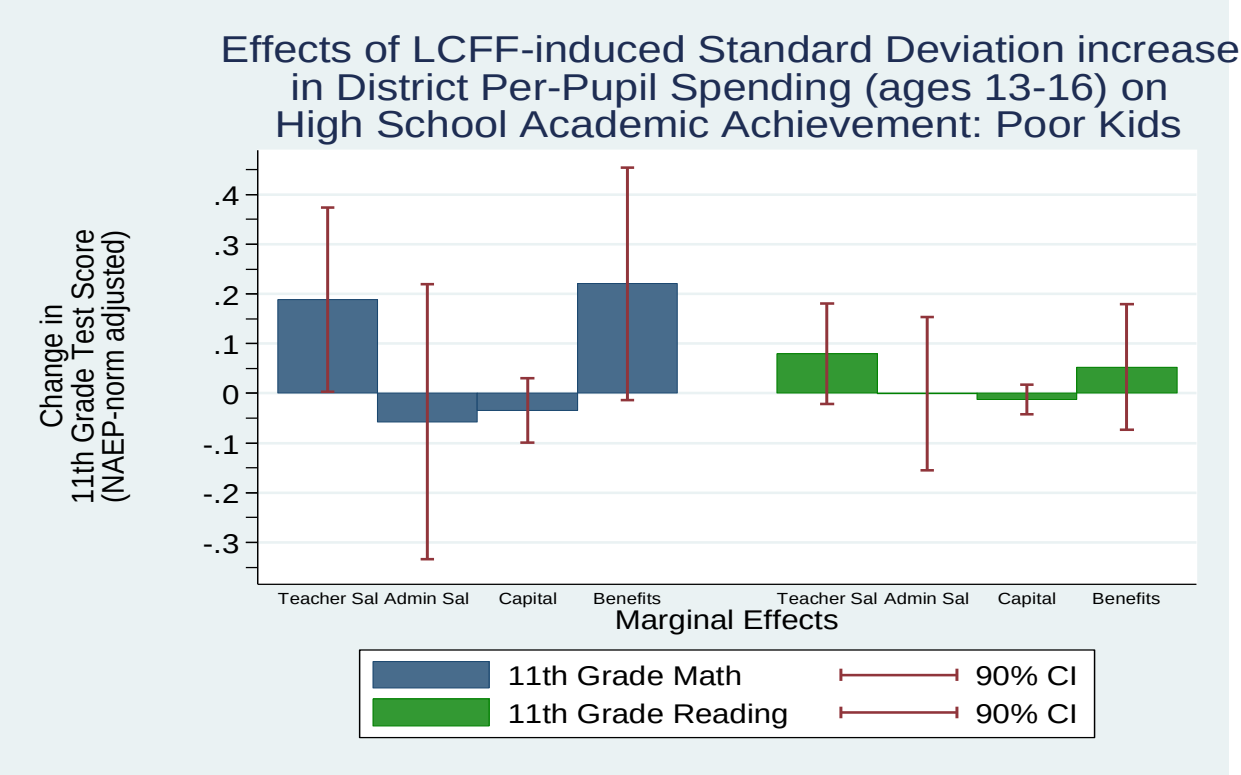
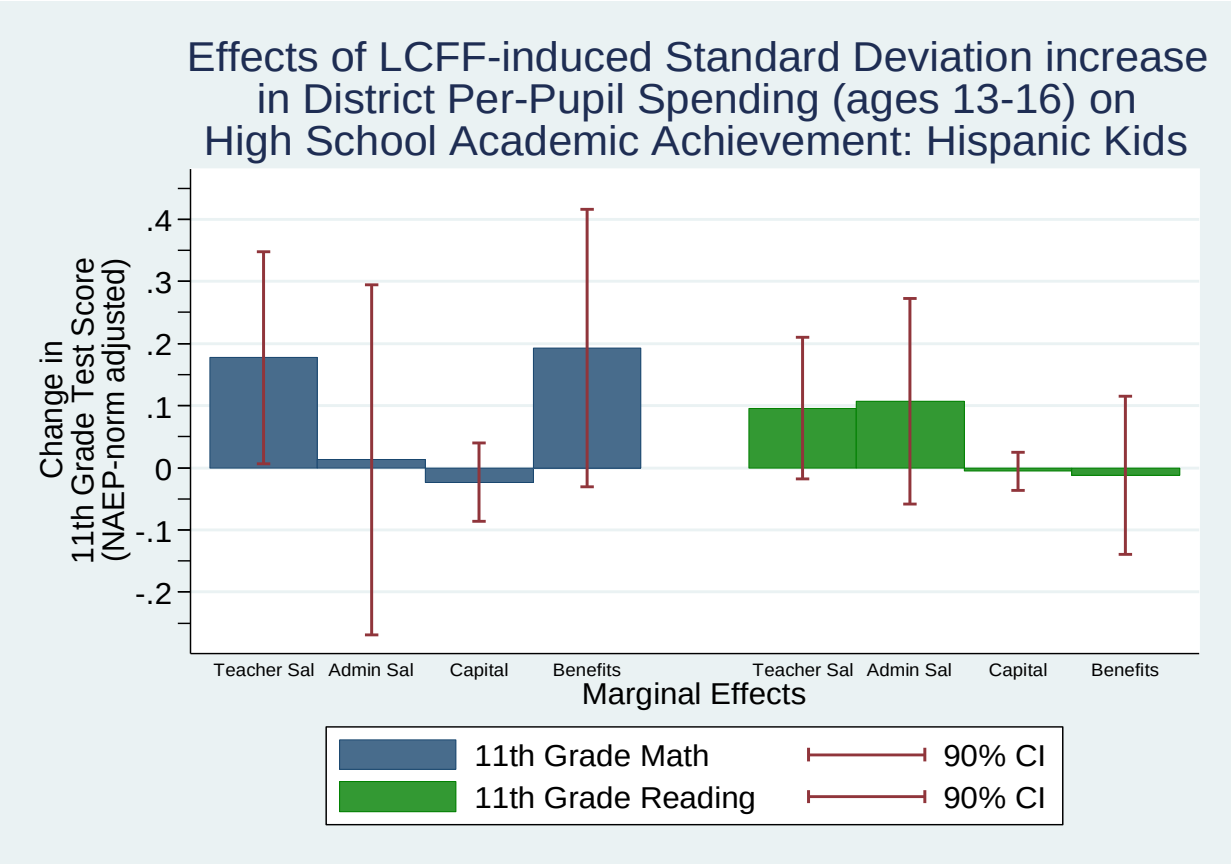


Figure 20: Effect of Expenditures by Category: Hispanic Children



Evidence of False Positives in Research Clearinghouses and Influential Journals: An Application of P-Curve to Policy Research

Section I - Introduction

Recent work has uncovered a troubling pattern in social science research: statistically significant results can often be attributed to selective reporting of results within studies, or "p-hacking" (Ioannidis, 2005; Simonsohn, Nelson, & Simmons, 2014). P-hacking occurs when a researcher, consciously or not, adjusts his model specifications or analysis sample in numerous ways until he finds a significant coefficient on the explanatory variable(s) of interest. This practice is given numerous labels, including "significance chasing," "specification search," "selective reporting," and "massaging the data." Implicit and explicit accusations of such behavior have generated heated controversy over numerous research questions, including which studies merit inclusion in a meta-analysis (Krueger, Hanushek, & Rice, 2002), how to define an African-American sub-sample from survey data (W. G. Howell, Wolf, Campbell, & Peterson, 2002; Krueger & Zhu, 2004), and how to dichotomize streams based on their width (C. Hoxby, 2007; Rothstein, 2007) to name but a few.[31] In each of these cases, the salient effect changed markedly when an alternative, equally defensible coding choice was employed, leading to partisan disagreements rather than scientific advancement.

The specter of p-hacking is worrisome in applied social science research, the results of which are relied on by policy makers in their deliberations and funding decisions (Bogenschneider & Corbett, 2010; Gueron & Rolston, 2013; Haskins & Margolis, 2015).[32] To take one example, the Obama administration's $1.1 billion "Investing in Innovation Initiative" allows $50 million for each educational program that has been validated by a single randomized control trial (Haskins & Margolis, 2015, Chapter 4). Though laudable in its effort to promote high internal validity, this funding rule incentivizes p-hacking by allocating substantial resources on the basis of a single statistically significant result. The aforementioned examples of p-hacking controversies were taken from research on the impacts of class size, school vouchers, and school competition on student achievement. Similar lists could be generated for policy-related research on labor, crime, health, and the environment. If p-hacking plagues such research, then the information social scientists provide to the policy community is unreliable. It is therefore of first order importance to investigate the reliability of policy-related scholarship. Until recently, however, it was difficult to distinguish p-hacking in empirical social science research from other forms of publication bias with less pernicious consequences for the accuracy of published research, such as journal editors' preference for rejection of null hypotheses. This article employs a novel estimation framework, "p-curve", to assess the veracity of reported effects in three influential policy-oriented journals and two research clearinghouses maintained

[31] The reference to these is meant to underscore the existence of controversy, not to imply that the original authors p-hacked their data.

[32] This is not a claim that policy makers rely *sufficiently* on academic research or that they rely on it as heavily as social scientists would prefer.

by the federal government. P-curve allows the analyst to distinguish amongst real effects, p-hacking, and false positives that occur by chance.

Section II - False Positive Social Science

Concern over dubious research practices is not new to the social sciences. Decades have passed since Leamer's (1983) foundational article on spurious causal inference in econometric work and Rosenthal's (1979) sobering illustration of the sensitivity of meta-analysis to unpublished null findings. The ensuing years have witnessed a vigorous debate on the merits of randomized and natural experiments in the production of durable scholarly knowledge (cite that debate), as well as a codification of Rosenthal's concerns in meta-analytic literature.[33] The hegemonic view in many disciplines is that of a "credibility revolution," wherein rigorous methods are allowing for more valid and reliable research than that produced with older methods (Angrist & Pischke, 2010).

A distressing realization has recently tempered the celebration of improved research design in empirical social science. The freedom of a researcher to choose amongst numerous specifications, variables, and sub-samples within a single study can lead to spurious results even when the method meets rigorous standards. This practice goes by many names: massaging the data, specification search, significance chasing, selective reporting. Following Simonsohn, Nelson, and Simmons (2014), the term p-hacking is used here to define attempts to generate statistically significant results from a distribution with a true null hypothesis. The following hypothetical examples illustrate, but are not exhaustive of, p-hacking: analyzing sub-groups (race, SES quintiles, gender), multiple dependent variables, or multiple estimators until a significant treatment effect is found. The operative phrase here is "until a significant treatment effect is found" as it underscores that p-hacking occurs when a researcher mines the data for significant effects, stops when they are found, and does not incorporate the full search procedure into p-value calculations. Doing so will cause p-values to cluster around significance thresholds, usually assumed to be .05 or .1. The clustering occurs because it is easier by chance to find a p-value that crosses .05 (or .1) than one that crosses more stringent significance thresholds such as .01. In the case of a single draw from a distribution with a true null hypothesis, a researcher is five times more likely to find a p-value between .05 and 0 than a p-value between .01 and 0. In 40 trials (hypothesis tests), a researcher is ten times as likely to find at least two p-values that cross .05 as he is to find at least two p-values that cross .01.

A number of articles provide evidence of clustered p-values in prominent sociology, political science, psychology, and economic journals. Gerber and Malhotra (2008a) find a three-fold jump in the number of p-values just under .05 in statistics taken from the *American Political Science Review* and the *American Journal of Political Science*. In a separate analysis of statistics taken from the *American Journal of Sociology* and the *Sociological Quarterly*, they find a two-fold jump in p-values that just cross the .05 threshold (2008b). Brodeur, Le, Sangnier, and Zylberberg (2013) find similar clustering around the .05 and .1 levels of significance in statistics taken from the *Quarterly Journal of Economics*, the *American Economic Review*, and the *Journal of Political Economy*. Psychology journals display the same general pattern, with "excess" p-values just below the .05 threshold in the *Journal of Experimental Psychology*, the

[33] The Cochrane Handbook does not recommend Rosenthal's particular solution of calculating a fail-safe N (2008, pp. 321–322), yet opens the section on publication bias with reference to Rosenthal's concern and suggests many alternatives to address it (2008, Chapter 10).

Journal of Personality and Social Psychology, and *Psychological Science* (Masicampo & Lalande, 2012).

Taken together, these four analyses provide evidence of publication bias in leading social science journals across several disciplines. However, as the studies' authors acknowledge, it is difficult to distinguish amongst various forms of publication bias with the strategies employed therein. Each article analyzes the distribution of test statistics across the .05 and .1 significance thresholds, comparing the relative representation of statistics close to either side of each boundary. An overrepresentation of test statistics on the significant side of the boundary is taken as evidence of publication bias. If journal editors favor articles that reject a null hypothesis, then one would find this empirical pattern whether or not the reported effects are real or artifacts of p-hacking. The consequences for the veracity of policy research are drastically different depending on which behavior is leading to the overrepresentation of statistically significant results. If all effects are real, journal editorial policies favoring null hypothesis rejection would censor true null results but still allow real effects into the literature. The research community would fail to learn about null effects, but would learn about true effects. However, if the effects reported are spurious artifacts of p-hacking or sampling variance, then the research community fails to learn anything, regardless of journal editors' preference for rejected null hypotheses.

An additional limitation of the previous work is that the authors included numerous p-values from each article they analyzed, leading to potential clustering of statistics at the study level and in some cases the inclusion of p-values from statistical tests that would not likely have been subjected to p-hacking. For example, though Brodeur et al. (2013) took pains to include only p-values from the main causal analyses in economics journals, they collected an average of over 50 p-values per article. Their dataset includes p-values from robustness checks, naïve OLS regressions in advance of more sophisticated designs, and the like. These analyses are not likely to suffer from p-hacking and are not the primary results of the articles. Including such p-values dilutes the proper analysis sample and will underrepresent p-hacking, as has recently been confirmed with the original analysis sample from the article that introduced the p-curve framework (Simonsohn, 2015).

Section III - Assessing the Evidential Value in a Body of Literature

The introduction of p-curve by Simonsohn, Nelson, and Simmons (2014) has made it possible to distinguish true effects from those that result specifically from p-hacking or sampling variance using only the observed results that cross a particular significance level. The null hypothesis begins with an assumption that all null hypotheses are true, which would create a uniform distribution of p-values, where the relative frequency of p-values between .05 and .04 is the same as the relative frequency of p-values between .03 and .02, which is also the same as the relative frequency of p-values between .02 and .01, and so on. The choices of bin width and significance thresholds do not change the uniformity of relative frequencies so long as the bin width is constant across to support of observed p-values that cross a particular significance threshold.

False null hypotheses (real effects) generate a right-skewed distribution of p-values, with more frequent values in the range .01 to .02 than .04 to .05. The greater the statistical power of individual studies, the more right-skewed a p-curve that results from those studies. There is no natural distribution of p-values that is (asymptotically) left-skewed: either the distribution is uniform (all null hypotheses are true) or is right-skewed (at least some null hypotheses are false). A left-skewed distribution of p-values can only occur under p-hacking, whereby the researcher

keeps altering the analysis until a significant p-value is found. This creates clustering of p-values around the chosen significance threshold and a left-skewed distribution.

This estimation framework has been used to detect spurious results in the *Journal of Personality and Social Psychology* (Simonsohn et al., 2014) and support the evidence in a foundational set of studies on anchoring (Simonsohn & Nelson, 2014). The research suspected to have been p-hacked generated a left-skewed distribution, whereas the research suspected to contain evidence generated a right-skewed distribution.

The logic of p-curve is somewhat intuitive: real effects should generate lower p-values more often than higher p-values. This logic underlies the estimation strategy favored by Simonsohn et al. (2014). In the first stage, each significant p-value is transformed into the probability of observing such a value at least as extreme if the null were true. For continuous tests, this is merely the p-value divided by the significance threshold chosen by the p-curve analyst (for example .1). This transformed value is labeled the pp-value by Simonsohn et al. (2014). The second stage aggregates the pp-values using Fisher's method in which -2 multiplied by the sum of the natural log of k uniform distributions is distributed $\chi^2(2k)$ under the null hypothesis that the pp-values are uniformly distributed (corresponding to no evidential value) (Fisher, 1932).

Section IV- Study Design

The entire study design and analytic structure has been articulated in a pre-analysis plan (PAP), which was published prior to the data collection (Tanner, 2015).[34] Several analytic decisions were altered after the PAP was published and are detailed in a deviation report. Such deviations are also denoted with "^" and an accompanying footnote in this analysis. The PAP, deviation report, data, and analysis code are all available on the author's website.[35]

Because of the well-known limitations of Fisher's method (Rosenthal, 1978, pp. 187–188)[36], the Kolmogorov-Smirnov (KS) test is the primary estimation technique used in this analysis to measure goodness of fit for a uniform distribution of p-values.[^] The KS test statistic is the maximum distance between the cumulative empirical distribution and the cumulative density of the given null distribution, which in this case is a discretized uniform distribution on the [0, .1] interval (Massey, Jr., 1951). In simulations, the KS test had approximately 80% power to reject a discretized uniform null with a sample of 60 studies powered at 33%. A second estimation strategy is also employed to ensure results are not local to the KS test. A simple method of estimating evidential value is to dichotomize p-values as low ($0<p<.05$) and high ($.05<p<.1$) and submit the data to a binomial test with a uniform null (50% low). This procedure is resistant to extreme values, simple, and transparent, yet it ignores variation of p-values within the high and low categories. This simple method is used as a robustness check on the primary analysis.

In testing multiple hypotheses, individual p-values from each KS and binomial test cease to have their intended meaning and underestimate the risk of false positives. Consequently, the

---

[34] Note that the PAP was submitted in October, 2014.

[35] The Open Science Framework web address for this author is https://osf.io/mn8gr/

[36] As of this writing, the original p-curve authors no longer use Fisher's method as their preferred technique, but instead use Stouffer's method (see note under "Computer Programs" here: http://www.p-curve.com/Supplement/ as well as page 22 of http://p-curve.com/Supplement/Supplement_pcurve1.pdf). They note that the technique employed here, the Kolmogorov-Smirnov test, should be considered as well in larger samples of p-values than they typically use (≈20).

[^] The PAP specified Fisher's method in the author's own code and the p-curve web app 2.0.

method of controlling the false discovery rate (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001) is employed here to further guard against spurious results.

Section V- Data Collection

This analysis evaluates the *evidential value* of recent policy-related scholarship in influential academic journals and research clearinghouses. Using the definition adopted by Simonsohn et al. (2014, p. 3), a set of findings in a body of research contains evidential value when p-hacking can be ruled out as their sole explanation. The sources were chosen as they are believed to be repositories of rigorous, policy-related analyses. If scholarship in these sources does not contain evidential value, then the credibility of applied social science research can be seriously questioned.

Applying p-curve to a set of studies requires selecting the studies and the sub-set of p-values from them to include in the analysis sample. Because this process is itself open to the p-hacking that p-curve is meant to uncover, a pre-specified data collection and estimation strategy are included in the published PAP.

The unit of analysis is conceptually a study but operationally a p-value, as usually only one p-value can be used from each study.[37] The sampling frame of studies was chosen to fit three primary criteria: (1) the perception of rigor, (2) policy-relevance, and (3) recency. These criteria allow for an analysis of the current state of prestigious, applied social science research and are features of research onto which policy makers have explicitly placed high value (Bogenschneider & Corbett, 2010, p. 35). The sampling frame is also narrow enough to allow for a census of studies that meet the following rules. All studies chosen include estimation of causal parameters in real data. Purely descriptive papers and modeling exercises were not included.

The studies taken into the sample were:

A. The 20 most recent articles in the *Journal of Policy Analysis and Management* (*JPAM*), *The Journal of Human Resources* (*JHR*), and *Education Evaluation and Policy Analysis* (*EEPA*).[^]

B. The entire group of studies included in the What Works Clearinghouse's (WWC) Single Study Review section in the following categories that meet the clearinghouse's standards of evidence (with and without qualifications): http://ies.ed.gov/ncee/wwc/

   a. Dropout Prevention

   b. Early Childhood Education

   c. Postsecondary Education

[37] The shape of p-curve relies on the assumption that all p-values are drawn from independent tests.

[^] The journal data were collected on March 9, 2015. *JPAM* was oversampled (31 articles) in order to span the editorial leadership change that began with Volume 33, Issue 4 in the fall of 2014.

d. School Choice

e. Teacher Incentives

C. The entire group of causal studies included in the Clearinghouse for Labor Evaluation and Research (CLEAR) database that have been labeled high or moderate strength.

http://clear.dol.gov/study_database[38]

The journals were chosen as representative of influential, policy-oriented scholarly research. *JPAM* is the sole journal published by the Association for Public Policy Analysis and Management and is widely considered to be the flagship journal in policy research. *JHR* and *EEPA* were both identified as having authors and readership similar to *JPAM* (Reuter & Smith-Ready, 2002).

The two clearinghouses were chosen for similar reasons – they were created to serve as repositories for high quality research. WWC is maintained by the Department of Education's Institute for Educational Sciences as a "central and trusted source of scientific evidence for what works in education to improve student outcomes" (Institute of Educational Sciences, Department of Education, 2014). As is typical of research clearinghouses, the WWC assigns standards ratings to individual studies. The studies used in this analysis met the WWC standards "without reservation" or "with reservation." In order to meet the WWC standards without reservation, "study participants must have been placed into each study condition through random assignment or a process that was functionally random", whereas the rating "meets standards with reservation" is applied to studies whose participants were not placed into study conditions through random assignment but can demonstrate baseline equivalence (Institute of Educational Sciences, Department of Education, n.d.). The ratings used by CLEAR, high causal evidence and moderate causal evidence, are determined in a similar manner and correspond to the WWC ratings of without reservation and with reservation, respectively (Department of Labor, 2014).

The p-values analyzed in these studies are those from the tests of the main hypotheses. All such p-values that cross the .1 significance level were analyzed.[39] This level of significance is justified for two reasons. First, it is common in this literature to call attention to "marginally" significant results for which the p-value falls between .05 and .1. Such results are publishable by most standards of policy-oriented social science research. It is therefore highly plausible that a researcher would p-hack the data until p-values cross the .1 or the .05 levels. Second, there is some evidence that significance levels cluster around both the .1 and .05 levels (Brodeur et al., 2013, p. 29).

The determination of "main" can be ambiguous when authors present multiple model specifications, analyze multiple dependent variables, and assess heterogeneity in treatment effects. This analysis used p-values that were mentioned in the abstract or introduction section

[38] The clearinghouse data were collected on December 18, 2014. Since then, new studies have been added to both clearinghouses.

[39] The set of robustness checks using Stouffer's method only use p-values of .05 or less and the results do not markedly differ (see footnote 9).

and from specifications favored by the authors. When this did not resolve ambiguity, a single p-value was drawn at random from the list of candidates for "main" hypothesis.[40] The median number of hypotheses in each article is 11 for the journal frame and 3 for the clearinghouse frame. Table 1 includes descriptive statistics for the analysis sample.

Exact p-values, test statistics, standard errors, and degrees of freedom are not uniformly reported in this literature. Comparing p-values in the face of this variation in reporting required the use of rounding.[^] Most often, p-values were reported to two decimal places. If possible, exact p-values were calculated using the test statistics, standard errors, and degrees of freedom in the relevant tables. These precise p-values were rounded to two decimal places to facilitate comparison with the format in which the majority were reported in the original studies. Similarly, it was often unclear whether tests were discrete or continuous. Given the large sample sizes in this literature, a decision was made to treat all tests as continuous.[^]

Section VI - Results

The distribution of p-values in the sample provides strong graphical evidence that the reported effects are not artifacts of selective reporting or random sampling variance. Figure 1 shows the histogram for the sample as a whole. The bins reveal a strong right skew and thus high evidential value. There are far more p-values in the neighborhood of .01 than .05 or .1. The dashed red line is how the distribution would appear if the effects contained in the articles were purely artifacts of random sampling variance: an even distribution with 10% in each bin. It is visually clear that the distribution of p-values is neither uniform nor skewed left, which formal hypothesis tests confirm. Both the Kolmogorov-Smirnov (KS) and binomial tests reject uniformity with $p<.001$.

These graphical patterns are broadly similar across all pre-specified subdivisions of the data: RCT vs. non-RCT, high vs moderate evidence ratings, clearinghouses vs. policy journals. Figure 16 contains the histograms for these divisions of the data. As is visually clear, the presence of p-hacking (left-skew) can be readily rejected with both KS and binomial tests. Table 2 contains the p-values for the tests against a uniform null for all groupings of the data as well as the sharpened q-values for each test after adjusting the entire set of tests[41] for the false discovery rate (Anderson, 2008; Benjamini & Yekutieli, 2001). The highest (least significant) sharpened q-value is .0256; most are less than .0001, providing compelling evidence that the results contained in this sample and all subdivision of it have "evidential value." More precisely, p-hacking can be ruled out as the sole explanation for the findings in this sample.

These results are reassuring for the validity of the research contained in this sample, yet several limitations are worth mentioning. 16% of studies (20) fail to report sufficient information to obtain p-values directly. 17 of those 20 report significance levels. These studies are included in the following robustness check wherein those reporting only significance levels are given the p-value of the significance level (10%=.1, 5%=.05, 1%=.01) and the three studies

---

[40] The precise decision rules for inclusion are contained in the pre-analysis plan. The data is available online and includes how many "main" p-values were reported as well as the order that the randomly chosen p-value appeared. In hindsight, each "main" p-value should have been recorded and the analysis should have been bootstrapped to estimate variance due to which p-values were chosen.

[^] The PAP made no reference to rounding.

[^] The PAP specified demarcating continuous and discrete tests, with separate p-curves for continuous tests and all tests.

[41] The sharpened q-values were calculated using all 28 p-values in the results section of the article, even the subsequent "worst case scenario" tests.

without readily available significance information were given a p-value of .1. This procedure will overstate the amount of clustering at significance thresholds as it assigns the right-most (highest) possible p-value to each missing entry- the worst-case p-hacking scenario. By design, these graphs reveal slightly more clustering around common significance thresholds. However, they do not alter the essential finding that this group of policy research is not overly plagued by either accidental false positives or p-hacking. Most of the distributions can be distinguished from uniform with $p<.001$ even in this worst-case analysis.

The second limitation is that the sample may not be reflective of work in the larger population of policy journals and research clearinghouses. As of this writing, there are more than 20 government-supported evidence reviews that could broadly be classified as research clearinghouses (Neuhoff, Axworthy, Glazer, & Berfond, 2015, pp. 51–53). However, many of these are specific to a narrow policy issue or location such as the Best Practices Registry for Suicide Prevention[42] and the California Healthy Kids Resource Center.[43] The two used in this analysis, WWC and CLEAR, are national in scope and topically broad. WWC in particular is the eldest and most established of its peers and should provide a representative case of a broad, developed, national clearinghouse. The policy journals were also chosen as representative of rigorous policy research, yet that definition is highly subjective and excludes many policy-related articles that appear in discipline-specific social science journals. It is probable that a different set of journals would contain a distribution of p-values with less right-skew (less credible evidence).

The third limitation of this article is that the sampling procedure collected only a single p-value from each article. This is a limitation of the p-curve framework in general, as analyzed p-values must be independent. This condition would almost certainly be violated if multiple p-values from a single article were included. The constraint of including only a single p-value from each article requires judicious decisions about what constitutes a primary hypothesis in addition to requiring a random selection in the face of multiple hypotheses that fit the definition. It could be the case that the decision rules employed here are too broad and that the p-hacking behavior in this literature is focused only on a sub-set of the tests that have been labeled as primary. If so, these results will understate the extent of p-hacking. This is a current tension in the literature on publication bias.

Finally, it should be stressed that the analysis is not a test of other forms of fraud or sophisticated p-hacking routines. If researchers purposefully altered their data, reported coefficients, or standard errors, then this analysis is based on fraudulent data and says only that the fraudulent data did not produce clustered test statistics.[44] Moreover, if researchers were p-hacking to cross .01 rather than .05 or .1, then this analysis will understate the extent of p-hacking.

Section VII- Discussion

This article finds encouragingly little evidence of false-positives or p-hacking in rigorous policy research. The findings are somewhat surprising given that prior work in several prestigious scientific disciplines has suggested publication bias is a large problem. Unlike the prior work, this analysis uses a method that can disentangle p-hacking from accidents of random sampling. It also analyzes a set of p-values in each article that are more likely to have been p-hacked. Leaving aside the issue of extrapolation to a more general body of policy scholarship,

[42] http://www.sprc.org/bpr
[43] http://www.californiahealthykids.org/index
[44] For a recent example of fraudulent research detected due to statistical abnormalities, see Simonsohn (2013).

these results are a welcome sign of credibility in light of recent scandals (Bhattacharjee, 2013; Carey, 2015; Godoy, 2015).

Table 1: Descriptive Statistics for the Analysis Sample

	Journals	Clearinghouses
Total # of Articles	71	75
# of Articles Reporting Effects	52	68
Statistics for those Reporting Effects		
# of Main Hypotheses- Mean (Med)	15 (11)	10 (3)
# Reporting Precise P-values	49	51
RCTs	11	49
Sample Size Mean (Med)	304,180 (8,523)	88,484 (1,160)
Published in Journal	75[a]	45
Publishing Date Range (Med)	2014-2015 (2014)	1979-2014 (2011)

[a] *The clearinghouse sampling frame included articles published in journals, leading to more journal articles than were collected in the journal sampling frame.*

Figure 1: Strong Evidence in Policy Research

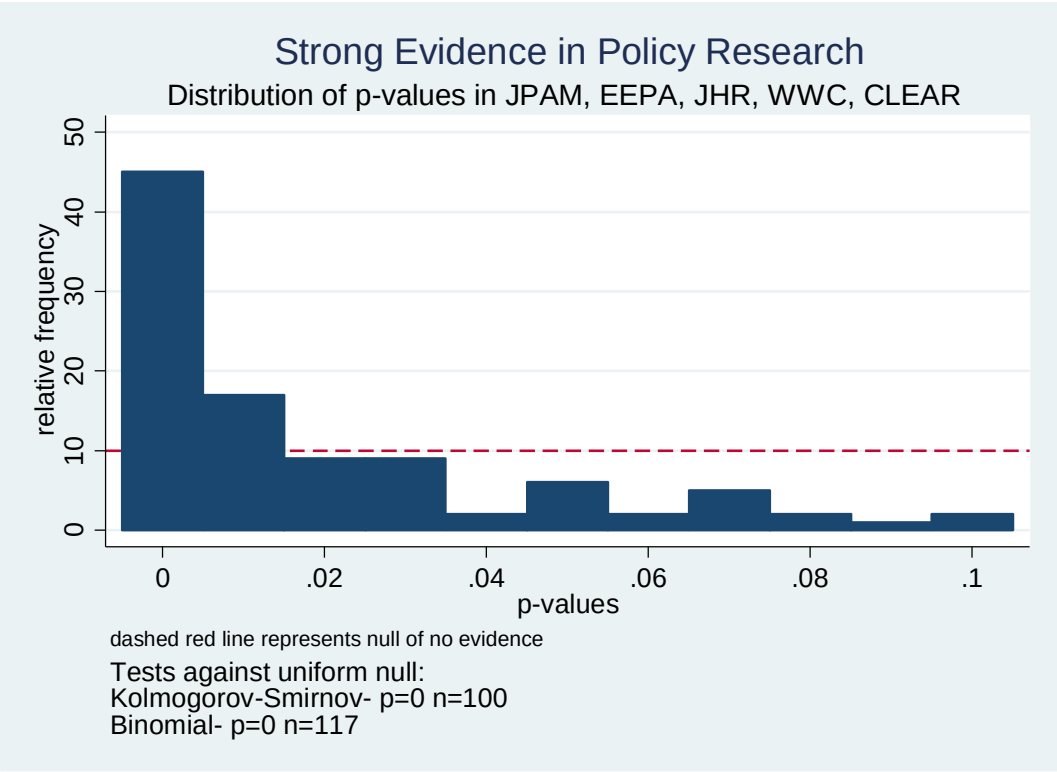


Figure 2: Robust Binomial Method Reveals Strong Evidence

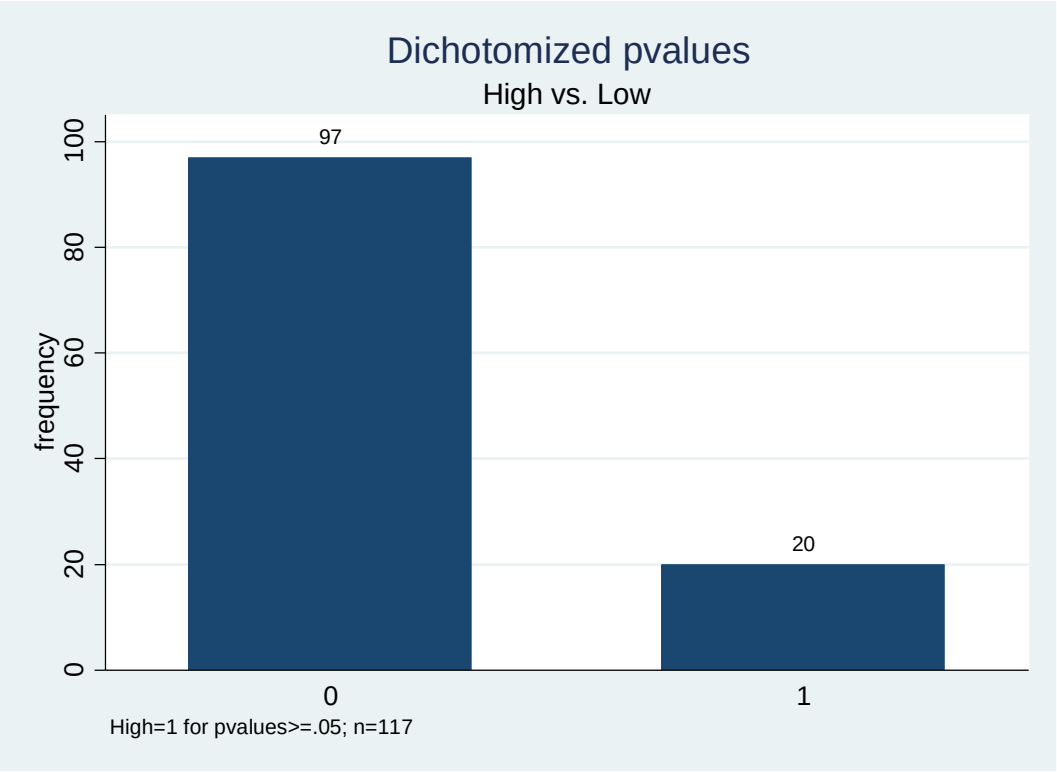


Table 2: P-values and sharpened q-values for the primary hypothesis tests.

Data	KS p-value	Binomial p-value	KS q-value	Binomial q-value
Entire Sample	<.0001	<.0001	<.0001	<.0001
Journals	<.0001	<.0001	<.0001	.0004
Clearinghouses	<.0001	<.0001	<.0001	<.0001
RCTs	<.0001	<.0001	<.0001	<.0001
Non-RCTs	<.0001	<.0001	<.0001	<.0001
Rating: Strong	.0004	<.0001	.002	<.0001
Rating: Moderate	.0128	.0015	.0256	.0046

Conclusion

Taken as a whole, the results contained in the preceding chapters provide reasons for cautious optimism in education policy and the social science that seeks to influence it. Despite well-founded fears, No Child Left Behind does not appear to have seriously harmed the non-cognitive skills of students targeted by the policy- at least as measured by teachers at the end of elementary school. Better still, California's recent school finance reform has so far been successful in strengthening the achievement of disadvantaged students during their secondary schooling years. The final chapter offers evidence that policy research such as this is not unduly burdened by false positives.

Several avenues for future research naturally present themselves. As for NCLB, the divergence between parents' and teachers' reports of non-cognitive skills suggests that gains in social emotional development from education policy might not spill over into home behavior. This underscores the importance of assessing differences across multiple measures of non-cognitive skills (Renk & Phares, 2004), particularly if they are to be components accountability or value-added regimes. At least several measures of non-cognitive skills should be included in state administrative data sets so that these skills can be measured more reliably than serendipitous panel studies allow. Doing so would allow research on the permanency of policy-induced changes in these skills. Future research should also establish the stability of the psychometric properties of these measures in the face of accountability and performance pay pressure. At very least, these measures should be routinely collected in educational randomized field trials. Currently, the Institute of Educational Science's What Works Clearinghouse lists 404 interventions focused on literacy and numeracy and only 56 focused on non-cognitive skills, many of which are focused on children with clinical behavior disorders.[45] The emerging research on non-cognitive skills is increasingly suggesting that this imbalance in research priorities is unjustified. As this analysis reveals, a more nuanced picture of NCLB arises when non-cognitive skills are included- future research on education policy should do the same.

Several aspects of the LCFF analysis also warrant further consideration. The first is the relative recency of the policy-- unintended negative consequences such as local recapture might take longer. Though tax payers may not be able to alter their local property tax rates, there are other means through which budget offsets may occur – such as reductions in parental financial support or a reduction in the willingness of voters to approve of parcel taxes. However, it is just as likely that improvement in student achievement will also grow as school and districts adjust to the new funding environment. There is some evidence that district administrators are hesitant to invest in more permanent inputs until they are assured that LCFF will not be repealed. Second, the variables used in this analysis are school- and district-level averages that do not reflect inequality or changes across student groups within schools and districts. The heterogeneity in graduation-rate impacts suggest that within-district resource allocation should be analyzed. Third, the district financial data may not capture the proper mechanisms that enhanced graduation rates, either because the "true" mechanism is not measured by the accounting code or our selective analysis has missed it. Future work should also analyze the effects of LCFF on student achievement in earlier grades as additional years of data become available.

[45] See http://ies.ed.gov/ncee/wwc/, accessed 9-15-15

The results presented in the closing chapter on research credibility open up two main avenues for further exploration: one methodological and the other substantive. Methodologically, it is crucial for the p-curve framework to be able to analyze all substantively import p-values in an article, dependent or not. Research and publishing norms for laboratory experiments in psychology might be amenable to methods that require a single p-value from each article, but economics, political science, sociology, and public policy research routinely has more than ten main hypothesis tests in a single article.[46] One potential solution is to weight articles by the number of p-values they contribute, so that five legitimate p-values from a single paper don't mask five p-hacked p-values from five separate papers. The behavior of such reweighted p-curves with and without p-hacking is not currently known, yet understanding this behavior is critical to the evolution of research on publication bias. At minimum, future p-curve analyses should capture all relevant p-values from each paper and bootstrap the p-curves to estimate their variance in the face of multiple potential p-values from each paper.

Substantively, it is important to understand whether and why the policy research used in this analysis differs from research contained in the prestigious journals analyzed in the prior work. This is best addressed with a two-fold approach. First, the estimation procedures used in this analysis should be applied to a subset of the p-values collected in the prior research from economics, political science, sociology, and psychology journals. A second and ideal step would be to prospectively collect p-values from a large body of social science research across multiple disciplines and analyze the data with p-curves.

While much of this future work will undoubtedly reveal new challenges, it is important to recognize that true progress is being made in both education policy and research. Some of the most common criticisms of accountability and school finance reforms appear unfounded. Accountability systems that focus attention and incentives on underperforming schools and students can indeed enhance academic achievement without commensurate losses in social emotional development. At the same time, student performance can also be elevated by providing districts with greater revenue and autonomy over their expenditures. Finally, credible policy research is indeed accumulating, unburdened by excessive p-hacking, and can therefore be relied on more confidently as social scientists and policymakers alike attempt to enhance human capital accumulation and reduce income inequality.

[46] This relatively simple analysis tested 7 primary hypotheses.

References

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *Review of Economic Studies*, *72*, 1–19.

Adamson, F., & Darling-Hammond, L. (2012). Funding Disparities and the Inequitable Distribution of Teachers: Evaluating Sources and Solutions. *Education Policy Analysis Archives*, *20*(37).

Almlund, M., Duckworth, A. L., Heckman, J. J., & Kautz, T. D. (2011). *Personality Psychology and Economics* (No. 16822). Cambridge. Retrieved from http://www.nber.org/papers/w16822

Anderson, M. L. (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, *103*(484), 1481–1495. doi:10.1198/016214508000000841

Angrist, J. D., & Pischke, J. (2010). The Credibility Revolution in Empirical Economics : How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, *24*(2), 3–30.

Barnett, W. S., Jung, K., Yarosz, D. J., Thomas, J., Hornbeck, A., Stechuk, R., & Burns, S. (2008). Educational effects of the Tools of the Mind curriculum: A randomized trial. *Early Childhood Research Quarterly*, *23*(3), 299–313. doi:10.1016/j.ecresq.2008.03.001

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Royal Statistical Society. Series B (Methodological), 57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, *29*(4), 1165–1188.

Bhattacharjee, Y. (2013, April 28). The Mind of a Con Man. *The New York Times*, p. MM44. New York. Retrieved from http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?pagewanted=all

Bogenschneider, K., & Corbett, T. (2010). *Evidence-Based Policymaking: Insights from Policy-Minded Researchers and Research-Minded Policymakers*. New York: Routledge.

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2013). *Star Wars: The Empirics Strike Back* (No. Discussion Paper Series, Forschungsinstitut zur Zukunft der Arbeit, No. 7268).

Burtless, G. (1996). Introduction and Summary. In G. Burtless (Ed.), *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success.* (p. 316). Washington, D.C.: Brookings Institution Press.

Candelaria, C. A., & Shores, K. A. (2015). *Court-Ordered Finance Reform on Spending and Graduation Rates*.

Card, D., & Krueger, A. B. (1992). Does School Quality Matter ? Returns to Education and the Characteristics of Public Schools in the United States. *The Journal of Political Economy*, *100*(1), 1–40.

Card, D., & Payne, A. A. (2002). School Finance Reform, the Distribution of School Spending, and the Distribution of SAT Scores. *Journal of Public Economics*, *83*(1), 49–82. Retrieved from http://www.nber.org/papers/w6766

Cardine, S. (2013, November 6). State revises school tests. *Los Angeles Times*. Los Angeles. Retrieved from http://www.latimes.com/tsn-vsl-state-revises-school-tests-20131106-story.html

Carey, B. (2015, May 29). Study on Attitudes Toward Same-Sex Marriage Is Retracted by a Scientific Journal. *The New York Times*, p. A16. New York. Retrieved from http://nyti.ms/1FFQkzj

Casey, K., Glennerster, R., & Miguel, E. (2012). RESHAPING INSTITUTIONS : EVIDENCE ON AID IMPACTS USING A PREANALYSIS PLAN. *Quarterly Journal of Economics*, *127*(4), 1755–1812.

Cellini, S. R., Ferreira, F., & Rothstein, J. (2010). THE VALUE OF SCHOOL FACILITY INVESTMENTS : EVIDENCE FROM A DYNAMIC REGRESSION DISCONTINUITY DESIGN. *Quarterly Journal of Economics*, *125*(1), 215–261.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from project star. *Quarterly Journal of Economics*, *126*(4), 1593–1660. doi:10.1093/qje/qjr041

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2013). *Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood* (No. 19424). Retrieved from http://www.nber.org/papers/w19424

Chung, I. H. (2015). Education finance reform, education spending, and student performance: Evidence from Maryland's Bridge to Excellence in Public Schools Act. *Education and Urban Society*, *47*(4), 412–432. doi:10.1177/0013124513498413

Clark, M. A. (2003). *Education Reform, Redistribution, and Student Achievement: Evidence From the Kentucky Education Reform Act*. Princeton, NJ.

Clotfelter, C. T. (2004). *After Brown: The Rise and Retreat of School Desegregation*. Princeton, NJ: Princeton University Press.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review*, *24*(4), 377–392. doi:10.1016/j.econedurev.2004.06.008

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2010). Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects. *Journal of Human Resources*, *45*(3), 655–681.

Clune, W. H. (1994). The Shift from Equity to Adequacy in School Finance. *Education Policy*, *8*(4), 376–394. doi:0803973233

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of Educational Opportunity*. Washington, D.C. Retrieved from http://www.eric.ed.gov/PDFS/ED012275.pdf

Corcoran, S. P., & Evans, W. N. (2015). Equity, Adequacy, and the Evolving State Role in Education Finance. In H. F. Ladd & M. G. Goertz (Eds.), *Handbook of Research in Education Finance and Policy* (2nd ed., p. 676). New York: Routledge.

Corcoran, S. P., Evans, W. N., Godwin, J., Murray, S. E., & Schwab, R. M. (2004). The Changing Distribution of Education Finance, 1972 to 1997. In K. M. Neckerman (Ed.), *Social Inequality* (p. 1017). New York: Russell Sage Foundation.

Darling-Hammond, L. (2007). Evaluating "No Child Left Behind." *The Nation*. Retrieved from http://www.thenation.com/article/evaluating-no-child-left-behind/

Davidson, E., Reback, R., Rockoff, J., & Schwartz, H. (2013). Fifty ways to leave a child behind: Idiosyncrasies and discrepancies in states' implementation of NCLB (No. w18988). *National Bureau of Economic Research*. Retrieved from http://www.nber.org/papers/w18988

Dee, T. S., & Jacob, B. (2010). *The Impact of No Child Left Behind on Students, Teachers, and Schools*. *Brookings Papers on Economic Activity* (Vol. 2010). Washington, D.C. doi:10.1353/eca.2010.0014

Dee, T. S., & Jacob, B. (2011). The Impact of No Child Left Behind on Student Achievement. *Journal of Policy Analysis and Management*, *30*(3), 418–446.

Department of Labor, U. S. (2014). *CLEAR CAUSAL EVIDENCE GUIDELINES, VERSION 1.1*.

Duckworth, A. L., Quinn, P. D., & Tsukayama, E. (2012). What No Child Left Behind leaves behind: The roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *Journal of Educational Psychology*, *104*(2), 439–451. doi:10.1037/a0026280

Duckworth, A. L., & Yeager, D. S. (2015). Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes. *Educational Researcher*, *44*(4), 237–251.

Figlio, D. N., & Ladd, H. F. (2015). School Accountability and Student Achievement. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of Research in Education Finance and Policy* (Second., p. 676). New York: Routledge.

Fisher, R. A. (1932). *Statistical Methods for Research Workers* (4th ed.). Edinburgh: Oliver and Boyd.

Gerber, A., & Malhotra, N. (2008a). Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals. *Quarterly Journal of Political Science*, *3*(3), 313–326. doi:10.1561/100.00008024

Gerber, A., & Malhotra, N. (2008b). Publication Bias in Empirical Sociological Research. *Sociological Methods & Research*, *37*(1), 3–30. doi:10.1177/0049124108318973

Gewertz, C. (2003, September 3). Hand in Hand. *Education Week*, pp. 38–41. Bethesda, MD. Retrieved from http://www.edweek.org/ew/articles/2003/09/03/01social.h23.html

Godoy, M. (2015). *Why A Journalist Scammed The Media Into Spreading Bad Chocolate Science*. *National Public Radio*. Retrieved from http://www.npr.org/sections/thesalt/2015/05/28/410313446/why-a-journalist-scammed-the-media-into-spreading-bad-chocolate-science

Goldhaber, D. (2015). Teachers Clearly Matter, but Finding Effective Teacher Policies Has Proven Challenging. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of Research in Education Finance and Policy* (2nd ed., p. 676). New York: Routledge.

Goldhaber, D., Lavery, L., & Theobald, R. (2015). Uneven Playing Field? Assessing the Teacher Quality Gap Between Advantaged and Disadvantaged Students. *Educational Researcher*, *44*(5), 293–307. doi:10.3102/0013189X15592622

Grissom, J. A., Nicholson-Crotty, S., & Harrington, J. R. (2014). Estimating the Effects of No Child Left Behind on Teachers' Work Environments and Job Attitudes. *Educational Evaluation and Policy Analysis*, *36*(4), 417–436.

Grubb, W. N. (2009). *The Money Myth: School Resources, Outcomes, and Equity*. New York: Russell Sage Foundation Publications.

Gueron, J. M., & Rolston, H. (2013). *Fighting for Reliable Evidence*. New York: Russell Sage Foundation.

Guryan, J. (2001). *Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts* (No. 8269). Cambridge, MA. doi:10.1007/s13398-014-0173-7.2

Hanushek, E. A. (2003). The Failure of Input-based Schooling Policies*. *The Economic Journal*, *113*(485), F64–F98.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, *30*(3), 466–479.

Hanushek, E. A., & Raymond, M. E. (2001). The Confusing World of Educational Accountability. *National Tax Journal*, *54*(2), 365–384.

Hanushek, E. A., Rivkin, S. G., & Taylor, L. L. (1996). Aggregation and the Estimated Effects of School Resources. *Review of Economics and Statistics*, *78*(4), 611–627.

Haskins, R., & Margolis, G. (2015). *Show Me the Evidence: Obama's Fight for Rigor and Results in Social Policy*. Washington, D.C.: Brookings Institution Press.

Heckman, J. J., Pinto, R., & Savelyev, P. (2013). Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review*, *103*(6), 2052–2086.

Higgins, J. P. T., & Green, S. (Eds.). (2008). *Cochrane Handbook for Systematic Reviews of Interventions*. Hoboken, New Jersey: Wiley-Blackwell.

Howell, P. L., & Miller, B. B. (1997). Sources of Funding for Schools. *Future of Children*, *7*(3), 39–50. doi:10.2307/1602444

Howell, W. G., Wolf, P. J., Campbell, D. E., & Peterson, P. E. (2002). School vouchers and academic performance: results from three randomized field trials. *Journal of Policy Analysis and Management*, *21*(2), 191–217.

Hoxby, C. (2007). Does Competition Among Public Schools Benefit Students and Taxpayers ? Reply. *American Economic Review*, *97*(5), 2038–2055.

Hoxby, C. M. (1996). Are Efficiency and Equity in School Finance Substitutes or Compliments? *Journal of Economic Perspectives*, *10*(4), 51–72.

Hoxby, C. M. (2001). All School Finance Equalizations are not Created Equal. *Quarterly Journal of Economics*, *116*(4), 1189–1231. doi:10.1162/003355301753265552

Hyman, J. (2013). *Does Money Matter in the Long Run? Effects of School Spending on Educational Attainment*.

Institute of Educational Sciences, Department of Education, U. S. (n.d.). *What Works Clearinghouse Procedures and Standards Handbook (Version 3.0)*.

Institute of Educational Sciences, Department of Education, U. S. (2014). *About Us: What Works Clearinghouse*. National Center for Education Statistics. Retrieved September 24, 2014, from http://ies.ed.gov/ncee/wwc/aboutus.aspx

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8).

Jackson, C. K. (2012). *Non-Cognitive Ability, Test Scores, and Teacher Quality: Evidence from 9th Grade Teachers in North Carolina* (No. 18624).

Jackson, C. K., Johnson, R. C., & Persico, C. (2015). *Money Does Matter After All*. *EducationNext*. Retrieved November 5, 2016, from http://educationnext.org/money-matter/

Jackson, C. K., Johnson, R. C., & Persico, C. (2016). THE EFFECTS OF SCHOOL SPENDING ON EDUCATIONAL AND ECONOMIC OUTCOMES: EVIDENCE FROM SCHOOL FINANCE REFORMS. *Quarterly Journal of Economics*, *131*(1).

Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*(5–6), 761–796. doi:10.1016/j.jpubeco.2004.08.004

Jacob, B. A., & Levitt, S. D. (2003). Rotten Apples : An Investigation of the Prevalence and Predictors of Teacher Cheating. *Quarterly Journal of Economics*, *118*(3), 843–877.

Junttila, N., Voeten, M., Kaukiainen, A., & Vauras, M. (2006). Multisource Assessment of Children's Social Competence. *Educational and Psychological Measurement*, *66*(5), 874–895. doi:10.1177/0013164405285546

Kirst, M. W., Goertz, M., & Odden, A. R. (2007). *The Evolution of California's State School Finance System and Implications from Other States*. Palo Alto, CA. Retrieved from https://cepa.stanford.edu/content/evolution-california%25E2%2580%2599s-state-school-finance-system-and-implications-other-states

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). EXPERIMENTAL ANALYSIS OF NEIGHBORHOOD EFFECTS. *Econometrica*, *75*(1), 83–119.

Koski, W. S., & Hahnel, J. (2015). The Past, Present, and Possible Future of Educational Finance Reform Litigation. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of Research in Education Finance and Policy* (2nd ed., p. 676). New York: Routledge.

Krueger, A. B., Hanushek, E. A., & Rice, J. K. (2002). *The Class Size Debate*. (M. Lawrence & R. Rothstein, Eds.). Washington, D.C.: Economic Policy Institute.

Krueger, A. B., & Zhu, P. (2004). Another Look at the New York City School Voucher Experiment. *American Behavioral Scientist*, *47*(5), 658–698.

Lafortune, J., Rothstein, J., & Schanzenbach, D. W. (2015). *School Finance Reform and the Distribution of Student Achievement* (No. 22011). Cambridge, MA. doi:10.1017/CBO9781107415324.004

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis. *Educational Evaluation and Policy Analysis*, *24*(1), 37–62. doi:10.3102/01623737024001037

Leamer, E. E. (1983). Let's Take the Con Out of Econometrics. *American Economic Review*, *73*(1), 31–43.

Lee, J., & Reeves, T. (2012). Revisiting the Impact of NCLB High-Stakes School Accountability, Capacity, and Resources State NAEP 1990–2009 Reading and Math Achievement Gaps and Trends. *Educational Evaluation and Policy Analysis*, *34*(2), 209–231. doi:10.3102/0162373711431604

Leuven, E., Lindahl, M., Oosterbeek, H., & Webbink, D. (2007). THE EFFECT OF EXTRA FUNDING FOR

DISADVANTAGED PUPILS ON ACHIEVEMENT. *Review of Economics and Statistics*, *89*(4), 721–736.
Manna, P. (2011). *Collision Course: Federal Education Policy Meets State and Local Realities*. Washington, D.C.: CQ Press.
Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279. doi:10.1080/17470218.2012.711335
Massey, Jr., F. J. (1951). Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, *46*(253), 68–78. doi:10.1080/01621459.1951.10500769
McGuinn, P. J. (2006). *No Child Left Behind and the Transformation of Federal Education Policy, 1965-2005*. Lawrence, KS: University Press of Kansas.
Murnane, R. J., & Willett, J. B. (2011). *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. New York: Oxford University Press.
Murray, B. S. E., Evans, W. N., & Schwab, R. M. (1998). Education-Finance Reform and the Distribution of Education Resources. *American Economic Review*, *88*(4), 789–812.
Neuhoff, B. A., Axworthy, S., Glazer, S., & Berfond, D. (2015). The What Works Marketplace Smarter Choices, (April).
Noguchi, S. (2015, September 9). California's school test scores reveal vast racial achievement gap. *The San Jose Mercury News*. San Jose. Retrieved from http://www.mercurynews.com/california/ci_28782503/califs-test-scores-reveal-yawning-achievement-gap
Papke, L. E. (2005). The effects of spending on test pass rates: Evidence from Michigan. *Journal of Public Economics*, *89*(5–6), 821–839. doi:10.1016/j.jpubeco.2004.05.008
Papke, L. E. (2008). The Effects of Changes in Michigan's School Finance System. *Public Finance Review*, *36*(4), 456–474. doi:10.1177/1091142107306287
Reardon, S. F. (2011). The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations. In G. J. Duncan & R. J. Murnane (Eds.), *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances* (pp. 91–116). New York: Russell Sage Foundation.
Reardon, S. F., Greenberg, E. H., Kalogrides, D., Shores, K. A., & Valentino, R. A. (2013). *Left Behind? The Effect of No Child Left Behind on Academic Achievement Gaps*. *working paper*. doi:10.1227/00006123-197907010-00030
Reardon, S. F., & Ho, a. D. (2015). Practical Issues in Estimating Achievement Gaps From Coarsened Data. *Journal of Educational and Behavioral Statistics*, *40*(2), 158–189. doi:10.3102/1076998615570944
Reback, R., Rockoff, J., & Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy*, *6*(3), 207–241. doi:10.1257/pol.6.3.207
Renk, K., & Phares, V. (2004). Cross-informant ratings of social competence in children and adolescents. *Clinical Psychology Review*, *24*(2), 239–254. doi:10.1016/j.cpr.2004.01.004
Reuter, P., & Smith-Ready, J. (2002). AssessingJPAM after 20 Years. *Journal of Policy Analysis and Management*, *21*(3), 339–353.
Rosenthal, R. (1978). Combining Results of Independent Studies. *Psychological Bulletin*, *85*(1), 185–193.
Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641.
Rothstein, J. (2007). Does Competition Public Schools Benefit Students and Taxpayers? Comment. *American Economic Review*, *97*(5), 2026–2037.
Roy, J. (2011). Impact of School Finance Reform on Resource Equalization and Academic Performance: Evidence from Michigan. *Education Finance and Policy*, *6*(2), 137–167. doi:10.2139/ssrn.630121
Sherlock, M. (2011). The Effects of Financial Resources on Test Pass Rates : Evidence from Vermont ' s

Equal Education Opportunity Act. *Public Fi*, *39*(3), 331–364. doi:10.1177/1091142110396500
Simonsohn, U. (2013). Just post it: the lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*(10), 1875–88. doi:10.1177/0956797613480366
Simonsohn, U. (2015). *Falsely Reassuring: Analyses of ALL p-values*. Retrieved from http://datacolada.org/2015/08/24/41-falsely-reassuring-analyses-of-all-p-values-2/
Simonsohn, U., & Nelson, L. D. (2014). Anchoring is Not a False-Positive : Maniadis , Tufano , and List ' s (2014) " Failure -to- Replicate " is Actually Entirely Consistent with the Original, 1–12.
Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology. General*, *143*(2), 534–47.
Somers, M.-A., Zhu, P., Jacob, R., & Bloom, H. (2012). *The Validity and Precision of the Comparative Interrupted Time Series Design and the Difference-in-Difference Design in Educational Evaluation*. Retrieved from http://appam.confex.com/data/extendedabstract/appam/2012/Paper_1758_extendedabstract_156_0.pdf
Sonstelie, J., Brunner, E., & Ardon, K. (2000). *For Better or For Worse ? School Finance Reform in California*. San Francisco, CA.
Springer, M. G., Houck, E. A., & Guthrie, J. W. (2015). History and Scholarship Regarding U.S. Education Finance and Policy. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of Research in Education Finance and Policy* (2nd ed., p. 676). New York: Routledge.
St.Clair, T., Cook, T. D., & Hallberg, K. (2014). Examining the Internal Validity and Statistical Precision of the Comparative Interrupted Time Series Design by Comparison With a Randomized Experiment. *American Journal of Evaluation*, *35*(3), 311–327. doi:10.1177/1098214014527337
Tanner, S. (2015). False Positives and Selective Reporting in Policy. *Observational Studies*, *1*(1), 18–29.
Timar, T. B. (1994). Politics, Policy, and Categorical Aid: New Inequities in California School Finance. *Educational Evaluation and Policy Analysis*, *16*(2), 143–160. doi:10.3102/01623737016002143
Timar, T. B. (2006). *How California Funds K-12 Education*. Palo Alto, CA. Retrieved from http://irepp.stanford.edu/documents/GDF/STUDIES/02-Timar/2-Timar(3-07).pdf
Vinovskis, M. A. (2009). *From a Nation at Risk to No Child Left Behind: National Education Goals and the Creation of Federal Education Policy*. New York: Teachers College Press.
Weston, M. (2011). *California's New School Funding Flexibility*. San Francisco, CA.
Wolf, R., & Sands, J. (2016). A Preliminary Analysis of California's New Local Control Funding Formula. *Education Policy Analysis Archives*, *24*(34), 1–37.
Wong, M., Cook, T. D., & Steiner, P. M. (2011). *No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series* (No. WP-09-11). *Institute for Policy Research*. Evanston, IL. Retrieved from http://www.ipr.northwestern.edu/publications/workingpapers/2009/wp0911.pdf