



Select Works on the Economics of Education

Citation

Lukes, Dylan John. 2022. Select Works on the Economics of Education. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences.

Link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37371978>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#)

HARVARD UNIVERSITY
Graduate School of Arts and Sciences



DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Education

have examined a dissertation entitled

Select Works on the Economics of Education

presented by Dylan J. Lukes

candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Signature _____

Typed name: Professor David Deming

Signature _____

Typed name: Associate Professor Eric Taylor

Signature _____

Typed name: Professor Andrew Ho

Date: 04/29/2022

Select Works on the Economics of Education

A Dissertation

presented by

Dylan J. Lukes

to

The Committee on Higher Degrees in Education

in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

in the subject of

Education

Harvard University

Cambridge, Massachusetts

April 2022

© 2022 Dylan J. Lukes
All Rights Reserved.

Select Works on the Economics of Education

Abstract

This dissertation comprises three essays in the field of economics of education. The first essay studies the short and long-run effects of Head Start, a federally funded early childhood education program that targets children from low-income families. This research replicates and extends previous evaluations of Head Start's impact on life cycle skill formation. My co-authors and I find primarily negative impacts of Head Start for more recent birth cohorts and null impacts of Head Start on school-age and early adulthood outcomes for all birth cohorts. Unpacking these results, we show that factors unrelated to Head Start, such as the mother's age at their child's birth, likely play an important role in mediating Head Start impacts across time.

The second essay explores the association between historic neighborhood inequality and modern-day educational outcomes. In this research, my co-author and I link 1935-1940 Home Owners' Loan Corporation A-D security ratings to present-day public schools and districts to determine whether educational institutions located in historically “D” neighborhoods have worse modern-day outcomes than those located in “A,” “B,” or “C” neighborhoods. In doing so, we find that schools and districts in “D” neighborhoods have, on average, less overall funding, student diversity, and worse academic performance than their “A,” “B,” and “C” peers.

The third essay evaluates the impact of the New York City Department of Education's cellphone ban on school discipline and school culture. Overall, I find that NYCDOE's removal

of their districtwide ban reduced school suspensions but also worsened student perceptions of school culture across various dimensions, including respect, student behavior, and school safety. These results show that while the removal of the New York City Department of Education's cellphone ban led to less punitive disciplinary measures in schools, it came at the cost of impaired student perceptions of school culture.

Table of Contents

Title Page	i
Copyright	ii
Abstract	iii
Table of Contents	v
Acknowledgements	ix
Chapter 1: Elusive Longer-Run Impacts of Head Start: Replications Within and Across Cohorts	1
1.1 Introduction	1
1.2 Present Study	4
1.3 Method	8
1.3.1 Head Start Program Background	8
1.3.2 Data	9
1.3.3 Family Background Statistics	10
1.3.4 Outcomes	11
1.3.5 Empirical Strategy	15
1.4 Results	19
1.4.1 Head Start Impacts on ASI	19
1.4.2 Head Start Impacts on Longer-Run Outcomes	22
1.4.3 Head Start Impacts on School-Age Outcomes	24
1.4.4 Robustness Checks	27
1.4.5 Reconciliation of Cross-Cohort Results	30
1.5 Discussion and Conclusion	34
Chapter 2: A Lingering Legacy: The Relationship Between 1930s HOLC “Redlining” Maps and School Funding, Diversity, and Performance	38
2.1 Introduction	38
2.2 Literature	42
2.2.1 HOLC Maps History	42
2.2.2 Modern Outcomes Linked to HOLC Maps	44
2.2.3 The Relationship Between Schools, Neighborhoods, and Student Outcomes	46
2.3 Data	48
2.3.1 HOLC Data	48
2.3.2 NCES Geospatial Data	49
2.3.3 NCES Non-Geospatial Data	50
2.3.4 National Education Resource Database on Schools Data	51
2.3.5 Stanford Education Data Archive Data	51
2.4 Analysis Samples	52
2.4.1 School-Level Analysis Samples	53
2.4.2 District-Level Analysis Sample	56
2.5 Empirical Strategy	58
2.6 Results and Discussion	60

2.6.1 District and School Finance	60
2.6.2 School Racial Diversity	72
2.6.3 School Student Performance	79
2.7 Conclusion	83
Chapter 3: The Impact of Lifting NYCDOE's Cellphone Ban on School Discipline and School Culture	87
3.1 Introduction	87
3.2 Literature	90
3.2.1 Computers and Learning	90
3.2.2 Smartphones and Learning	92
3.3 Data	93
3.3.1 Administrative Data	94
3.3.2 School Characteristics	95
3.3.3 Sample	100
3.4 Empirical Strategy	103
3.4.1 Traditional 2x2 DiD Models	103
3.4.2 Threats to Identification	106
3.5 Results and Discussion	111
3.5.1 Main Results	111
3.5.2 Discussion	114
3.5.3 Sensitivity Analysis	119
3.6 Conclusion	124
Appendix	127
Appendix A: Additional Tables and Figures from Chapter 1	127
Appendix B: Additional Tables and Figures from Chapter 2	129
Appendix C: Additional Tables and Figures from Chapter 3	140
Appendix D: Co-Authorship Statement	147
References	149

List of Tables

1.1 Household Characteristics Averaged Over Head Start and No Preschool Status	14
1.2 Head Start Impacts on Cohorts' Adulthood Summary Index	19
2.1 Summary of Analysis Samples by Outcome Groups	54
2.2 District Per-Pupil Revenues and 1935-1940 HOLC A-D, SY2017-18 (USD 2018)	61
2.3 School Per-Pupil Expenditures and 1935-1940 HOLC A-D, SY2018-19 (USD 2018)	63
2.4 Margins & Differences Over Time, Finance Outcomes, 1989-2018 (USD 2018)	71
2.5 School Racial Diversity and 1935-1940 HOLC A-D Grades, 2018-2019	72
2.6 Margins & Differences Over Time, Racial Diversity Outcomes, 1988-2019	78
2.7 School Student Performance and 1935-1940 HOLC A-D Grades, Pooled 2008-2019	80
3.1 NYCDOE School Survey Questions on School Culture, Teachers & Students	96
3.2 Descriptive Statistics for NYC Public Middle and High Schools	101
3.3 Descriptive Statistics Pre- and Post-Policy Change	102
3.4 Impact of NYCDOE Cellphone Ban on Suspensions and School Culture (Students)	113
3.5 Impact of NYCDOE Cellphone Ban on School Culture (Teachers)	113
3.6 Impact of NYCDOE Cellphone Ban on School Safety and Behavior by Individual Outcome (Student)	118
3.7 Impact of NYCDOE Cellphone Ban on School Culture, Cutoff Comparison	120
B.1 Diversity Sample Overview: All 2018-19 U.S. Public Schools vs. HOLC Sample	130
B.2 SEDA Sample Overview: All 2017-18 U.S. Public Districts vs. HOLC Sample	131
B.3 Finance Sample Overview: All 2017-18 U.S. Public Districts vs. HOLC Sample	132
B.4 HOLC A-D Means, SEs and Differences, Full vs. Panel, Finance (USD 2018)	133
B.5 HOLC A-D Means, SEs and Differences, Full vs. Panel, Racial Diversity Outcomes	134
B.6 HOLC A-D Averages Over Time, Full and Panel, Exposure Index, 1988-2019	135
C.1 Summary of Cellphone Policies and Metal Detector Status in Responding Schools	141
C.2 Impact of NYCDOE Cellphone Ban on Suspensions and Select School Culture Outcomes, SY2006-2016	142
C.3 Impact of NYCDOE Cellphone Ban on School Safety (Students) by Individual Outcome, SY2006-2016	143

List of Figures

1.1 Birth and Outcomes Time Range by Cohorts	12
1.2 Pretreatment Index Kernel Density Estimation by Preschool Status Across Cohorts	17
1.3 Head Start Impacts on the Adulthood Index Individual Outcomes Across Cohorts	23
1.4 Head Start Longer-Run Impacts	24
1.5 Head Start Impacts on School-Age Outcomes Across Cohorts	26
2.1 1935-1940 HOLC CBSAs by US Census Bureau Region	49
2.2 HOLC A-D School Distributions by HOLC A-D District Grades	65
2.3 HOLC Averages, Finance Outcomes, 1989-2018 (USD 2018)	70
2.4 HOLC Pairwise Comparisons, Finance Outcomes, 1989-2018 (USD 2018)	70
2.5 HOLC A-D Averages Over Time, Racial Diversity Outcomes, 1988-2019	77
2.6 HOLC Pairwise Comparisons Over Time, Diversity Outcomes, 1988-2019	77
3.1 Per-Pupil Suspensions by Scanning Status, SY2012-2016	107
3.2 School Culture (Students) by Scanning Status, SY2012-2016	107
3.3 School Culture (Students) by Scanning Status, Individual Outcomes, SY2012-2016	109
3.4 School Culture (Teacher) by Scanning Status, SY2012-2016	110
B.1 Los Angeles County Area Descriptions for Nos. A-1 and D-1	136
B.2 Los Angeles and Surrounding Districts & 1935 – 1940 HOLC Maps	137
B.3 Cleveland and Surrounding School Districts & 1935 – 1940 HOLC Maps	137
B.4 School Total PPE by HOLC A-D District Grades – Only “D” Schools	138
B.5 HOLC A-D Coefficient Plots, Full vs. Panel, 2017-18, Finance (USD 2018)	138
B.6 HOLC A-D Coefficient Plots, Full vs. Panel, Diversity Outcomes	139
B.7 HOLC A-D Averages Over Time, Exposure Index (1988-2018)	139
C.1 Per-Pupil Suspensions by Scanning Status, SY2006-2016	144
C.2 School Culture (Students) by Scanning Status, SY2006-2016	144
C.3 School Culture (Students) by Scanning Status, Individual Outcomes, SY2006-2016	145
C.4 School Culture (Teacher) by Scanning Status, SY2006-2016	146

Acknowledgments

First, I thank my dissertation committee, David Deming, Eric Taylor, and Andrew Ho, for their steadfast support, feedback, and encouragement throughout my doctoral studies. I am also appreciative of the mentorship and professional support of many others, including Joshua Goodman, Martin West, Felipe Barrera-Osorio, Drew Bailey, Greg Duncan, Emmerich Davies, Peter Blair, Ashley Craig, and David Martin.

I am grateful for the amazing group of instructors and professors at Harvard Graduate School of Education that I had the honor of teaching for throughout my doctoral career, including Joseph McIntyre, Jon Fullerton, Andrew Ho, and James Kim. Your commitment to your students, teaching fellows, and pedagogy is an inspiration and model to follow for those teaching or desiring to teach in higher education.

My dissertation benefited from the financial support provided by Harvard University, Harvard Graduate School of Education, and Harvard Graduate School of Arts and Sciences. I would also like to thank the Bill & Melinda Gates Foundation for their generous financial support during the final year of my doctoral program.

Finally, I am immensely grateful to my family, friends, and classmates for their kindness and support. Christopher Cleveland, thank you for being such a wonderful colleague, co-author, and friend throughout my Ph.D. candidacy. Ellen, words cannot capture your immense contribution throughout my studies. You have been a champion of my research aspirations and love of learning since I have known you, and I am so grateful to have you by my side on this journey.

Chapter 1

Elusive Longer-Run Impacts of Head Start: Replications Within and Across Cohorts^{1 2}

1.1 Introduction

Understanding the causal effects of early childhood programs implemented at scale on long-term adult outcomes is challenging. However, as early childhood is considered by many economists to be a key launching period for lifelong human capital accumulation (e.g., Chetty et al., 2011; Cunha et al., 2006; Currie & Almond, 2011; Heckman & Mosso, 2014; Hoynes et al., 2016), considerable attention has been devoted to research attempting to estimate the short- and longer-run impacts of early education programs (Duncan & Magnuson, 2013).

One such early childhood education program is Head Start, the United States' oldest and largest early childhood education program to be offered at scale. Given this, it is hardly surprising a great deal of the research has been devoted to it.³ Studies of the longer-run impacts of Head Start attendance have generally shown positive, albeit sometimes mixed, results (Bauer & Schanzenbach, 2016; Carneiro & Ginja, 2014; Deming, 2009; Garces et al., 2002; Ludwig & Miller, 2007; Thompson, 2018). In a recent study analyzing data from the Panel Study of Income Dynamics (PSID) survey, Johnson and Jackson (2019) suggest that some of these inconsistencies can be attributed to complementarities between Head Start attendance and subsequent K–12 spending.⁴ One important study on the long-term impacts of Head Start attendance is Deming (2009). Using data from the National Longitudinal Survey of Youth 1979 Children and Young

¹ This chapter has been published at *Educational Evaluation and Policy Analysis* (<https://doi.org/10.3102/0162373720948884>). Full citation: Pages, R., Lukes, D., Bailey, D., & Duncan, G. (2020). Elusive Longer-Run Impacts of Head Start: Replications Within and Across Cohorts. *Educational Evaluation and Policy Analysis*, 42(4), 471–492.

² Dylan J. Lukes and Remy Pages are co-equal first authors. Please refer to Appendix D for the official co-authorship statement.

³ An overview of the literature on Head Start's short- and medium-term impacts is available in the Supplementary Section S1 in the online version of the journal.

⁴ For example, Johnson and Jackson (2019) calculated that for a child attending Head Start, a 10% increase in K–12 spending boosted educational attainment by 0.4 years, earnings by 20.6%, and reduced the probability of being incarcerated by 8 percentage points.

Adults (CNLSY), his paper builds on the approach of Currie and Thomas (1995; also, Garces et al., 2002) by comparing both school-age and young adulthood outcomes between children who attended Head Start and their siblings who either attended other non-Head Start preschools or did not attend any preschool program. Most of the cohorts analyzed in Deming (2009) were born between 1976 and 1986 and had outcomes tracked through the survey's 2004 interviewing wave.⁵ The study found that, compared with siblings who did not attend any preschool, children who attended Head Start averaged 8.5 percentage points higher rates of high school graduation and 0.23 standard deviation (SD) higher scores on an index of adult outcomes.⁶ Deming (2009) is noteworthy both for its sibling comparison design, which controls for some unmeasured time-invariant factors of the family environment, and because of its use of a reasonably large and relatively recent national longitudinal sample followed through childhood into early adulthood.

However, recent research calls into question the sibling comparison design both in terms of its external validity (D. L. Miller et al., 2019) and potential to produce biased estimates from sibling spillover effects (Heckman & Karapakula, 2019). To correct for nonrandom selection into the family fixed effects (FEEs) identifying sample, D. L. Miller et al. (2019) found that reweighting on observables attenuated many of the original Deming (2009) FFE estimates of Head Start's impact on long-term outcomes. D. L. Miller et al. (2019) document similar attenuated FFE estimates of Head Start's long-term impact in the PSID. Highlighting threats to construct validity, Heckman and Karapakula (2019) found siblings who participated in the Perry Preschool Project had large positive spillovers on their nonparticipating siblings. This was particularly true for male siblings.

⁵ Nearly 95% of children included in the Deming (2009) family fixed effect (FFE) estimation sample were born between 1976 and 1986. The remaining 5% of children were born between 1970 and 1975.

⁶ The adulthood summary index used in Deming (2009) included high school graduation, college attendance, teenage parenthood, idleness (i.e., neither working nor attending school), crime, and poor health status.

Importantly, evidence of Head Start’s lasting positive impact into adulthood is not limited to Deming (2009) or the FFE design. Over the past several decades, a sizable body of evidence that leverages a variety of empirical methods, including FFE, regression discontinuity (RD), and difference-in-difference (DID), have shown Head Start’s ability to improve adolescent and longer- term outcomes (Bailey et al., 2018; Barr & Gibbs, 2019; Bauer & Schanzenbach, 2016; Carneiro & Ginja, 2014; Currie & Thomas, 1995, 1999; Garces et al., 2002; Ludwig & Miller, 2007; D.L. Miller et al., 2019; Thompson, 2018). Most consistent across this body of research is Head Start’s positive impact on educational attainment, health outcomes, and reduced criminal activity with estimated impacts tending to be larger and more robust for males, siblings from earlier birth cohorts, and those born to mothers with less than a high school education.

On the whole, these studies predominantly focused on 1970s and 1980s birth cohorts. Notable exceptions include Carneiro and Ginja’s (2014) RD analysis of 1977–1996 birth cohorts and Bauer and Schanzenbach’s (2016) FFE analysis of 1970–1990 birth cohorts. Although not part of their main results, Barr and Gibbs’s (2019) supplementary FFE analysis sample (contained in their appendix) included CNLSY 1970–1992 birth cohorts. Results are mixed on Head Start’s impact for more recent birth cohorts. Carneiro and Ginja (2014) indicated that a robustness check showed that the positive effects for males age 12 to 13 in their overall sample were driven by the earlier 1980s birth cohorts. Similarly, in an appendix FFE analysis, Barr and Gibbs (2019) found no significant impact of Head Start on high school graduation, some college, crime, teen parenthood, or their index of adulthood outcomes. However, Head Start impacts were positive and significant for males on high school graduation, crime, and their index of adulthood outcomes (Barr & Gibbs, 2019). In both cases, each overall sample included birth cohorts from 1970s through the early to mid-1990s. In contrast, Bauer and Schanzenbach’s (2016) FFE

analysis found positive impacts of Head Start on high school graduation, some higher education, postsecondary completion, self-control index, self-esteem index, and positive parenting index. These results more closely followed Deming (2009) and included birth cohorts up to 1990.⁷ A detailed synthesis of these studies and more—including birth cohorts analyzed, identification strategies, and findings—can be found in the Supplementary Table A1 in the online version of the journal.

1.2 Present Study

The present work builds upon this rich existing literature by expanding Deming's (2009) evaluation of Head Start's longer-run impacts. By appending 10 additional years to the original 1976–1986 birth cohorts analyzed in Deming (2009), we were able to estimate impacts on outcomes measured later in adulthood and not previously considered: educational attainment, college graduation, and earnings. Second, the additional data provided us with an opportunity to apply the methods used in Deming (2009) to 10 additional birth cohorts in the National Longitudinal Survey of Youth (NLSY) to address whether his results generalized to cohorts born to older mothers and into somewhat different historical conditions. Third, we estimated impacts on both school-age and adulthood outcomes for a sample combining all possible cohorts to provide estimates based on the broadest population base.

We found that extending the measurement period for Deming's cohorts and early-adult outcomes decreased the estimated impact on the adulthood summary index (ASI) of Head Start attendance relative to not attending any pre- school program from 0.23 to 0.17 SD, standard error (SE) = 0.07. Of the longer-run outcomes we were able to consider, the largest impact of attending Head Start was on years of completed schooling (0.30 years; SE = 0.15). This is

⁷ The FFE analysis sample used in this article includes 1970–1996 National Longitudinal Survey of Youth 1979 Children and Young Adults (CNLSY) birth cohorts.

notable and, taken by itself, could indicate a sizable return on investment for the program. However, we estimated relatively small, nonsignificant impacts on gains on other later life outcomes, including college graduation and earnings.⁸ For the children born after Deming’s cohorts, Head Start impacts were mostly null and sometimes negative. In fact, positive impacts on ASI generated by Deming’s cohorts were matched by nearly symmetric negative impacts for the complement cohorts (-0.15 SD; $SE = 0.07$). For the final sample that combined the two sets of cohorts, the point estimate of Head Start’s impact on the ASI was close to zero and not statistically significant.

In light of recent work by D. L. Miller et al. (2019), following these initial analyses, we checked whether our FFE identifying samples exhibited “selection into identification” (SI) across a variety of observable characteristics, including family size and mother’s age at child’s birth. Finding evidence of SI in both the Deming and combined cohorts, we used the one-step reweighting-on-observables procedure outlined in D. L. Miller et al. (2019) to correct for any potential bias. Similar to D. L. Miller et al. (2019), after reweighting the Deming cohort, we found attenuated estimates of Head Start’s impact on long-term outcomes. However, for the combined cohorts sample, we found limited evidence that reweighting attenuated Head Start impact estimates on long-term outcomes.

Our article concludes with a discussion of what is driving these cross-cohort differences in Head Start’s impact. Although we found differences in baseline human capital between cohorts, we found no evidence that the impact of Head Start varied for different levels of human capital within cohorts. Similarly, we observed differences in other pretreatment covariates between cohorts, but also found limited evidence that they drove variation in Head Start impacts.

⁸ Earnings are computed as the log of 1994–2014 averaged earnings, adjusted for age and survey year.

Finally, to better understand whether the effects of Head Start were changing across cohorts, we performed a Blinder–Oaxaca decomposition (Blinder, 1973; Oaxaca, 1973) using a threefold decomposition (Jann, 2008). In line with the above analyses, the results from this exercise also highlighted key differences between the Deming and complement cohort samples, but importantly showed how these differences—most notably in the pretreatment index and mother’s age at child’s birth—were associated with variation in estimated Head Start impacts across cohorts. These results indicated that if mother’s age at child’s birth was set fixed at Deming’s cohort level, the ASI mean for the complement cohort’s Head Start attendees would be similar to that of Deming’s cohort counterparts.

These final analyses shed light on the potential importance non-Head Start factors play in moderating the direction and magnitude of Head Start impacts—most notably the pretreatment index and mother’s age at child’s birth. The pretreatment index is composed of a wide range of within-family covariates that vary between siblings and occur before or at the age of a child’s Head Start eligibility. These variables, many of which have been shown to improve with each marginal year a young mother delays childbirth beyond her teenage years (e.g., Augustine et al., 2015; Duncan et al., 2018; Hotz et al., 2005; A. R. Miller, 2011), include maternal work history, maternal income, maternal and infant health, child care arrangements, and household structures. Holding all else equal, changes in these underlying measures could alter a given parent’s need of and benefit from Head Start’s wraparound services which both educate and, in some cases, employ parents of participating children (Currie & Neidell, 2007; Zigler & Valentine, 1979). Beyond this, children from better resourced households participating in Head Start may also need and benefit less from participation, particularly if household resources such as parental human capital, parental income, or parental socioemotional skills—each of which typically improve

over a parent’s lifetime—substitute for the benefits that otherwise would have been provided by the program.

Although we are unable to conduct definitive tests of them, these hypotheses are consistent with extant literature documenting heterogeneity in the impacts of Head Start on short- and long-term outcomes, with impacts, concentrated at the left tail of the outcome distribution suggesting Head Start benefits participants most in need, particularly children of parents with low skill or low social backgrounds (Bitler et al., 2014; De Haan & Leuven, 2020). Although results did not show robust evidence of heterogeneity of Head Start estimates within cohorts across time, our analyses hint at the disproportionate positive impact Head Start has for those children from less advantaged households relative to their more advantaged participating peers. Still, additional factors could help explain cross-cohort differences in program effects, including shifting Head Start counterfactuals, changes in Head Start programming and quality of offerings, and an altered social context. Although our article provides no definitive evidence regarding these claims, they could be promising avenues of future exploration.

Thus, although the past several years have seen a resurgence of research on the long-run impacts of Head Start, this study adds value in several notable ways. First, we used a well-established FFE design, which hitherto has estimated positive long-run outcomes of Head Start, to estimate predominantly negative or null long-run outcomes of the program and showed how, if at all, these results were sensitive to “selection into identification” by performing the one-step reweighting-on-observables procedure as outlined in D. L. Miller et al. (2019). Second, despite major changes to the program and social context of Head Start–eligible children during this period, ours is the first paper to estimate the impact of Head Start for the most recent set of

CNLSY birth cohorts and to compare their program effects with those of earlier cohorts.⁹

Finally, we attempted to reconcile why Head Start impacts were different across cohorts, finding suggestive evidence that between-cohort differences in the ages of mothers at the time of their child's birth played an important explanatory role.

1.3 Method

1.3.1 Head Start Program Background

Part of the Johnson Administration's Great Society policies, Head Start was launched in 1965 to provide educational and health-related services to children living in low-income families. As of 2017, about 900,000 children were enrolled in Head Start, 97% of whom were between the ages of 3 and 5, at an annual cost of around uS\$9 billion in federal funding (U.S. Department of Health and Human Services, Administration for Children and Families, 2018). Enrollment and funding have varied greatly since Head Start's 1965 inception. Participation grew until the early 1980s, plateaued through the early 1990s, and then grew again when funded enrollments almost doubled (i.e., from around 500,000 in 1990 to 900,000 in 2000). Appropriations (in 2018 US\$) grew from about uS\$3 billion in 1990 to US\$9 billion in 2000. After 2000, both enrollment and inflation-adjusted funding remained steady (U.S. Department of Health and Human Services, Administration for Children and Families, 2018).

Between the 1989–1990 (which are typical Head Start attendance years for Deming's cohorts) and 1996–1997 (which are typical attendance years for our complement cohorts), enrollment increased by about 60%. However, the proportion of teachers or assistant teachers with at least a Child Development Associate credential increased very little—by about 5

⁹ The Head Start impact estimates of Carneiro and Ginja (2014), Bauer and Schanzenbach (2016), and Barr and Gibbs (2019) were based on samples that included 1990 birth cohorts. However, these studies did not systematically estimate Head Start impacts by birth cohorts across time. Instead, primary results were for analyses based on overall samples which include both more recent birth cohorts from the late 1980s and early 1990s, and older birth cohorts from 1970s and early to mid-1980s.

percentage points over this period (U.S. Department of Health and Human Services, Administration for Children and Families, 2018). More generally, the 1990s and 2000s were a time of rapid increases in preschool enrollment, including Head Start, but also state-run prekindergarten programs (Duncan & Magnuson, 2013).

1.3.2 Data

Figure 1.1 provides an overview of birth years and years in which childhood and adult outcomes are measured for the two sets of cohorts that form our analytic samples. Deming's cohorts were born between 1970 and 1986 and attended Head Start no later than 1990. Moreover, Deming's sibling fixed effect (FE) analyses were estimated for a sample of siblings discordant on Head Start attendance and who enrolled in Head Start no later than 1990. Deming's sample eligibility rules were (1) at least two children aged 4 or older by 1990 within the same family, and (2) at least one pair of siblings in a family had to be discordant across Head Start, other preschool, or neither statuses. The median age of individuals in Deming's analytic sample was 23 years (21 and 25 years for first and third quartile, respectively) by 2004, the most recent CNLSY survey round year available for his study.

For our complement and combined cohorts, Deming's sample restrictions were moved forward by 10 years: samples were restricted to siblings who were at least 4 by 2000 (i.e., at least 19 by 2014). Sample restrictions produced sample sizes of $N = 1,251$ for Deming's cohorts, $N = 2,144$ for our complement cohorts, and $N = 3,768$ for our combined cohorts.¹⁰ It is important to

¹⁰ A third restriction was applied to the complement cohort: Siblings were considered for eligibility up to 2000, excluding those already part of Deming's cohort (i.e., selected under Rules 1 and 2); it is in that sense that this new cohort is the complement of Deming's cohort. Of all siblings comprising the complement cohort, 78% had reached 4 years of age post-1990 (75%, for Head Start participants). As in Deming (2009) and for all cohorts, the original National Longitudinal Survey of Youth 1979 (NLSY79) over sample of low-income Whites was excluded. One final point, the sum of Deming's cohort and the complement cohort is smaller than the size of the combined cohorts sample as there are more opportunities for siblings to meet the FFE eligibility criteria for the latter. For example, there were cases where families included in Deming's cohort had an additional child or children in later years that were not age eligible for Deming's cohort but were old enough to be a candidate for the complement cohort. However, to be included in the complement cohort, these age-eligible children needed to exhibit differential participation in Head Start. Thus, the family with one child was automatically disqualified from the complement cohort, and the family with multiple new children would also be excluded unless these children exhibited differential participation in Head Start. Regardless of their inclusion within the complement cohort, the children in each of these scenarios would be included in the combined cohorts sample.

note that the sampling design of the CNLSY (i.e., all children were born to women who were between ages 14 and 22 in 1979) led children in our complement cohort to be born to older mothers than is the case for children in Deming’s cohort. Later, we consider the role this factor may have played in explaining differences in Head’s Start impact between the Deming and complement cohorts. Furthermore, because we wanted to both estimate Head Start’s impacts on educational attainment, college graduation and earnings and assess their robustness on ASI, we both replicated and extended Deming’s analysis for these cohorts up to 2014, the latest CNLSY survey round year available to us at the time of our analyses.

1.3.3 Family Background Statistics

In Table 1.1, household characteristics are presented by cohort and preschool status (Head Start vs. the counterfactual of no preschool), permanent income, maternal education and cognitive test score, and grandmother’s highest grade completed.¹¹ Across these variables and for all three cohorts, there was a clear pattern of selection of more disadvantaged children into Head Start for samples of siblings under Rule 1 only—a less restricted sample, more representative of the CNLSY sample— and samples with Rule 2 added (i.e., the FE subsamples). Discrepancies between the two samples were small, suggesting that the demographic characteristics of the FE subsamples were similar to the less restricted, larger samples.¹²

As shown in the column “Difference HS- None”—reporting mean differences in standard deviation units for Deming’s cohorts, the complement cohorts and the combined cohorts,

¹¹ NLSY79 derived from the Armed Services Vocational Aptitude Battery of tests, the Armed Forces Qualification Test (AFQT) comprising items in arithmetic reasoning, mathematics knowledge, word knowledge, and paragraph comprehension.

¹² Deming (2009) presented these characteristics over three preschool statuses (i.e., Head Start; other pre-school; no preschool) and by racial/ethnic subgroups. To keep Table 1.1 manageable, only overall means for Head Start and no preschool status (the counterfactual) are displayed. For details with other preschool status included (and also by racial/ethnic subgroups), see Supplementary Tables S2 through S4 in the online version of the journal. There was a reduction in sample sizes when restricting on families that had differential preschool status for at least two siblings (see “Data” subsection, Rule 2): by 66% for Deming’s cohort, by 41% for the complement cohort, and by 45% for the combined cohorts. However, variation on selected household characteristics appeared to be very similar across both type of samples, and across all cohorts (Table 1.1).

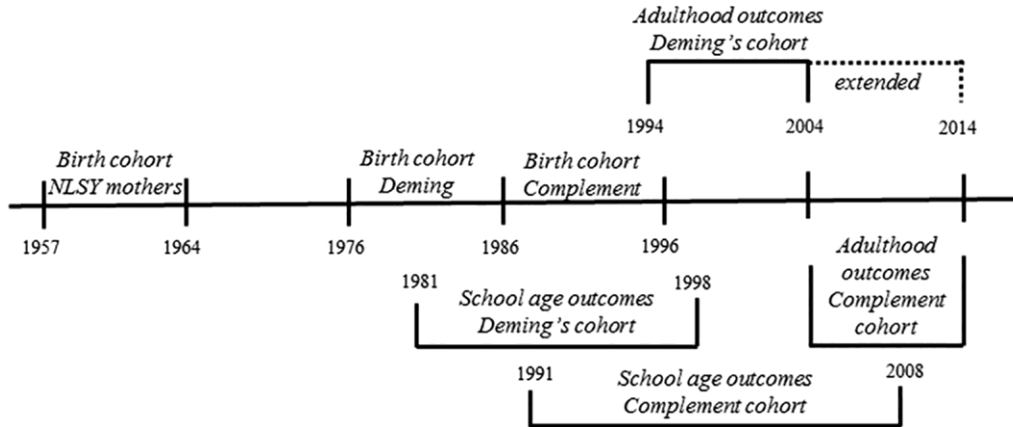
respectively—selection into Head Start was similarly associated with socioeconomic disadvantage for Deming’s cohort as well as for the complement cohort. For example, Head Start participants had a 0.44 SD lower permanent income and a 0.59 SD lower maternal Armed Forced Qualification Test (AFQT) than children not attending any preschool. Overall, Head Start children came from relatively more disadvantaged households.¹³ As noted by Deming (2009), because his cohort of Head Start participants had been born to younger mothers (their median age was 20), they might have benefited more from the program (which, in addition to early education, includes services for parents). In contrast, for the complement cohort, mothers were older (median age was 28), and household characteristics more favorable on all of the dimensions included in Table 1.1.

1.3.4 Outcomes

As part of our replication and extension of Deming (2009), we assessed the impact of Head Start on the same set of three test scores, two nontest outcomes, and six young adulthood outcomes for each of the Deming, complement, and combined cohorts. The three test scores covered ages 5 to 14 and included the Peabody Picture Vocabulary Test (PPVT), the Peabody Individual Achievement Math (PIATMT) subtest, and the PIAT Reading Recognition (PIATRR) subtest. Following Deming (2009), due to the biannual survey design of the CNLSY, we pooled PPVT tests scores of 5- and 6-year-olds to get the first post-Head Start score for each child in our sample. Both the PIATMT and PIATRR were administered annually for respondents ages 5 to 14 and resulted in considerably more observations compared with the PPVT.

¹³ Differences were even more pronounced when comparing between Head Start and the other preschool status (see Supplementary Table S2 in the online version of the journal).

Figure 1.1: Birth and Outcomes Time Range by Cohorts



Notes: Timewise, the combined cohorts sample (not shown) encompasses Deming's and complement cohorts. While boundary end points are approximate, they nonetheless include the bulk of each distribution: around 95% and 85% for Deming's cohort and complement cohort, respectively.

The two nontest outcomes covered ages 7 to 14 and included grade retention and learning disability diagnosis. As in Deming (2009), grade retention is a dichotomous variable based on survey respondents' answers to whether their child has ever been retained at grade level while in school. This question was asked biannually in the NLSY from 1988 to 2014. Grade retention was coded as a 1 if parents ever answered "yes" to this question across any of the survey years. Learning disability was based off a "yes" or "no" NLSY survey question that asked parents if their child had a learning disability. We coded our learning disability variable as 1 if respondents ever answered "yes" to this question, discounting a small number of children who were diagnosed with a learning disability prior to age 5.

Finally, as in Deming (2009), we included the same six young adulthood outcome variables: high school graduation, teen parenthood, some college attended, idleness, involvement with the justice system, and poor health status. All outcomes were measured up to the CNLSY 2014 survey round. Individuals were considered "idle" if they were not enrolled in school or had reported zero annual earnings—by 2004 for the Deming cohort and by 2014 for the complement

cohort. The “involvement with the justice system” variable was constructed as a dichotomous variable, coded as one if a respondent ever answered “yes” to any survey question related to conviction, probation, sentencing, and prison. Teen parenthood was operationalized also with an indicator equal to one if a respondent’s age at the birth of their first child was before 20 years old and applied to both female and male respondents. Finally, our poor health status variable was constructed by averaging a respondent’s self-reported health framed by a 1-to-5 Likert-type scale (lower responses equating to poorer self-reported health). Poor health status was flagged by a dichotomous variable coded as one, if a respondent’s average self-reported health score was less than three on the Likert-type scale.

Just as in Deming (2009), to reduce the risk of multiple-inference inflated Type I errors and mitigate measurement error, we constructed summary indices for the three test scores, an index for the two nontest score outcomes, and a final adulthood summary index (ASI) for the six young adulthood outcomes. Outcomes were normalized to have mean of zero and standard deviation of one, with positive index values signaling “good” outcomes and negative index values signaling “bad” outcomes. The final index was then created by taking a simple average of all the normalized and, where appropriate, resigned outcomes.

Comparing the distribution of these outcomes across the Deming and complement cohorts, we found evidence of substantial distributional shifts often favoring the complement cohort (see Supplementary Table S5 & Figure S6 in the online version of the journal). These between-cohort distributional changes might explain why Head Start impacts were dramatically different for the more recent set of siblings in the complement cohort. To illustrate, on the ASI, complement cohort siblings attending Head Start and not attending preschool were respectively 0.22 SD and 0.50 SD higher than the Deming cohort siblings. Similarly, although there was a

Table 1.1: Household Characteristics Averaged Over Head Start and No Preschool Status

Household characteristic	Head Start		No preschool		Diff. HS-none	Diff. HS-all
	Deming's cohort	Complement cohort	Deming's cohort	Complement cohort		
Permanent income						
Full sample	32,884 (21,810)	39,800 (27,539)	35,970 (24,830)	42,764 (30,000)	61,857 (49,704)	52,445 (41,989)
Fixed effects subsample	34,672 (25,443)	40,465 (29,118)	37,571 (26,961)	41,587 (27,968)	61,793 (53,019)	53,938 (45,831)
Mother < high school						
Full sample	0.24 (0.43)	0.12 (0.33)	0.19 (0.39)	0.24 (0.43)	0.12 (0.32)	0.18 (0.38)
Fixed effects subsample	0.28 (0.45)	0.14 (0.35)	0.20 (0.40)	0.22 (0.42)	0.12 (0.33)	0.16 (0.37)
Mother some college						
Full sample	0.28 (0.45)	0.40 (0.49)	0.33 (0.47)	0.25 (0.43)	0.44 (0.50)	0.34 (0.48)
Fixed effects subsample	0.25 (0.43)	0.42 (0.49)	0.34 (0.47)	0.27 (0.44)	0.43 (0.50)	0.37 (0.48)
Maternal AFQT						
Full sample	-0.61 (0.61)	-0.50 (0.71)	-0.56 (0.66)	-0.36 (0.80)	0.11 (1.03)	-0.12 (0.95)
Fixed effects subsample	-0.62 (0.61)	-0.48 (0.70)	-0.54 (0.66)	-0.21 (0.81)	-0.01 (0.99)	-0.18 (0.92)
Grandmother's education						
Full sample	9.16 (3.09)	9.69 (3.08)	9.39 (3.09)	9.45 (3.23)	10.41 (3.39)	9.92 (3.35)
Fixed effects subsample	9.13 (3.12)	9.79 (3.00)	9.50 (3.04)	9.69 (3.14)	10.29 (3.30)	10.07 (3.22)
Sample size	779	637	1,491	1,931	1,857	3,658
Sample size FE	435	475	972	769	1,098	1,799

Note. Means and standard deviations were obtained for the full and sibling fixed effects samples, across cohorts. Differences in means (in SD units) between *Head Start* vs. *No preschool* status (Difference HS-None) and between *Head Start* vs. *No preschool* + *Other preschool* status (Difference HS-All) were reported for Deming's cohort/Complement cohort/Combined cohort (in that order). Permanent income is the average over reported years of household net income (in 2014 dollars). The AFQT was age normed based on the NLSY79 empirical age distribution of scores, and then standardized ($M = 0$; $SD = 1$). Household characteristics for the *Other preschool* status group were documented in the Supplementary Table S2 in the online version of the journal. Mean differences between Deming's cohort and Complement cohort were all significant at the 1% level (with the exception of Grandmother's education; see Supplementary Table S2 in the online version of the journal). HS = Head Start; FE = fixed effect; AFQT = Armed Forces Qualification Test.

small and marginally significant 0.10 SD difference on the cognitive test index across cohorts for Head Start attendees, that difference was more pronounced for the No Preschool status children (0.30 SD; $p < .001$). As these distributional shifts in outcomes were between cohorts across time and not within a cohort across Head Start treatment statuses, these results are not inconsistent with our later findings that the Deming and complement cohorts showed no signs of within-family selection into Head Start or No Preschool status (see Supplementary Table S7 in the online version of the journal).

Finally, the longer time series of NLSY data enabled us (but not Deming) to estimate Head Start effects for Deming's cohort on completed years of schooling and college graduation, and on earnings.¹⁴ The earnings composite for each sample member was obtained by first pooling all person-year earnings observations (in 2014 US\$) and then regressing them on dummy-variable indicators for birth cohort and calendar year to purge earnings of birth cohort and measurement year effects.¹⁵ From the coefficients in this regression, we generated a set of person-year earnings residuals for all individuals in the analysis sample. We then averaged these earnings residuals for each individual, added them to the grand mean earnings in the sample, and took the natural logarithm of this earnings average.

1.3.5 Empirical Strategy

As noted, families selecting into Head Start were relatively more disadvantaged on a series of selected household characteristics. Consequently, Head Start estimates relative to other preschool status based on cross-family variation may be negatively biased. An FFE design

¹⁴ Both variables were derived from CNLSY cross-round item asking respondents which highest grade they had completed at the date of the latest survey round interview. Responses were recoded as equivalent years of completed schooling (e.g., if respondent answered "high school graduate," it was recoded 12; "completed an associated degree" was recoded 14; etc.).

¹⁵ We opted to combine all respondent age 35+ into a single birth cohort dummy to ensure a sufficient sample size ($N = 589$) before regressing for adjustment. For all other birth cohorts, sample sizes were of at least 300 observations. The arguably arbitrary 35-year threshold was chosen such that a priori valid inputs would be available.

mitigates some of these biases by separating the potentially confounding influence of family environment variance shared among siblings from estimations of interest. This was the empirical strategy undertaken in Deming (2009), which we reproduced in the present study and formalized in the same fashion:

$$Y_{ij} = \alpha + \beta_1 HS_{ij} + \beta_2 PRE_{ij} + \delta X_{ij} + \gamma_j + \epsilon_i$$

In this model, i and j respectively, index individuals and families. Thus, HS_{ij} (PRE_{ij}) stands as an indicator for participation in Head Start (Preschool) where β_1 (β_2) denotes Head Start (Preschool) impact estimates on outcome Y_{ij} , for some sibling i within family j , relative to a sibling (within family j) attending neither. Next, $\mathbf{X}_{ij}$ represents the vector of “pretreatment” family covariates pertaining to sibling i within family j ; family j fixed effect is captured by γ_j , while ϵ_i represents sibling i ’s residual.

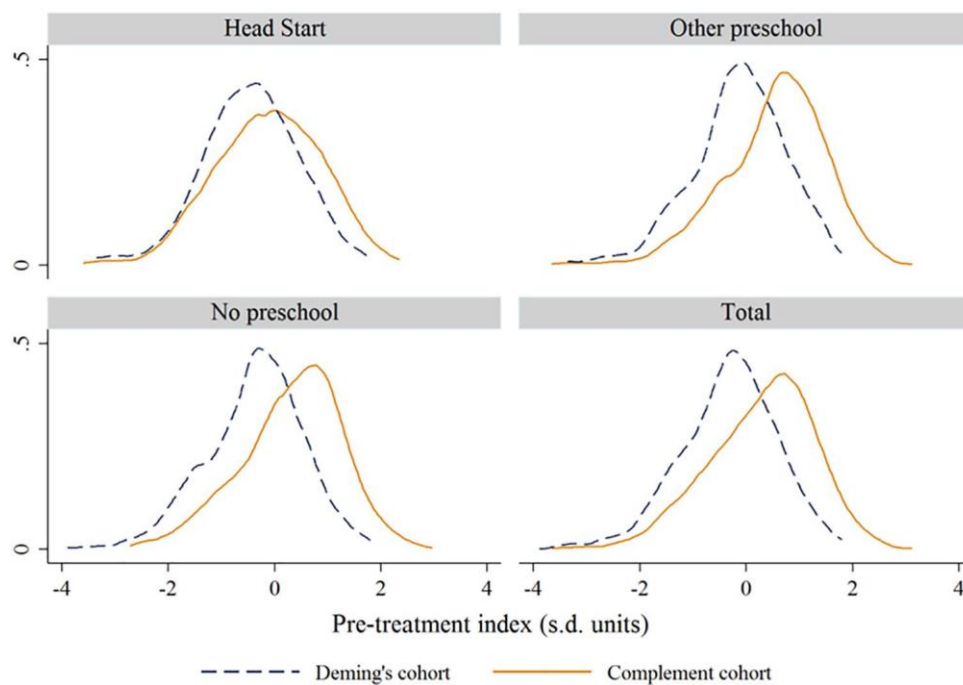
1.3.5.1 Selection Bias

Within-family comparisons remove the effects of time-invariant family characteristics on siblings’ outcomes. There remains, however, a strong possibility of within-family selection bias. Reasons for different within-family patterns of care in early childhood were not recorded in the CNLSY. To mitigate such potential for bias, Deming (2009) opted for a series of sibling-specific family-level covariates—the ones represented by the vector $\mathbf{X}$ in Equation 1—measured before or at the age of Head Start eligibility (3 years old). We examined these covariates for the complement and combined cohorts and tested whether siblings within a given family systematically differed on these covariates.¹⁶

¹⁶ As in Deming (2009), when estimating Head Start impacts in the regression models, missing data for these covariates were imputed with their corresponding sample mean value. For each, a dichotomous indicator for imputed responses was also included

Within our FFE framework, each covariate was thus regressed on siblings' preschool status, either Head Start or other preschool. A statistically significant and substantial variation from no preschool (the reference status) would then signal a potential selection bias regarding the relation between that pretreatment characteristic and the regressed-on preschool status. These estimates are reported in the Supplementary Table S7 in the online version of the journal.

Figure 1.2: Pretreatment Index Kernel Density Estimation by Preschool Status Across Cohorts



Notes: Distributions were smoothed, and densities estimated via the Epanechnikov kernel function. Kolmogorov–Smirnov tests all indicated nonequality, at the 0.1% level, between compared densities.

Focusing first on the complement cohort, siblings attending Head Start were on average older by 1 year, and by almost 2 years for siblings attending other preschools, than their counterparts not attending any preschool program. This was consistent with the probability being greater for first-born sibling to be enrolled in preschool, by 10 *pps* for Head Start enrollees, and by 28 *pps* for other preschool participants. Both groups were 6 to 8 *pps* less likely to receive

maternal care from birth to age 3, and so more likely to receive care from a nonrelative (5 and 6 *pps*, respectively). Attrition was low for both the complement and the combined cohorts, averaging about 4% and 3%, respectively. Moreover, both Head Start and other preschool participants were somewhat less likely (by about 3 and 2 *pps*, respectively) to be part of observations lost to attrition.

To characterize selection bias as it pertains to overall disadvantage, a summary index of all pretreatment covariates was constructed in the same way as the multiple outcomes adulthood index described earlier.¹⁷ For all cohorts, as with the covariates, the pretreatment index was regressed on the two preschool indicators, keeping the no preschool status as the reference category. We found that Head Start and other preschool within-family effects on the pretreatment index were close to zero and never statistically significant. As with Deming's (2009) selection bias analysis (which we replicated, see Supplementary Tables S8–S9 in the online version of the journal), we could not reject the null hypothesis of equality between preschool statuses.

While there was no evidence of parental selection into Head Start (or any other preschool status) within each cohort and household, Figure 1.2 illustrates what might be one explanation for cross-cohort differences of Head Start impacts on longer-run adulthood outcomes. Complement cohort respective kernel densities of pretreatment index scores were shifted to the right, that is, toward more favorable household characteristics for the complement cohort. Complement cohort siblings having attended Head Start later would have then, on average, benefited from more household resources. Compared with Deming's cohort, such a shift might

¹⁷ As in Deming (2009), variables comprising the pretreatment covariates index were all first positively oriented with respect to the adulthood summary index. For example, variables like gender (male), age (older), or grandmother living in household between child's birth and age 3 were negatively correlated with the outcome. Their correlational direction was reversed multiplying their sign by -1 . All covariates were then standardized and aggregated into an index, which in turn was also standardized ($M = 0$; $SD = 1$).

stand as a substitute for whatever impact the program would have otherwise yielded.

Table 1.2: Head Start Impacts on Cohorts' Adulthood Summary Index

Source of estimate	Model	Head Start	Other preschool	P value (HS = other)	R ²
Deming (2009) ^a	(1)	0.14 (0.07)	0.08 (0.08)	.47	.12
Measurement period ^b : 1994–2004	(2)	0.27 (0.08)	0.11 (0.08)	.12	.59
Sample size = 1,251 [364/364]	(3)	0.23 (0.07)	0.07 (0.07)	.08	.62
Deming's cohort ^c	(1)	0.14 (0.06)	0.08 (0.06)	.40	.21
Measurement period: 1994–2014	(2)	0.18 (0.07)	0.02 (0.09)	.07	.64
Sample size = 1,251 [364/364]	(3)	0.17 (0.07)	0.03 (0.07)	.13	.69
Complement cohort ^d	(1)	– 0.12 (0.06)	0.03 (0.06)	.01	.24
Measurement period: 2004–2014	(2)	–0.16 (0.10)	–0.05 (0.08)	.30	.71
Sample size = 2,144 [497/795]	(3)	– 0.15 (0.07)	–0.04 (0.05)	.15	.73
P value for model (3) ^e (Deming's = complement)		.01	.64		
Combined cohorts ^f	(1)	–0.02 (0.04)	0.06 (0.03)	.05	.24
Measurement period: 1994–2014	(2)	–0.01 (0.01)	0.01 (0.04)	.78	.61
Sample size = 3,738 [951/1,275]	(3)	–0.01 (0.04)	–0.003 (0.04)	.86	.63

Notes: Adulthood summary index (standardized) is a composite of six indicators: high school graduation, college attendance, teenage parenthood, either working or attending school, involvement with the justice system, and poor health status. Model (1): adulthood index is regressed on Head Start and other preschool participation indicators, along with pretreatment covariates and standardized permanent income; maternal AFQT score; one indicator for maternal high school graduation and one for some college attendance; siblings' gender and age. Model (2): same as model (1) but with family fixed effect only, no pretreatment covariates. Model (3): same as model (2) with pretreatment covariates included. Standard errors are in parenthesis and clustered at the family level. Estimates in bold case were significant at the 5% level or less. HS = Head Start; AFQT = Armed Forces Qualification Test. ^aDeming published results. ^bOutcomes measurement period. ^cFor Deming's cohort (compared with Deming, 2009), individual outcomes composing the adulthood index were extended up to 2014. ^dComplement cohort includes siblings fitting the same criteria as in Deming (2009) but found eligible from 1990 to 2000. ^ep value = estimates' difference testing between Deming's and complement cohorts' impacts estimated in model (3). ^fCombined cohorts integrate both Deming's and the complement cohorts.

1.4 Results

Each of the following subsections is organized by cohorts. We first present Head Start impacts on the adulthood summary index (ASI), along with its individual composing outcomes. Second, longer-run Head Start impacts on ASI, educational attainment, college graduation, and earnings are described. Third, estimates for school-age outcomes are shown. Robustness checks and a reconciliation of results are presented in the final subsection.

1.4.1 Head Start Impacts on ASI

In Table 1.2, the FFE model was implemented in steps—with three model specifications; repeated across the complement and the combined cohorts—to gauge the relative directions of

biases from observed covariates and unobserved household-level confounders.¹⁸ Model (1) included no FFEs but included the pretreatment covariates (see Supplementary Table S7 in the online version of the journal), along with household predictors (Table 1.1)—namely, standardized permanent income, maternal AFQT score, one indicator for maternal high school graduation, and one for some college attendance. By contrast, model (2) includes only FFEs. Model (3) includes both FFEs and pretreatment covariates. Moving from model (1) to model (2), the explained variance (R^2) was larger for all cohorts. Hence, error variance from unobserved variables was smaller than that from the selected observed variables. Moreover, moving from model (2) to model (3) added some precision to the estimates, with the R^2 increasing from .64 to .69, from .71 to .73, and from .61 to .63 for the Deming, complement, and combined cohorts, respectively.¹⁹ From the middle panel, for the complement cohort, Head Start impacts on ASI were negative at -0.15 SD (significant at the 5% level). This value was in clear contrast ($p < .01$) with Deming’s cohort estimate of 0.17 SD ($SE = 0.07$). In the bottom panel, for the combined cohorts model (3), no Head Start impact estimates were statistically significant; most were negative and close to zero.²⁰

We investigated the decrease in Head Start impact on ASI from 0.23 SD—in Deming (2009); outcomes measured up to CNLSY 2004 survey round—to 0.17 SD; outcomes measured up to 2014. This change was due in part to the impact estimate on “Idle” changing sign and

¹⁸ Deming (2009) used a similar approach in estimating impacts on the school-age cognitive tests index.

¹⁹ From similar trends, Deming (2009) concluded—based on Altonji et al.’s (2005) seminal work on the topic that estimates obtained from model (3) stood as lower bounds for Head Start causal impacts.

²⁰ Throughout (see Supplementary Tables S10–S13 in the online version of the journal), we considered identical demographic subgroups as in Deming (2009). Looking at the “Adulthood index” column in Supplementary Table S10 in the online version of the journal, for the combined cohorts sample, most impacts of Head Start were also close to zero. One exception being for siblings with low maternal AFQT subgroup—1 SD below NLSY79 AFQT empirical sample average (see bottom panel of Supplementary Table S10 in the online version of the journal). For these siblings ($N = 810$), Head Start appeared to have had a marginally significant positive impact on the adulthood summary index (0.11 SD; $SE = 0.08$). This estimate was greater for Deming’s cohort (0.38 SD, significant at the 1% level). For the complement cohort, impact was much smaller (0.03 SD, $SE = 0.15$), although difference testing between this cohort’s estimate and that of Deming’s cohort did not fall below the 10% level of statistical significance ($p = .15$). For the complement cohort, the proportion of siblings with low maternal AFQT background was also smaller (0.21 compared with 0.41 within Deming’s cohort) which is consistent with the overall observed favorable shift over household characteristics between the two cohorts (see Table 1.1).

ceasing to be statistically significant. For example, while Head Start participants in Deming's cohort were 7 *pps* (SE = 0.04) less likely to be "idle" in 2004, by 2014, the estimated impact disappeared (−3 *pps*; SE = 0.04). Overall, with the passage of the additional decade, Head Start participants were not, on average, better positioned to pursue a college degree or to have a job, relative to their siblings not having attended any preschool program.

Figure 1.3 shows Head Start impacts on all individual outcomes composing the ASI.²¹ For Deming's cohort, "Poor health status" stayed favorable by 5 *pps* (SE = 0.03), decreasing slightly from 2004 (7 *pps*; SE = 0.03). Impacts on "Some college attended" rose to statistical significance (11 *pps*; SE = 0.04).²² On "Crime," Head Start participants did not appear to have had more involvement with the justice system than their siblings. Yet, impacts on teenage parenthood shifted unfavorably. As impacts on ASI are based on scores averaged across the composing indicators, Deming's cohort overall decline on this index were captured by changes on the individual outcomes just described.

In contrast to Deming's cohort, Head Start impacts on the complement cohort were mostly negative and larger in absolute value; when positive, they were smaller in absolute value. Head Start's estimated impact on "Idle" was relatively large, negative (−0.08; SE = 0.03), and significant at the 1% level. Thus, in the complement cohort, siblings who have attended Head Start were less likely by about 8 *pps* to be employed or enrolled in school (by age 19 or older), compared with their siblings who received home care. Impact on "Some college attended" went in the opposite (negative) direction for the complement cohort (−0.07 SD; SE = 0.04), as well as for "Crime" (reversed scaled; −0.03; SE = 0.03).²³ In sum, the discrepancies between the two

²¹ For all estimates on individual outcomes, over- all and by subgroups, see Supplementary Tables S11 and S12 in the online version of the journal.

²² Education data were obtained using CNLSY 2014 survey-round cross-round variable for respondents' highest grade completed.

²³ For both Deming's and complement cohorts, Head Start impacts on "Poor health status" were positive (i.e., not self-identifying as being of poor health). This is in line with results found on a range of health outcomes in Carneiro and Ginja's (2014) Head Start evaluation study. The

cohorts over Head Start impacts on these individual outcomes are aligned with the difference observed earlier over ASI (Table 1.2). Once more, impact estimates for the combined cohorts sample were small and never statistically significant.

Finally, consistent with M. L. Anderson's (2008) study of early childhood interventions life cycle impacts, females appeared to have benefited more than males from Head Start across the board of outcomes considered here (see Supplementary Tables S10–S13 in the online version of the journal). Over the extended ASI (Deming's cohort), Head Start impact was estimated at 0.23 SD (SE = 0.11) for females versus 0.10 SD (SE = 0.10) for males.²⁴ Furthermore, females possibly carried more of Head Start's impact on educational attainment with an estimate at 0.34 SD (SE = 0.21) against 0.27 SD (SE = 0.21) for males (see Supplementary Table S13 in the online version of the journal).

1.4.2 Head Start Impacts on Longer-Run Outcomes

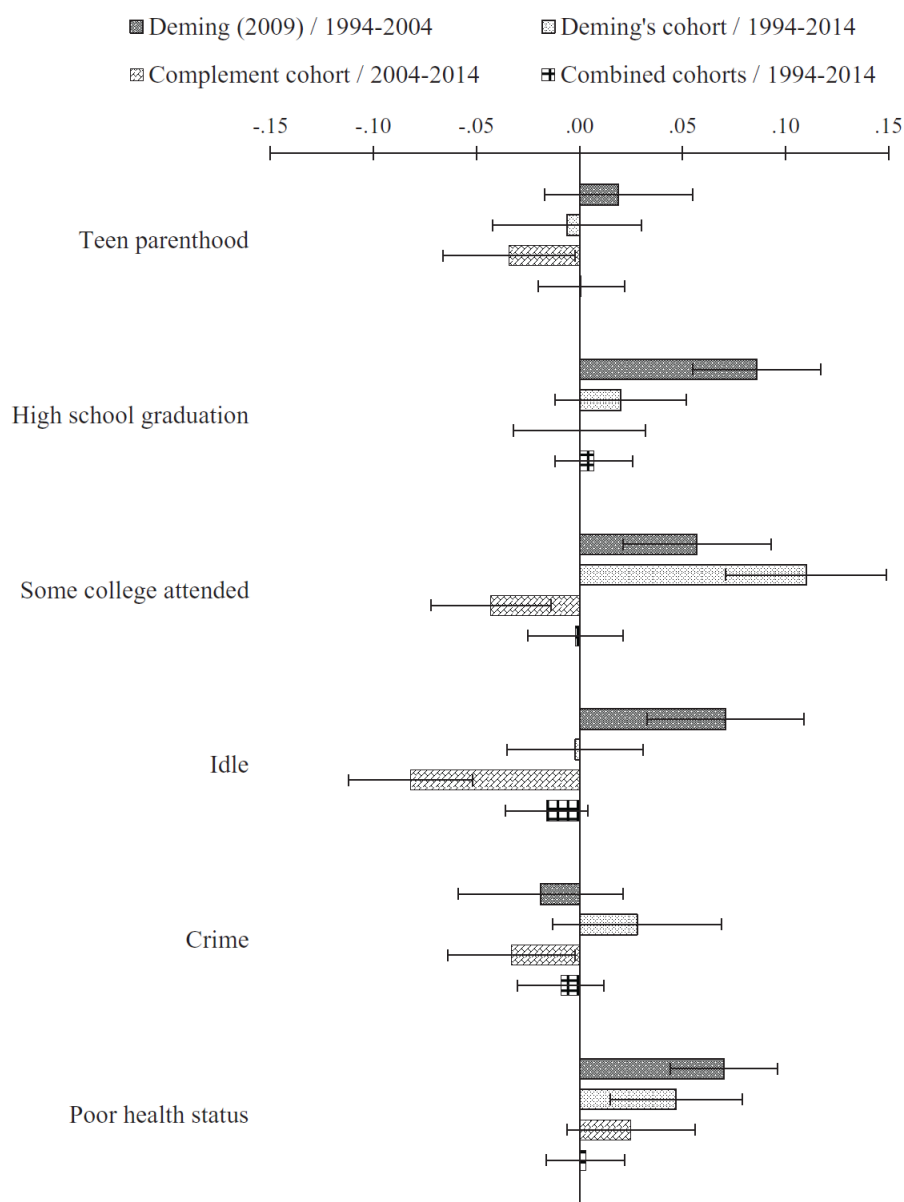
Head Start longer-run impacts are displayed in Figure 1.4 (the complete set of estimates is presented in Supplementary Table S13 in the online version of the journal). As described previously, impacts on ASI declined from Deming's (2009) published results as his study's cohort grew older by a decade (second bar from the top in Figure 1.4). Yet, by 2014, Head Start attendees went to school 0.3 years longer than their siblings not attending any preschool. This positive, potentially important impact, however, did not translate into either higher college graduation rate or to significantly higher adulthood earnings.²⁵

estimate for the combined cohorts was very small though, with a standard error well balanced across zero (Figure 1.3; Supplementary Table S12 in the online version of the journal).

²⁴ Although we could not reject equality between these estimates, we detected a favorable and statistically significant Head Start impact difference (see Supplementary Table S12 in the online version of the journal) of about 8 *pps* between genders for the "Idle" individual outcome (i.e., neither working nor in school) for the combined cohorts sample. Similarly, for the complement cohort, females had a 14-*pp* advantage on the "Crime" outcome (i.e., whether involved with the justice system).

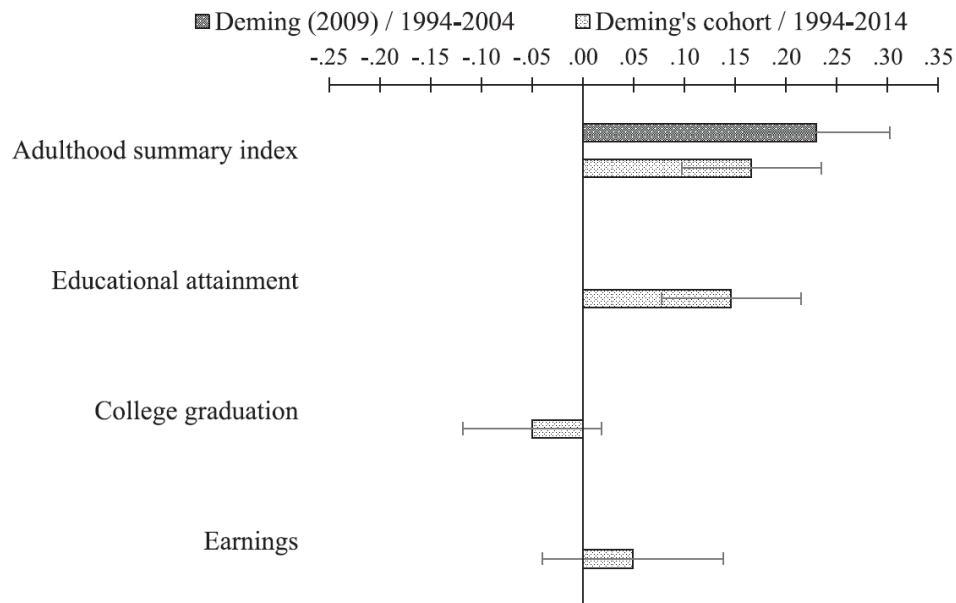
²⁵ Head Start impacts were positive and statistically significant for the subgroup of siblings whose maternal IQ (intelligence quotient) background was 1 SD below the M: adulthood index (0.38 SD; SE = 0.12); educational attainment (0.45 years of schooling completed; SE = 0.23); and earnings (0.44 log points; SE = 0.20). See Supplementary Table S13 in the online version of the journal.

Figure 1.3: Head Start Impacts on the Adulthood Index Individual Outcomes Across Cohorts



Notes: Measurement periods are displayed to the right of each label. Impacts are expressed as proportions. Deming's cohort, N = 1,251; complement cohort, N = 2,144; combined cohorts, N = 3,768. The counterfactual was a no preschool attendance. Error bars represent standard errors which were clustered at the family level. Estimates were oriented such that a positive value represents a more favorable outcome.

Figure 1.4: Head Start Longer-Run Impacts



Notes: Measurement periods are displayed to the right of each label. Impacts are expressed in standard deviation units. For both cohorts, $N = 1,251$. Recall that the Deming (2009) study did not estimate Head Start impacts on educational attainment, college graduation, and earnings, hence no Deming (2009) bar-estimates for these outcomes. The counterfactual was a no preschool attendance. Error bars represent standard errors which were clustered at the family level.

1.4.3 Head Start Impacts on School-Age Outcomes

Could Head Start impacts on school-age outcomes explain the cohort differences in adult outcomes shown above? For example, are Head Start impacts on achievement generally positive for Deming's cohorts but negative for the complement cohorts? Although a full econometric mediation analysis (e.g., Heckman & Pinto, 2015) was not the focus of this article, school-age outcomes might nonetheless be considered as potential mediators (e.g., as cognitive or noncognitive inputs) impacting adulthood outcomes. Estimating Head Start impacts on these earlier outcomes could thus be informative about the processes underlying the pattern of later impacts.²⁶

²⁶ We conducted this section's analysis as in Deming (2009) and considered identical outcomes and age groups. The full set of estimates was compiled in the Supplementary Tables S12–S13 and S19–S20 in the online version of the journal.

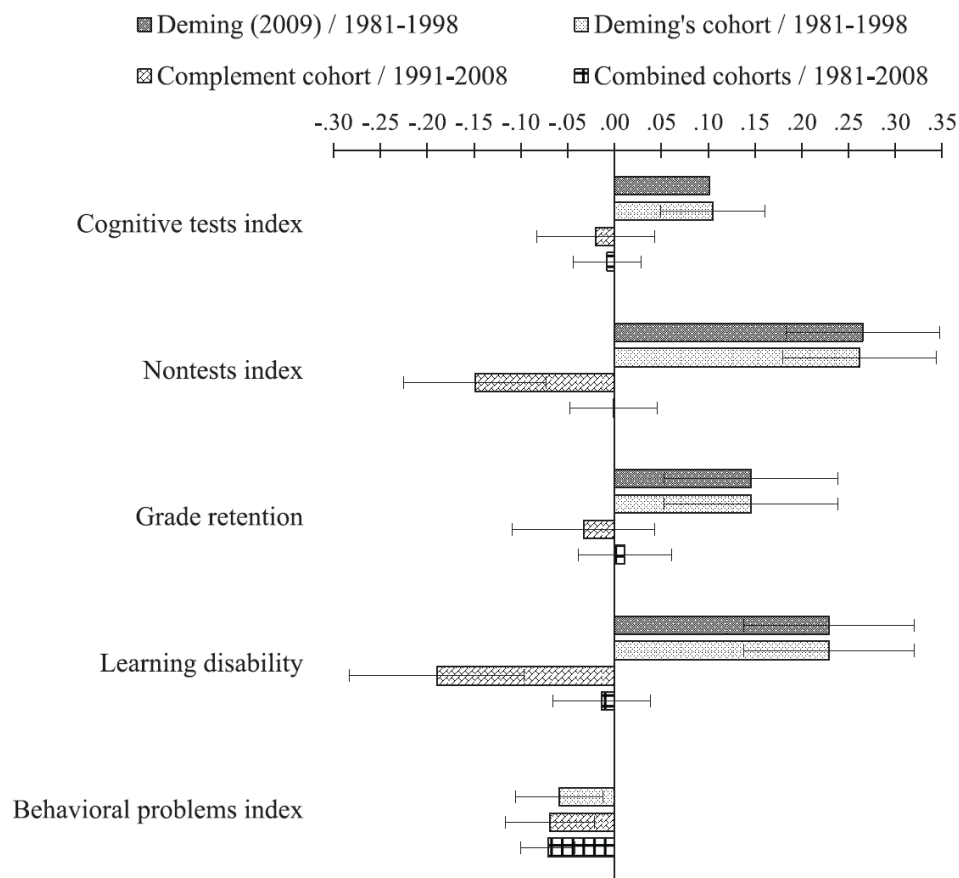
As shown in Figure 1.5, estimates from Deming (2009) and our replication of Deming (2009) were aligned. For the complement cohort, patterns of impacts on school-age outcomes mirrored impacts on the adulthood outcomes: They went in the opposite direction. This was also the case for the nontest index (-0.15 SD; $SE = 0.08$), with Head Start's impact on the learning disability diagnosis indicator (reverse scaled; -0.04 SD) being statistically significant at the 5% level. This is in line with the adverse impacts recorded on the Behavior Problems Index (reverse scaled; -0.07 SD; $SE = 0.05$). Yet, we could not detect any Head Start impact on "Grade retention" (whereas for Deming's cohort, the impact of being grade retained was at a lesser 7 *pps*, significant at the 10% level).²⁷ Regarding the cognitive tests index, the relatively sustained gains generated by Head Start for Deming's cohort (0.11 SD; $SE = 0.06$) did not reflect those for the complement cohort (-0.02 ; $SE = 0.06$), while equality between the two estimates could not be rejected ($p = .24$). The cognitive tests and behavioral problems indices considered in Figure 1.5 were scored as the overall average of all corresponding index scores measured from age 5 to 14. We also considered age periods 5 to 6, 7 to 10, and 11 to 14 (see Supplementary Tables S14 S15 in the online version of the journal). Head Start impacts on the BPI index were stable across age groups and cohorts and were of similar magnitude as the 5 to 14 average.

Deming (2009) reported some fadeout of Head Start impact on the cognitive tests index by ages 11 to 14: from an estimate of 0.15 SD ($SE = 0.09$) by age period 5 to 6 to one of 0.06 SD ($SE = 0.06$). In contrast, for the complement cohort, a fadeout from a small but positive estimate (0.06 SD; $SE = 0.07$) might have occurred faster by age period 7 to 10 (-0.03 SD; $SE = 0.06$): The difference in impact with Deming's cohort for this age group (0.13 SD, $SE = 0.06$) was of marginal significance ($p = .12$). Complement cohort estimates ended at -0.05 SD ($SE = 0.07$) by

²⁷ Other preschool impact estimates had similar trends on all these "noncognitive" outcomes (see Supplementary Tables S12–S13 in the online version of the journal).

age period 11 to 14. Finally, the combined cohorts sample faced a similar trend as the complement cohort; overall, impacts approached zero earlier for later Head Start cohorts.

Figure 1.5: Head Start Impacts on School-Age Outcomes Across Cohorts



Notes: Measurement periods are displayed to the right of each label. Impacts are expressed in standard deviation units. Deming's cohort, N = 1,251; complement cohort, N = 2,144; combined cohorts, N = 3,768. The counterfactual was no preschool attendance. Error bars represent standard errors which were clustered at the family level. Estimates were oriented such that a positive value represents a more favorable outcome. Deming (2009) estimated impacts on Behavior Problems Index (not statistically significant) but did not report them. "Grade retention" and "Learning disability" composed the "Nontest index."

In our analytic model (see Equation 1), other preschool was also included as a within-family predictor of adult outcomes. Considering later and combined cohorts patterns of school-age outcomes by age period, we find that for both the Head Start and other preschool measures, impacts on cognitive outcomes were positive at treatment outset (age 5–6), followed by fadeout,

possibly occurring earlier for Head Start participants (see model (5) in Supplementary Table S14 in the online version of the journal). Second, impacts on the nontest score index were unfavorable and statistically significant, for both preschool statuses (see Supplementary Table S10 in the online version of the journal). Third, impacts on the Behavior Problems Index were sometimes significant, mostly similar in magnitude, and unfavorable over all age periods across both preschool groups (see model (5) in Supplementary Table S15 in the online version of the journal). Overall, we could never statistically reject the equality of estimates between Head Start and other preschool status for any of the considered school-age and adulthood outcomes for the later cohorts of siblings (see respective top panels of Supplementary Tables S10–S13 in the online version of the journal).

1.4.4 Robustness Checks

As noted above, while the FFE design has been a workhorse empirical strategy to estimate the causal impact of Head Start (Bauer & Schanzenbach, 2016; Currie & Thomas, 1995; Deming, 2009; Garces et al., 2002), recent research has called into question this approach both in terms of the external (D. L. Miller et al., 2019) and construct validity of the FFE design (Heckman & Karapakula, 2019). The following section will discuss these threats and how, if at all, our primary findings change as a result.

1.4.4.1 Selection Into Identification

D. L. Miller et al. (2019) showed FEs can induce nonrandom selection of individuals into the FE identifying sample, leading to biased FE estimates relative to the average treatment effect (ATE) unless reweighting on observables is completed. In the FFE context, it may be that families with differential sibling participation in Head Start are systematically different across a variety of measures compared with those families with siblings that do not have variation in

preschool status. If present, this “selection into identification” (SI) is a threat to the external validity of our results and could lead to a biased FFE estimate of Head Start’s impact compared with the ATE (D. L. Miller et al., 2019).

To address this potential problem, following our above FFE analysis, we performed the reweighting-on-observables procedure discussed in D. L. Miller et al. (2019), first checking whether the combined cohorts FFE identifying sample exhibited SI across a variety of observables, including child’s birth cohort, family size, mother’s age at child’s birth, permanent income, maternal AFQT, and whether a child was African American. using the “switcher” and “non- switcher” naming convention found in D. L. Miller et al. (2019)—where (non-)“switcher” represents those families with (no) sibling variation in preschool status—and estimating propensity scores via multinomial logistic regression, we found differences over all family characteristics between Head Start participating “switcher” families and non-Head Start “non switcher” families. By contrast, the only statistically significant difference ($p < .001$) in predictors between Head Start “switcher” versus “non-switcher” families (i.e., in which all siblings attended the program) was the indicator whether a child was African American.

Given that the FFE identifying sample for the combined cohorts exhibited some degree of SI, we corrected the potentially biased FFE estimates using the one-step reweighting-on-observables procedure in D. L. Miller et al. (2019). Overall, we found no evidence for the combined cohorts that reweighting changed the estimates of Head Start’s impact on young adult outcomes, including high school graduation and ASI. We also performed the D. L. Miller et al. (2019) reweighting-on-observables procedure for the Deming cohort sample. Similar to D. L. Miller et al. (2019), we found reweighting attenuated the FFE estimates of Head Start for high school graduation, idleness, learning disability, and poor health, but found little change on other

young adult outcomes or ASI. These results closely replicate D. L. Miller et al.'s (2019) findings and can be found in Supplementary Table S16 in the online version of the journal.

1.4.4.2 Spillovers

The construct validity of FFE estimates has also recently come into question. Revisiting the Perry Preschool Project, Heckman and Karapakula (2019) showed siblings who participated in the program had large positive spillovers on their nonparticipating siblings, particularly for male siblings. Although this problem is material to the FFE design, it is of less concern for this article as we were interested in comparing FFE results across cohorts and between short-run and long-run outcomes. In addition, it is unclear how sibling spillovers could explain this article's main findings, unless sibling spillover effects had become much stronger in the later as opposed to earlier cohorts.

That said, given Head Start's provision of wraparound services, some of which may influence parenting practices, some degree of spillover across children is likely (Deming, 2009; Ludwig & Miller, 2007; U.S. Department of Health and Human Services, Administration for Children and Families, 2019). Garces et al. (2002) and Deming (2009) test for Head Start treatment spillovers by interacting an indicator for first-born status with an indicator for Head Start treatment status. If spillover effects are present from an older to younger sibling, one would expect the impact of Head Start to be larger for non-first-born siblings (Deming, 2009). Consistent with Deming (2009), we found sparse and inconsistent evidence of spillovers. This holds true for both the Deming cohort and combined cohorts samples (see Supplementary Tables S17–S18 in the online version of the journal).

1.4.5 Reconciliation of Cross-Cohort Results

1.4.5.1 Human Capital Index

One possible explanation for cohort differences in the estimated impacts of Head Start is that the more recent cohort was more advantaged, and thus less likely to benefit from Head Start. To check if this was the case, a household human capital factor was constructed (Cronbach's $\alpha = .83$) by combining standardized measures of maternal, and both grandparents', education levels, maternal AFQT, the natural logarithm of family permanent income, and the CNLSY Home Observation Measurement of the Environment short-form (HOME). This human capital factor was then interacted with indicators for Head Start and other preschool from Equation 1. The interaction impact estimate on ASI, for the combined cohorts sample, was not statistically significant for any value within the range of the household human capital factor.

1.4.5.2 Cohort Covariates

To check whether the difference between Deming's cohorts and complement cohorts was due to some of the covariates we presented above, we interacted Head Start with an indicator for whether siblings belong to the complement cohort, along with interactions of Head Start with a series of covariates: first interacting one covariate at a time with the main effect included, then interacting all covariates with all main effects included. These covariates included pretreatment index, family human capital index, mother's age at child's birth, child's age at outcome measurement, indicators for gender, whether White/Hispanic or Black, and whether maternal AFQT score was 1 SD below the mean. In addition, the 2007–2009 Great Recession could have negatively impacted complement cohort siblings who lived through its aftermath as teenagers. Thus, an indicator created for complement cohort siblings between age 12 and 18 in 2008 was also added. Had any of these interactions substantially reduced the estimate from the

Head Start $\times$ Cohort interaction, then these covariates would have explained some of the cross-cohort change in Head Start impacts between the Deming and complement cohort. We did not find any evidence that this was the case (see Supplementary Table S19 in the online version of the journal).

1.4.5.3 Blinder–Oaxaca Decomposition

A Blinder–Oaxaca decomposition approach (Blinder, 1973; Oaxaca, 1973) allowed us to consider how much a group mean difference on an outcome of interest (Y) is explained by group differences in predictors. using the Deming cohort group (d) and complement cohort group (c), we formalized a threefold decomposition (Jann, 2008) as follows:

$$Y_d - Y_c = E + C + I$$

where $E = (E_d - E_c)$ measures the part of complement cohort's expected change in ASI, with complement cohort's predictors' means (i.e., endowments) fixed at Deming's cohort levels; $C = (C_d - C_c)$ measures the part of complement cohort's expected change in ASI, with complement cohort coefficients fixed at Deming's cohort levels; and $I = (E_d - E_c) \times (C_d - C_c)$ measures the contribution of the interaction between endowments' and coefficients' respective cross-cohort differences. The results from this analysis highlight key differences between Deming's and complement cohort samples, and showed how these differences most notably in the pretreatment index and mother's age at child's birth— drive variation in estimated Head Start impacts across cohorts.

Breaking down the threefold decomposition results, we first found that the mean difference in ASI for Head Start attendees between the Deming cohort and complement cohort

($Yd - Yc$) was -0.22 SD (SE = 0.09).²⁸ The direction of this mean difference favored the complement cohort and was statistically significant. Second, having chosen the pretreatment index and mother's age at child's birth as predictors in the analysis,²⁹ we found that predictors' endowment parts of the decomposition (E) explained all of the outcome group mean difference (-0.27 SD; SE = 0.10), with mother's age at child's birth endowment recovering the near totality of it (-0.21 SD; SE = 0.10). The coefficient (C) and interaction (I) parts of the decomposition were negligible and not statistically significant ($p > .84$; $p > .95$). In sum, with mother's age at child's birth fixed at Deming's cohort level, complement cohort Head Start attendees would share an ASI expected value similar to that of their Deming's cohort counterparts.

Furthermore, if, for siblings of the complement cohort's counterfactual no preschool group, the expected change in ASI mean was also explained by mother's age at child's birth (as endowments, coefficients, or interaction effect), then, keeping this factor equal, Head Start would have had an impact of similar magnitude across cohorts. Thus, a threefold decomposition was conducted for counterfactual no preschool group siblings. Cross-cohort mean difference on the outcome of interest ($Yd - Yc$) was moderate and statistically significant (-0.56 SD; SE = 0.09). However, the E and C components of the decomposition were negligible ($p > .88$; $p > .74$), whereas the I component—that is, the interaction of mother's age at child's birth endowments' and coefficients' differences—recovered the outcome differential (-0.61 SD; SE = 0.18).

Because of the NLSY design, later cohorts of children had, on average, older mothers. Earlier cohorts (i.e., Deming's cohort) were born to relatively younger mothers, many of which may have disproportionately benefited from Head Start due to the program's provision of

²⁸ In this complementary analysis, standard errors were clustered at the family level.

²⁹ These predictors were selected based on both the magnitude and direction of the mean covariate differences between the Deming and complement cohort. Mother's age at child's birth was around 7 years higher, and the mean pretreatment index statistically more favorable (0.20 SD; SE = 0.08) for complement cohort's Head Start attendees.

wraparound services that emphasize parental involvement (Currie & Neidell, 2007; Deming, 2009). This factor alone might have contributed to the discrepancy in our estimated Head Start impacts between Deming’s and later cohorts. In particular, within Head Start participating families, siblings who did not attend a preschool program appeared to have on average benefited the most from having an older mother during their early years. Finally, the explanatory power of the mother’s age at child’s birth predictor remained robust after adding other covariates to the Blinder–Oaxaca decomposition, including permanent family income, whether the mother attended college (Table 1.2), a family human capital index, and family size (see Notes of Supplementary Table S19 in the online version of the journal).

1.4.5.4 Secular Trends

Our original cohort analysis looked only at the impact of Head Start across two cohorts—the Deming cohort and the complement cohort—obscuring whether over time there was a gradual secular decline or precipitous drop in the effect of Head Start. To address this, we decomposed our overall sample into three new birth-cohort-year groupings: C1—families with all siblings born before 1983, C2—1983–1987, and C3: post-1987. More than 90% of the C1 siblings and around a third of the C2 siblings were part of the Deming cohort sample. Two thirds of the C2 siblings and all of C3 siblings were part of the complement cohort sample.

Across virtually all of our outcomes, Head Start more favorably affected the C1 cohort compared with C2 and C3 cohorts. Moving from older to more recent cohorts, we observed unfavorable sign changes in the direction of Head Start’s impact for a variety of short-term and longer-term outcomes, including cognitive and noncognitive school-age measures, ASI, educational attainment, and earnings.³⁰ In addition, the estimated trends indicate a comparatively

³⁰ We also checked for impacts on an alternative earnings variable, taking this time the natural logarithm of the most recent yearly earnings available in CNLSY 2014 survey round. Head Start impacts were never statistically significant.

larger decline in adulthood outcomes for individuals belonging to the most recent post-1987 cohort. In contrast, for school-year nontest outcomes, the decline was observed in much earlier cohorts. As reported in the Supplementary Table S20 in the online version of the journal, Head Start impacts between C1 and C3 on ASI went from positive and statistically nonsignificant (0.09 SD; SE = 0.09) to negative and significant at the 10% level (-0.25 SD; SE = 0.14). By contrast, Head Start impact on the nontest index went from a statistically significant 0.29 SD (SE = 0.10) advantage for C1 to a negligible -0.06 SD (SE = 0.15) for C3, with the drop occurring between C1 and C2 (-0.10 SD; SE = 0.15).

1.5 Discussion and Conclusion

In this study, we replicated and extended Deming's (2009) evaluation of Head Start impacts over life cycle skill formation. We found mixed results for Deming's cohort of siblings, after having extended adulthood individual outcomes with an additional decade of CNLSY data. Second, replicating Deming's analytic framework on children born to CNLSY mothers after the children in Deming's cohort revealed contrasting patterns of impacts. In general, for this new cohort, impacts were negative. In fact, for our study of more recent cohorts, Head Start participation might have been detrimental, relative to home care, on noncognitive and behavioral measures or on the adulthood summary index. Third, combining both cohorts produced Head Start impact estimates on all measured outcomes that were small and not statistically significant. Finally, we conducted a handful of empirical tests to determine what is driving the differences in Head Start impacts across birth cohorts. Although not definitive, our results suggest the important role factors outside of Head Start, notably mother's age at child's birth and the pre-treatment index, played in moderating the program's impacts on participating children's young adult outcomes.

Given the benefit of time and 10 additional years of NLSY data, this article provides an updated estimate of Head Start’s impact on observed adult wages for the Deming cohort. Deming (2009) originally estimated a 0.11 log points Head Start impact on adult wages; however, as survey participants were too young at the time of the calculation to report actual wages, this estimate was based on Deming’s projection of future adult wages. Using the now available realized adult wages for siblings 25 years old or more, unconditional on employment and averaged across each survey round up to 2014, we follow up on his initial projection finding a nonstatistically significant smaller impact at 0.07 log points ($SE = 0.12$). Using dollar (2014 US\$) instead of log earnings produces a negative and still nonstatistically significant impact of $-\text{US\$}999$ ($SE = \text{US\$}1,507$). In sum, Head Start generated no clear adult earnings gain for the Deming (2009) cohort siblings, although the large confidence interval on our estimate includes Deming’s original estimated impact.³¹

Johnson and Jackson (2019) analyzed earlier cohorts of siblings born before 1976 with a dynamic complementarity design (i.e., capitalizing on two exogenous sources of variation separated in time). They found a larger, more precise estimate: Attending Head Start at age 4—that is, facing an average Head Start spending versus no spending, coupled with (and sensitive to) an average public K–12 spending—boosted earnings of poor children (measured from age 20–50) by 0.10 log points ($SE = 0.02$).³² These positive estimates, generated from a different identification strategy and earlier cohorts, are close to Deming’s projections, but the variation and imprecision associated with our estimates underscores the importance of monitoring cross-

³¹ Earnings (log transformed or not) regression estimates were very similar—0.05 log points; $SE = 0.11$; $-\text{US\$}1,030$; $SE = 1488$, respectively—whether age was controlled for instead of year of birth. As described in the “Results” section, for Deming’s cohort, attending Head Start versus no preschool yielded 0.3 years increase of completed schooling. From this and based on Card’s (1999) review on returns to education of about 5% to 10% per year of completed schooling, we might have expected to find evidence of an impact on earnings between 1% and 3.5%.

³² That estimate appeared to be sensitive to subsequent K–12 spending level: coupled with a 10% decrease in K–12 spending, the estimate fell to 0.03 log point ($SE = 0.03$) and was no longer statistically significant at the 10% level. In contrast, with a 10% increase in K–12, the estimate jumped to 0.17 log point ($p < .01$).

cohort trends in Head Start impacts.

Apart from revisiting and extending Deming (2009), this article adds to the current Head Start literature by finding heterogeneous program affects across CNLSY birth cohorts, perhaps due to cohort differences in resources that might serve as substitutes for Head Start. We find evidence of a notable rightward shift in the distribution of mothers' age at child's birth and in the pretreatment index across time, the latter indicating more favorable household conditions across time for mothers and their children in the years leading up to Head Start eligibility. As discussed in Deming (2009), this feature is an artifact of the CNLSY sampling design. Subsequent findings, primarily from our threefold Blinder–Oaxaca decomposition analysis, suggested that the changing effects of Head Start across cohorts were due largely to changes in covariates of the identifying sample, notably mother's age at child's birth, and not due to changing effectiveness of Head Start across time.

Although this study found that the counterfactual conditions of household characteristics (e.g., maternal age) predicted cross-cohort differences on long-term outcomes, additional factors may be driving these variations. Both the United States' substantial increase in spending on means-tested programs between the 1980s and 2010s and possible changes to program quality (e.g., due to a steep and continuous enrollment increase) could also explain these observed changes (Ludwig & Phillips, 2008), as could changes in labor market conditions and the return to specific forms of human capital over time. For example, changes in human capital returns over time may have induced individuals from more recent birth cohorts to invest less in educational attainment relative to earlier older birth cohorts. Our finding that Head Start negatively affected “Idleness” (i.e., not employed or enrolled in school) and “Some College Attended” for more recent birth cohorts, but not for earlier cohorts, lends support to this claim. Each of these

explanations is potential avenue future research could take to further explore these results.

Overall, this article suggests that understanding and eliciting pathways of early skill formation with potential subsequent complementarities could be an important priority for basic human capital research and education policies. The novelty of these findings, combined with the possibility of unobserved changes in the selection process into Head Start during this period, necessitates further research on recent cohorts of Head Start attendees using complementary identification strategies.

Chapter 2

A Lingerin Legacy: The Relationship Between 1930s HOLC “Redlinin” Maps and School Fundin, Diversity, and Performance^{33 34}

2.1 Introduction

The United States has a long history of racially discriminatory policies and practices. Redlinin, where the Home Owners’ Loan Corporation (HOLC) assigned A-D security ratings to nearly 240 cities across the United States between 1935 and 1940, is a small but important part of this history.³⁵ With the advent of recently digitized HOLC maps by the “Mapping Inequality Project” headed by the University of Richmond’s Digital Scholarship Lab, the long-run impact of HOLC A-D security ratings on social and economic outcomes has become increasingly well documented by social scientists. While much of the literature today shows HOLC maps’ adverse effects on modern outcomes such as credit scores, family structure, homeownership, home values, household income, neighborhood segregation, and incarceration (Appel & Nickerson, 2016; Krimmel, 2018; Anders, 2018; Aaronson et al. 2020; Aaronson et al., 2021), very few studies, if any, look at the long-term relationship between HOLC maps and contemporary educational outcomes. We hypothesize that the association between educational outcomes and neighborhood quality is intertemporal such that U.S. neighborhood inequality from the early 20th century, often resultant from historical discriminatory practices, is predictive of present-day educational outcomes. With recent work by Aaronson et al. (2020) and Aaronson et al. (2021)

³³ This chapter is published as a working paper in *EdWorkingPapers* (<https://doi.org/10.26300/qeer-8c25>). Full Citation: Lukes, D., & Cleveland, C. (2021). The Lingerin Legacy of Redlinin on School Fundin, Diversity, and Performance. *EdWorkingPaper* No. 21-363. Providence, RI: Annenberg Institute at Brown University.

³⁴ Dylan J. Lukes and Christopher Cleveland are co-authors. Please refer to Appendix D for the official co-authorship statement.

³⁵ HOLC and Federal Housing Administration (FHA) maps are not the same. HOLC maps were created first in 1935 under the Federal Home Loan Bank Board (FHLBB) as part of a systematic appraisal process by the HOLC to evaluate the lending risk for loans they had already issued to mitigate non-farm mortgage defaults and foreclosures during the Great Depression (Fishback et al., 2020; Aaronson et al., 2020). The FHA, also under the FHLBB, created its maps and used them to determine whether homes qualified for mortgage insurance. While historians agree that the HOLC maps influenced the FHA maps, it is unknown to what degree since all but two FHA maps (Chicago, IL; Greensboro, NC) have been lost or destroyed (Light, 2010).

establishing a causal relationship between HOLC maps and neighborhood segregation, housing, and socioeconomic outcomes, it would be surprising if these results did not spill over to measures of educational quality given the deep-rooted connection between neighborhoods and the schools and districts that serve them (Massey & Denton, 1993; Aaronson, 1998; Chetty et al., 2016; Bayer et al., 2020; Dalane and Marcotte, 2020). To this end, our paper addresses the following questions: (1) How do historic HOLC D neighborhoods relate to existing patterns of district-level and school-level funding, student racial diversity, and student performance? (2) Do these patterns vary by region? And (3) How, if at all, do these patterns vary over time? In doing so, we hope to contribute to the existing body of research connecting HOLC maps to modern-day outcomes by providing some of the first evidence of the long-term association between 1935-1940 HOLC A-D grades and educational outcomes.

To link historical HOLC maps to contemporary educational institutions, we map 1935-1940 HOLC A-D neighborhood grades to present-day schools and districts. For schools, individual HOLC A-D grades are assigned based on overlapping school-level latitude and longitude geolocations and HOLC A-D geospatial polygons. For districts, we determine HOLC A-D mappings using the area, in square miles, of HOLC A-D polygons that overlap with each respective district boundary. We use these mappings to analyze the relationship between HOLC A-D security ratings and our district and school outcomes, both modern-day and over time, using various plots and statistical tests.

Our analysis identifies a few important findings. First, for district-level funding, present-day districts located in predominantly D neighborhoods have less district-level per-pupil total revenues but generally higher per-pupil federal and state revenues than those mapped to HOLC A, B, and C security ratings. However, the differences in per-pupil federal and state revenues are

not large enough to overcome the sizeable gaps in per-pupil local funding that favor those districts mapped to higher HOLC security ratings. These local, state, and federal dynamics drive the differences in per-pupil total revenue by HOLC A-D grades. Conversely, at the school level, we find that schools mapped to HOLC D ratings have, on average, more per-pupil total expenditures and per-pupil federal expenditures than schools mapped to HOLC A, B, and C security ratings. Differences in aggregated per-pupil state and local expenditures between D vs. A, B, and C schools are statistically insignificant and likely signal the countervailing force that state funding has on local funding, which favors higher-rated HOLC security ratings. In reconciling the district and school finance results, we find that schools mapped to HOLC D grades have the largest shares of students qualifying for free and reduced-price lunch, making them eligible for Title I funding, the largest federal funding program for U.S. public schools. This finding provides a plausible mechanism for why HOLC D schools, regardless of what HOLC A-D district they fall within, have the most per-pupil funding of all HOLC A-D mapped schools. Furthermore, our district-level results remain pertinent as they represent an upper limit on how much a district can allocate to their most-in-need schools (i.e., D schools) while still maintaining funding levels for their other district schools.

Second, for student racial diversity, we find that modern-day schools located in HOLC D neighborhoods have larger school-level shares of Black and non-White student bodies and less diverse student populations than their more highly rated HOLC A, B, and C counterparts, nationally and by region. While schools in higher-rated HOLC A, B, and C neighborhoods have relatively larger shares of White student bodies, they also have, on average, more diverse student populations. Even more, school diversity is monotonically increasing in HOLC rating, such that A schools have the greatest levels of diversity, followed by B, C, and ending with D.

Third, for student performance, we find that schools today located in historic HOLC D areas have worse school-level average math and reading scores than their more highly rated A, B, and C peers, nationally and by region. We also observe virtually no differences in either measure of average learning rates or educational opportunity changes by A-D HOLC grade. These results are generally true nationwide and by region. Finally, we document positive time trends for the finance and diversity outcomes across all HOLC A-D grades from the late 1980s to today, but persistent and widening gaps between schools in historically redlined D neighborhoods and those in A, B, and C neighborhoods. We do not document the time series trends for our school performance outcomes since measures are only available as pooled estimates spanning 2009 to 2018.

In summary, our paper sheds light on the previously unexplored relationship between 1935-1940 HOLC maps and modern district and school-level outcomes. Overall, our paper provides some of the first evidence documenting the association between historic HOLC A-D grades, which captured neighborhood inequality in the 1930s and subsequently contributed to neighborhood inequality in the following generations, and modern-day educational outcomes. While we cannot definitively say whether HOLC redlining caused the modern-day educational inequalities we show in this paper, HOLC A-D grades' predictive ability hint at a stubborn historical legacy. We believe these results indicate the need for educational policymakers to consider the historical implications of past neighborhood inequality on present-day neighborhoods when designing modern education interventions focused on improving the life outcomes of students of color and students from socioeconomically disadvantaged communities.

The remainder of this paper is organized as follows. Section 2.2 provides an overview of the history of HOLC maps, the current set of literature that links HOLC security ratings to social

and economic outcomes, and related literature on neighborhoods and schools. Section 2.3 describes the data used for this research. Section 2.4 and Section 2.5 outlines this paper's methods, including how analysis samples were constructed, and details the analytic approach. Section 2.6 provides the results and a discussion. Section 2.7 concludes.

2.2 Literature

2.2.1 HOLC Maps History

In 1932, the Federal Home Loan Bank Board (FHLBB) was created to manage federal savings and loan associations. In 1933, under the FHLBB, the Home Owners' Loan Corporation (HOLC) agency was subsequently made to oversee the troubled U.S. mortgage market and tasked with purchasing and refinancing non-farm mortgages to limit foreclosures and defaults. After the HOLC completed issuing loans to distressed properties, to evaluate their lending risk, they created a systematic appraisal process that included neighborhood-level characteristics such as race, ethnicity, immigration status, household income, homeownership rates, access to public services, and occupation type (Hillier, 2003; Crossney & Bartelt, 2005; Fishback et al., 2020). Between 1935 and 1940, HOLC's department of Research and Statistics used thousands of realtors, developers, lenders, and appraisers to create neighborhood-by-neighborhood security ratings of 239 cities and made over 5 million appraisals (Hillier, 2003; Crossney & Bartelt, 2005). Neighborhoods were graded on a scale of A (i.e., least risky/most stable) to D (i.e., most risky/least stable) based upon the perceived risk of making housing loans in different neighborhoods. These grades were later solidified as color-coded maps in which D-graded areas were colored red and prompted the later coinage of the term "redlining."

Each neighborhood had a standardized assessment sheet used to assign the HOLC grades. For example, area descriptions used in Los Angeles County in 1939 included eight sections. 1)

Population asked for a record of whether the population was increasing, decreasing, or static. It also asked for the class and occupation of residents, the percentage of foreign families and their nationalities, the percentage of negro families, and whether the population trends reflected shifting or infiltration. 2) *Buildings* asked for the type and size of the building, construction, average age, repair status, occupancy rate, owner-occupied, 1935 price bracket, 1937 price bracket, 1939 price bracket, sales demand, predicted price trend, 1935 rent bracket, 1937 rent bracket, 1939 rent bracket, rental demand, and predicted rent trend. 3) *New Construction* captured the number of new properties built within the past year, the prices for these units, and how they were selling. 4) *Overhang of Home Properties* captured unsold HOLC properties. 5) *Sale of Home Properties* captured sold HOLC properties 6) *Mortgage Funds* captured mortgage funds' availability. 7) *Total Tax Rate per \$1000* captured the local tax rate. 8) *Description and Characteristics of Area* captured other qualitative detail about the terrain and population.³⁶

A current debate exists as to whether HOLC A-D grade assignments were racially biased or merely a geographic snapshot of an outcome caused by America's long history of racial discrimination before it (Fishback et al., 2020). There is evidence that race, immigration status, household income, and ethnicity were often explicit factors in the HOLC grading process, such that non-White racial groups, immigrants, low-income households, and ethnic minorities were more likely to receive lower grades (Jackson, 1980; Connolly, 2014; Nelson et al., 2020). However, the degree to which HOLC maps influenced the lending practices and underwriting standards of the FHA and other private lenders is an ongoing debate. Some authors argue that access to HOLC maps was limited, while others argue that access to HOLC maps was ubiquitous and materially influenced public and private lending policies and practices (Jackson, 1980;

³⁶ The area descriptions for Los Angeles A-1 and D-1 can be viewed in Figure B.1 in Appendix B.

Hillier, 2003; Light, 2010; Woods, 2012). Regardless, it is clear that HOLC encouraged the general rule of using maps to classify the creditworthiness and lending risks of neighborhoods and broadly are considered to have redirected public and private capital and homeownership for intergenerational wealth-building to native-born white families and away from African American and immigrant families (Appel & Nickerson, 2016; Krimmel, 2018; Anders, 2018; Aaronson et al., 2020; Aaronson et al., 2021).

2.2.2 Modern Outcomes Linked to HOLC Maps

A literature base that leverages both descriptive and causal empirical strategies is developing that links neighborhoods' HOLC grades to various contemporary outcomes. Mitchell and Franco (2018) descriptively analyze HOLC-graded areas' modern demographic and residential patterns. They find that many neighborhoods rated high-risk or “Hazardous” by HOLC eight decades ago are low-to-moderate income (~74%) and minority neighborhoods (~64%) today. Additionally, the authors find greater economic inequality, higher levels of interaction between Black and White residents, and a stronger positive association between gentrification and economic change in neighborhoods rated “Hazardous” by HOLC in their sample. Regionally, the South showed the smallest change in the HOLC-rated “Hazardous” neighborhoods with relatively lower incomes and more majority-minority residents. In addition, the Midwest region was found to closely mirror the South in the persistence of low-to-moderate income neighborhoods in HOLC “Hazardous” areas.

Three papers have highlighted the impacts of the HOLC maps on housing and socioeconomic outcomes. Appel and Nickerson (2016) use a spatial regression discontinuity boundary design to study long-term impacts on home prices. They show that housing characteristics varied smoothly at the boundaries when the maps were created. Despite this initial

smoothness, they find that HOLC "redlined" neighborhoods have about 4.8% lower home prices, fewer owner-occupied homes, and more vacant buildings relative to adjacent areas. Similarly, Aaronson, Hartley, and Mazumder (2020) use a spatial RD boundary propensity score design to study the effects of the HOLC A-D maps on the long-run trajectories of neighborhoods. They find that the maps led to reduced homeownership rates, house values, and rents and increased racial segregation in later decades. They conclude that the HOLC maps had sizeable and persistent impacts on urban-neighborhood development driven by reduced credit access and the subsequent neighborhood disinvestment (Aaronson et al., 2020). In a follow-up paper, Aaronson, Faber, Hartley, Mazumder, and Sharkey (2021) extend Aaronson et al. (2020) to study the impact of HOLC A-D maps on a variety of socioeconomic outcomes such as income rank, family structure, incarceration, and geographic mobility. Using data from the Opportunity Atlas and an identification strategy similar to Aaronson et al. (2020), they find sizeable and statistically significant differences across all outcomes that favor higher-rated HOLC neighborhoods over lower-rated ones.

Other papers have explored a broader set of outcomes connected to the HOLC maps. Jacoby, Dong, Beard, Wiebe, and Morrison (2018) use descriptive spatial analysis to evaluate the relationship between HOLC grades and modern firearm violence in Philadelphia, PA. Using data from the 1940 U.S. Census the authors adjust for socio-demographic factors at the time of HOLC-map creation and find that firearm injury rates are highest in historically HOLC D areas of Philadelphia (Beard et al., 2017; Jacoby et al., 2018). Nardone, Casey, Rudolph, Karasek, Mujahid, and Morello-Frosch (2020) descriptively explore how birth outcomes within California vary based upon HOLC grade. The authors find that worse HOLC ratings are associated with adverse birth outcomes. However, their findings are inconsistent when using propensity score

matching and stratifying by metropolitan area (Nardone et al., 2020). Hoffman, Shandas, and Pendleton (2020) use descriptive spatial analysis to study the association between a neighborhood's HOLC grade and modern-day surface temperatures. They find that for nearly all areas in their study, previous HOLC redlined areas have consistently elevated land surface temperatures compared to non-redlined areas. Cities in the Southeast and Western regions display the largest differences in land surface temperatures, while those in the Midwest region show the least. Overall, the authors find that nationally, land surface temperatures in previously HOLC redlined locations are about 2.6 °C hotter than those in non-redlined ones.

These recent studies suggest a strong association between historical practices for assigning grades to neighborhoods and modern outcomes. The HOLC maps and related historical policies have contributed to racial disparities in neighborhood diversity, income, health care, access to healthy food, incarceration, and public infrastructure investment. This paper's main contribution is to explore the relationship between HOLC map grades and present-day educational outcomes of school and district funding, school racial diversity, and school performance. As discussed in the following section, these outcomes are intricately connected to U.S. housing and neighborhoods and have yet to be explored.

2.2.3 The Relationship Between Schools, Neighborhoods, and Student Outcomes

The academic literature highlights the inter-relatedness of school funding, school racial diversity, and broader neighborhood contexts on students' long-term outcomes. Education funding in the U.S. operates at federal, state, local, and district levels. The two primary federal revenue sources are Title I and IDEA, which generally distribute dollars based on student population size and poverty concentration. At the state level, each state has different mixes of revenue streams dedicated to education and implements a unique funding formula to direct

dollars to school districts. At the local level, districts primarily use property taxes to generate revenue. Within a district, funding is generally distributed through a traditional centralized model in which the district deploys resources to schools in the form of staff, programs, and services (Urban Institute, 2017; Roza, Hagan & Anderson, 2020; Baker et al., 2018; Brittain, Willis & Cookson, 2019; Green et al., 2021).

A recent study has illuminated the connection between school quality and neighborhood value. Using multiple decades of U.S. housing price data, Bayer et al. (2020) conduct a national study on the causal effect of school spending and local taxes on housing prices. The authors find that households greatly value school spending, including spending on teachers and staff salaries. Additionally, they show that salary spending is allocated inefficiently throughout the U.S. and find that within-district-salary-expenditure raises funded via local taxes would increase home prices.

Interrelatedly, there has been a growing focus on race- and income-based segregation within districts and schools. Dalane and Marcotte (2020) find that classroom segregation increased by around 10 percent between 2007 and 2014 for those elementary and middle schools in their study. Further, they find that segregation of economically disadvantaged students within schools is associated with the level of segregation between schools in districts and that this correlation grew stronger across the panel years. These findings of increasing segregation in certain parts of the country have been highlighted in several other recent studies (Clotfelter et al., 2019; Alcaino & Jennings, 2020; Clotfelter et al., 2021; Monarrez et al., 2020). The salience of neighborhoods became well documented through the efforts of Massey & Denton (1993) to link persistent poverty among blacks in the United States to the unparalleled degree of deliberate segregation they experience in American cities. Chetty, Hendren, and Katz (2016) found from

the Moving to Opportunity experiment that a reduction in neighborhood poverty had no impact on short-term reading and math test scores for Black children but improved early-adult outcomes, including educational attainments and labor market earnings. More broadly, the recent development of the Opportunity Atlas shed light on the average adult outcomes of individuals who grew up in each U.S. Census tract to trace back the roots of poverty and incarceration (Chetty et al., 2018; Chetty & Hendren, 2018a; Chetty & Hendren, 2018b; Murnane, 2021).

Given this literature base, this paper's primary contribution is to extend the roots of perceived issues in school funding, racial diversity, and performance to the inception of HOLC redlining to demonstrate the long-term relationship between HOLC A-D security ratings and present-day educational outcomes.

2.3 Data

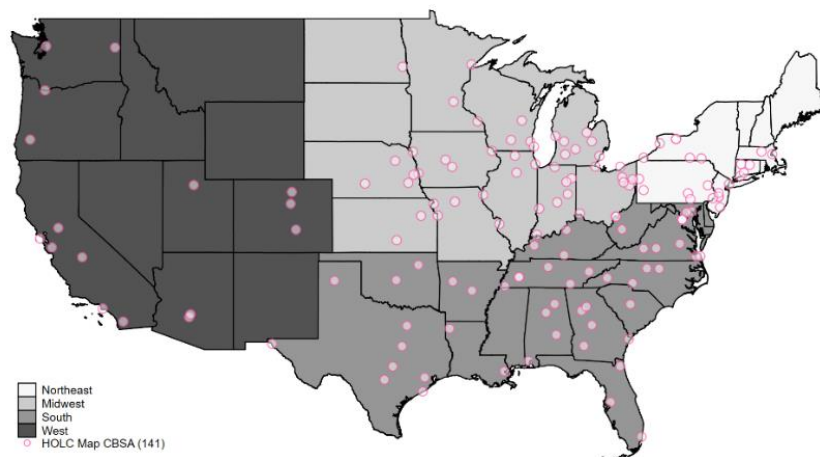
In what follows, we provide a succinct overview of the district, and school-level data sources we used in our analysis, in addition to the digitized Home Owners' Loan Corporation (HOLC) maps created by the "Mapping Inequality" project (Nelson et al., 2020). For further details, please refer to Appendix Table B.1-B.3.

2.3.1 HOLC Data

The 1935-1940 HOLC maps were preserved by the U.S. National Archives and recently digitized by a team at the University of Richmond, Virginia Tech, University of Maryland, and Johns Hopkins University as part of the "Mapping Inequality" project (Nelson et al., 2020). This data contains digitized versions of every available city-level HOLC map from 1935 to 1940 in Shapefile or GeoJSON format. The spatial data includes over 7,000 neighborhoods and nearly 240 unique cities across the United States. For each HOLC city, the data contains HOLC Geographic Information System (GIS) polygons, HOLC grade assignment text (i.e., A, B, C, D),

and detailed area description transcriptions (Nelson et al., 2020). For our study, we aggregate the data to 144 unique core-based statistical areas (CBSAs), which span the United States and include all four U.S. Census Bureau Regions (i.e., Midwest, Northeast, South, West). These CBSAs are captured below in Figure 2.1. Finally, we include all CBSAs with at least one HOLC grade in our analysis sample to maximize geographic coverage. As a result, some CBSAs do not contain a full A-D HOLC rating set.

Figure 2.1: 1935-1940 HOLC CBSAs by US Census Bureau Region



Notes: The above pink dots represent each unique CBSA in our analysis sample. These total to $n = 144$ present-day CBSAs mapped to the 1935-1940 HOLC Residential Security Maps and are broken down by regions as follows: Northeast ($n=30$), Midwest ($n = 54$), South ($n = 45$), and West ($n = 15$).

2.3.2 NCES Geospatial Data

We combine the HOLC data with geospatial district and school-level data from the National Center for Education Statistics (NCES). Our geospatial data includes district boundaries and public school-level latitude and longitude point locations. The 2018-2019 district boundary data derives from the Census TIGER/2019 Line geospatial data (U.S. Census Bureau, 2020).³⁷

³⁷ We supplement this data with 2018-2019 NYC public school district boundaries from the NYC Department of City Planning (DCP). This supplemental geospatial boundary data was required since all NYC public school districts were reported as one unified district in the NCES EDGE geospatial data. Thus, combining the NYC DCP and NCES EDGE geospatial data allow us to account for each of NYCDOE's 32 school districts and more accurately capture the within-district variation in HOLC grades across the city (NYC Department of City Planning, 2020).

We use the 2018-19 NCES EDGE school-point location data to complete our school-level analyses, which provide latitude and longitude coordinates for public elementary and secondary schools from the NCES EDGE Common Core of Data (CCD) (U.S. Department of Education, 2020).³⁸

2.3.3 NCES Non-Geospatial Data

The NCES non-geospatial data we use is also at the district and school levels. This data includes fiscal and non-fiscal data, which we leverage in our analyses below on district-level financing (i.e., local, state, federal, total) and school-level racial diversity by HOLC grade. We leveraged the most recent NCES 2017-18 F-33 survey data for the fiscal data, which provides general financing information (e.g., revenue and expenditure totals and subtotals) at the district level. This data is provided through NCES CCD (U.S. Department of Education, 2020). We use F-33 reported total general revenues for district-level revenues and their associated first-level revenue subtotals (e.g., local, state, federal). We transform all district-level financial data into per-pupil terms to account for differences in enrollment.³⁹

For the non-fiscal district and school data, we use NCES 2017-2018 and NCES 2018-19 datasets. We analyze each school's Simpson's Diversity Index (1-D) for our within-school-between-student racial diversity outcome. This index captures the likelihood that two randomly selected students from a given school will belong to different racial groups, ranging from 0 to 1, with larger values representing greater within-school-between-student racial diversity (Simpson,

³⁸ Geospatial school boundary files also exist in the NCES School Attendance Boundary Survey (SABS). SABS was an experimental survey led by NCES and supported by the U.S. Census Bureau to collect school boundaries for the 2013-14 and 2015-16 school years. This effort led to the collection of over 70,000 school boundaries across the United States but is now discontinued (U.S. Department of Education, 2020). Due to incomplete nationwide coverage, we opt not to use SABS in this study.

³⁹ As a result, we use district enrollment weights to calculate per-pupil weighted averages by HOLC A-D security rating. In addition, our focus on district revenues versus district expenditures is largely an artifact of variable coverage. NCES district expenditures are not disaggregated at the local, state, and federal levels, but NCES district revenues are. When comparisons are possible, district expenditure results (i.e., per-pupil total, instructional salaries, benefits) mirror district revenues (i.e., per-pupil total).

1949; Hirschman, 1964).⁴⁰ To complement Simpson's Diversity Index, we also consider the Exposure Index, a popular segregation index in the social sciences, and approximates between-group contact and interaction for a given race-group pair (Massey & Denton, 1988). We use this index to evaluate segregation for all schools in our sample by A-D HOLC grade for various majority-minority race-group pairs, including White-Asian, White-Black, White-Hispanic, and White-Non-White.

2.3.4 National Education Resource Database on Schools Data

To complement our district-level finance analysis, we combine the HOLC data with school-level finance data from the Edunomics Lab at Georgetown University (Edunomics Lab, 2021). Their newly released National Education Resource Database on Schools (NERD\$) data includes state-reported 2018-19 school-level expenditures across 49 states and the District of Columbia and captures per-pupil total expenditures, federal expenditures, and combined state and local expenditures.⁴¹ Each per-pupil outcome we use in our analysis was provided directly by NERD\$ and based on state-reported school-level student enrollment versus NCES school-level student enrollment. As a result, there are some discrepancies between the state-reported and NCES school enrollment data, likely due to varying enrollment metrics (e.g., September Count Day, October County Day, Average Daily Attendance (ADA), Average Daily Memberships (ADM), Weighted Enrollment) used by states. However, these differences are often immaterial and do not change our overall school-level findings.

2.3.5 Stanford Education Data Archive Data

⁴⁰ The equation is $1 - D_i = 1 - \sum_{r=1}^R p_r^2$, where p_r represents the probability that two randomly selected students from a given i^{th} school will belong to the same race, and r represents the seven commonly used racial categories by the U.S. Census Bureau, including American Indian or Alaska Native, Asian, Black or African American, Hispanic, Native Hawaiian or Other Pacific Islander, White, and Multiple Races.

⁴¹ By law, districts must report actual dollars spent instead of estimations based on teacher FTE counts and average teacher salaries. From my discussions with NERD\$, nearly every district says they report actual dollars; however, NERD\$ occasionally notices within-district trends that suggest otherwise. Thus, overall NERD\$ data will be actual dollars spent, but in some cases, it will be based on teacher FTE and average teacher salaries.

We used school-level Stanford Education Data Archive (SEDA) data for all our school-performance analyses. This data provides students' academic outcomes in grades 3-8, spanning SY2008-09 to SY2017-18, and includes students' average test scores, test score trends, and learning rates. Average test scores are a school's mean test-based achievement pooled across Math and ELA subjects. Learning rates are the school-grade slope of a school's mean test-based achievement or how test scores change across grades within a cohort. Test score trends are the school-cohort slope of a school's mean test-based achievement or how test scores change across student cohorts within grades (Reardon et al., 2021). Like SEDA, we use "educational opportunities" and "changes in educational opportunities" interchangeably with "average test scores" and "test score trends," respectively. The terms are substituted for another in the sense that test score changes reflect educational opportunities in a community and are influenced by occasions to learn at home, in neighborhoods, in child-care centers, at preschool and after-school programs, and from peers at school (Reardon et al., 2021). We rely on SEDA's cohort standardized (C.S.) scale achievement estimates based on the Spring 2009 4th grade cohort for each student achievement measure. The C.S. scale achievement estimates are measured in S.D. units relative to the national average and are calculated using OLS and Empirical Bayes (Reardon et al., 2021).⁴² Pairing these measures with 1935-1940 HOLC maps allows us to look at how HOLC redlining maps from over eight decades ago are associated with the current state of educational opportunity and student learning rates today.

2.4 Analysis Samples

We use an approach motivated by Hoffman, Shandas, and Pendleton (2020), which documents the association of HOLC redlining policies on resident exposure to intra-urban heat.

⁴² As per SEDA, OLS estimates are more appropriate here than E.B. estimates since we use precision weights in our regression models. Regardless, our results are robust to the underlying estimation procedure.

While our papers' methods deviate, the overarching analytic templates are similar.⁴³ In what follows, we provide an overview of how we constructed each school and district-level analysis sample and describe our analytic approach.

2.4.1 School-Level Analysis Samples

A key piece of this paper is the connection between the geospatial HOLC maps and the geospatial NCES public district and school-level data. For the NCES school-level geospatial data, our mapping approach is straightforward. If a modern-day school is contained within a historic HOLC A-D polygon based on its latitude and longitude point location, it is assigned that historic HOLC A-D polygon grade. If not, no grade is assigned, and the school is dropped from the sample. Thus, at most, a school is assigned one HOLC A-D grade.⁴⁴ This approach, applied to all 2018-19 U.S. public primary and secondary schools, links $s = 10,013$ public schools to a unique HOLC A-D HOLC grade. We merge this data with the NCES 2018-19 school-level data to create our primary cross-sectional "Student Racial Diversity" analysis sample. After accounting for missing data, the "Student Racial Diversity" analysis sample has $s = 9,709$ U.S. public primary and secondary schools in $d = 1,955$ unique districts (Table 2.1).⁴⁵ This represents just over 10% of all active 2018-2019 U.S. public primary and secondary schools. We also link these schools to their late 1980s, 1990s, and 2000s school demographics for our subsequent longitudinal analysis. Doing so creates a "Student Racial Diversity" longitudinal analysis sample

⁴³ Hoffman, Shandas, and Pendleton (2020) map intra-urban land surface temperature anomalies to HOLC ratings, whereas this paper maps school and district outcomes to HOLC ratings. In addition, instead of calculating a "delta" variable that consists of demeaned HOLC A-D averages within a given city by their respective city-wide average, we use CBSA fixed effects.

⁴⁴ Specifically, we execute a "one to many" join via the ArcGIS "completely contains" spatial relationship option. This option matches features from disparate data only if a target layer features from one dataset completely contain join-layer features from another. HOLC A-D ratings and public primary and secondary schools are the target and join layers for this exercise, respectively.

⁴⁵ The initial merge with 2018-19 NCES school-level data matches 98.5% of schools. The remaining unmatched 1.5% of schools are subsequently matched with NCES 2017-18 data. After matching 2018-19 orphans with 2017-18 NCES schools, only one school remains. This school is dropped from the data, giving $s = 10,012$. Finally, two schools mapped to the E HOLC rating, where E represents "other," were also dropped, giving us a total of $s = 10,010$ observations for the school-level diversity analysis sample. After accounting for missing school demographic data across our outcome variables (i.e., Percent Black, White, Non-White, Simpson's Diversity Index), we are left with a total of $s = 9,709$.

with $s = 4,677$ schools in $d = 590$ districts covering 1989-2019.

Table 2.1: Summary of Analysis Samples by Outcome Groups

	Fiscal		Student Racial Diversity		Student Performance
	Cross-Sectional	Time Series	Cross-Sectional	Time Series	Cross-Sectional
Outcomes	Per-Pupil Revenue – Total, Local, State, Federal		% Black, White, Non-White, and Simpson’s Diversity Index		SEDA Outcomes
Level	District & School		School		School
Year	2017-2018	1989-2018	2018-2019	1989-2018	2009-2018
Schools	N/A		9,709	4,677	5,124
Districts	1,760	1,109	1,955	590	1,006
CBSAs	144	143	141	118	137
<i>District Matches</i>					
F33	1,760	1,109	1,283	581	891
Racial Diversity	1,283	581	1,955	590	1,001
SEDA	891	N/A	1,001	N/A	1,006

Notes: For a more detailed breakdown of outcome group samples please refer to Appendix Tables B.1-B.3. These tables cover shares of schools, districts, and CBSAs included in the study and a variety of characteristics (e.g., urban locale, demographics) of the units represented in our study. The time series sample is not applicable for the student performance outcome group. The count of schools is not applicable for “Fiscal” outcomes since unit of analysis is at the district-level.

We follow the same exercise described above to construct the school-level “Student Performance” analysis sample. However, instead of linking the HOLC A-D dataset to NCES 2018-19 demographic data, we now link it to SEDA SY2008-09 to SY2017-18 student performance data. Doing so reduces the sample for the school-level performance analysis by just over one quarter, from $s = 10,013$ to $s = 7,303$ U.S. public primary and secondary schools mapped to unique HOLC A-D grades. Nearly all unmatched schools in the SEDA dataset to the HOLC A-D dataset are due to grade-level mismatches (i.e., SEDA data only covers grades 3-8). Some unmatched schools are also due to school year coverage (i.e., SEDA data only covers schools up to Spring 2018). As a result, all grade levels K-2 and 9-12 and any new public 3-8

grade level schools created after 2018 are excluded from this sample. Finally, the sample drops to $s = 5,124$ schools across $d = 1,006$ districts after removing matched schools with missing SEDA student performance data. The “Student Performance” analysis sample, except for a small school set due to missing student enrollment data, perfectly overlaps with the more extensive “Student Racial Diversity” sample.⁴⁶

While the “Student Performance” and “Student Racial Diversity” analysis samples differ notably in size, on average, they share similar student demographics and other school characteristics. For example, Black students makeup about 30 percent of the student population in each sample, double that of the national average for all U.S. public schools. Hispanic students are also consistently oversampled in “Student Performance” and “Student Racial Diversity” samples relative to the U.S. national average, making up around 37 percent of the student population relative to 27 percent. In contrast, White students are underrepresented, making up just over 22 percent of student populations in both samples. At the regional level, the samples have very similar Black, Hispanic, and White splits across each U.S. Census Bureau region, with consistently oversampling of Black and Hispanic students and under-sampling of White students (Table B.1-B.2). Unsurprisingly, since HOLC A-D maps predominantly targeted larger urban areas, schools in both samples are also almost exclusively located in urban (75 percent) or suburban (20 percent) locales, with less than 5 percent in rural areas or towns. On average, roughly 55 percent of all U.S. public schools are in urban (27 percent) and suburban locales (32 percent), indicating our HOLC A-D samples contain a disproportionate number of these locales relative to national averages (Table B.1-B.2). Finally, while the “Student Racial Diversity”

⁴⁶ The “Student Performance” sample has five districts and seven schools that the “Student Diversity” sample does not. This is due solely to schools in the “Student Diversity” sample being dropped due to missing student enrollment and demographic data, not because of failed NCES school matches.

sample captures a few additional CBSAs compared to the “Student Performance” sample, both have very similar coverage across all U.S. cities included in the sample.⁴⁷ For a more thorough overview of the school-level analysis samples, please refer to Appendix B.

2.4.2 District-Level Analysis Sample

For the NCES district-level geospatial data, our approach is more nuanced. Unlike school point locations, district boundaries often span large metropolitan areas, or in some cases, entire counties, and as a result, at times envelope multiple HOLC A-D polygons. Thus, while we could use a one-to-one mapping procedure based on school point locations for schools, we use a one-to-many for districts. We first overlay all 2018-19 U.S. public school district boundaries with 1935-1940 HOLC A-D maps. If a HOLC A-D polygon overlaps, even in part, with a district boundary, we link it to that overlapping district and calculate the intersecting area of the HOLC A-D polygon. Any district that does not intersect with at least one HOLC A-D polygon is dropped from the sample. Once this mapping exercise is complete, we have a dataset consisting of (1) U.S. public school districts, (2) HOLC A-D polygons that overlapped, entirely or in part, with said U.S. public school districts, and (3) HOLC A-D polygon areas, in square miles, based on the intersecting area between a HOLC A-D polygon and a district boundary.

We use this data to calculate the HOLC A-D weighted average for each public school district, where weights are derived from the overlapping areas (in square miles) of the HOLC polygons and the district boundaries. The weighted average formula can be viewed below:

$$\overline{HOLC}_{LEA} = \frac{1 * \sum_a^A Area_{a \cap LEA} + 2 * \sum_b^B Area_{b \cap LEA} + 3 * \sum_c^C Area_{c \cap LEA} + 4 * \sum_d^D Area_{d \cap LEA}}{\sum_a^A Area_{a \cap LEA} + \sum_b^B Area_{b \cap LEA} + \sum_c^C Area_{c \cap LEA} + \sum_d^D Area_{d \cap LEA}}$$

⁴⁷ Importantly, the distribution of HOLC A-D grades remains stable across both school-level diversity and SEDA samples, both nationwide and by CBSA. HOLC A-D distributions are nearly identical at the national level, varying only by a few percentage points. Comparing samples by CBSA, we see almost the same set of CBSAs represented across the school-level diversity and SEDA samples. In addition, while there is a bit more variation in HOLC A-D distributions across samples at the CBSA level, these are generally driven by CBSAs with already low district counts in the “School Diversity” analysis sample.

As expected, for each district, the weights taken together sum to one, where (1) the denominator is the total sum of all overlapping HOLC-district areas in a district and (2) the numerator is the sum of the overlapping HOLC-district areas for each HOLC A-D grade multiplied by the numeric of that HOLC grade. This procedure produces a continuous set ranging from one to four that we then discretize and link directly to a singular HOLC A-D grade which we use later in our analysis of district-level finances.

Now, each public school district included in the NCES 2018-19 geospatial data now has a HOLC A-D grade assigned to it; however, not all public-school districts have polygon boundaries (e.g., charter schools). We use NCES 2018-19 school district geospatial data with district latitude and longitude point locations to capture these omitted districts. Following a similar procedure outlined above for our school-level point location mapping to HOLC A-D grades, we obtain HOLC A-D grades for each previously omitted district point location. Next, we append the polygon district datasets and point location district datasets, each with their respective HOLC grade mappings, to create our district to HOLC grade mapping dataset.⁴⁸ After merging this dataset with the 2017-18 NCES district-level F-33 fiscal data and removing districts with missing F-33 finance data, we get our final cross-sectional “Fiscal” district-level analysis sample of $d = 1,760$ unique districts, which captures just over 10% of all U.S. public school districts. Nearly all unmatched NCES districts between the HOLC grade mappings dataset and 2017-18 F33 finance data are due to a mismatch in school year coverage (i.e., F-33 data only covers the 2017-18 school year).

Like the school-level analysis samples described above, the “Fiscal” analysis sample

⁴⁸ Given the natural overlap between the NCES 2018-19 district polygon boundary dataset and the NCES 2018-19 district point location dataset, duplicates from the point location dataset were removed once appended.

oversamples districts from urban and suburban areas and those that educate larger shares of Black and Hispanic students relative to the U.S. national average (Table B.3). This is once again true both overall and by U.S. Census Bureau region. The overlap between the "Fiscal" and school-level analysis samples described above is considerable. Around 80 percent of schools in the "Student Racial Diversity" analysis sample are captured by those districts in the "Fiscal" sample, and just over 90 percent of schools in the "Student Performance" analysis sample. We provide a more thorough overview of analysis samples in the Appendix B (Table B.1-B.3). In addition, to view examples of HOLC A-D maps overlaid with NCES district boundaries, please refer to Figure B.2 and Figure B.3 in Appendix B, where we provide NCES district boundary maps superimposed on 1935-1940 HOLC maps for a small selection of CBSAs in our samples (i.e., Cleveland Municipal School District, Los Angeles Unified School District).

2.5 Empirical Strategy

Across our outcomes, our primary empirical strategy is inspired by a difference of means approach taken by Hoffman, Shandas, and Pendleton (2020), which documents the association of HOLC redlining policies with resident exposure to intra-urban heat.⁴⁹ While similar, we base our approach on a CBSA fixed effects regression model with HOLC grade indicators, where we are interested in determining whether, if at all, school and district outcomes for those located in once redlined HOLC D neighborhoods differ from those found in historically HOLC A, B, or C neighborhoods.⁵⁰ This approach allows us to quantify, for a given urban region, how much better or worse off schools or districts with a particular HOLC grade assignment (e.g., D) are from the

⁴⁹ Hoffman, Shandas, and Pendleton (2020) map intra-urban land surface temperature anomalies to HOLC ratings, whereas this paper maps school and district outcomes to HOLC ratings. In addition, instead of calculating a "delta" variable that consists of demeaned HOLC A-D averages within a given city by their respective city-wide average, we use CBSA fixed effects.

⁵⁰ Recall, for districts, this is not quite the case. Districts are assigned a composite HOLC A-D grade based on the weighted average of HOLC A-D polygons they contain, where weights are based on the area, in square miles, of overlapping district boundaries and HOLC A-D polygons. See the above section for details.

set of all other HOLC schools or districts (e.g., A, B, C) in that urban region. Given the heterogeneity in state funding mechanisms across the U.S. (which have implications for school and district funding levels and distributions locally) and nationwide differences in demographic compositions, a within CBSA comparison of our outcomes by HOLC A-D grades is preferred to an unadjusted one. Our use of CBSA fixed effects also ensures that differences in any time-invariant covariates at the city-level are differenced out in our model, which we present below:

$$Y_{ij} = \alpha + \beta_1 HOLC_{A_{ij}} + \beta_2 HOLC_{B_{ij}} + \beta_3 HOLC_{C_{ij}} + \gamma_j + \epsilon_i$$

where i indexes schools or districts, j indexes CBSAs, and γ_j is the CBSA fixed effect. For schools, $HOLC_A$, $HOLC_B$, and $HOLC_C$ are the estimated average differences in the outcome Y_{ij} between schools in what once were HOLC D graded neighborhoods and their contemporaries located in what once were HOLC A, B, or C HOLC graded neighborhoods, respectively. For districts, the interpretation is similar but with the caveat that HOLC A-D grades represent HOLC A-D weighted averages, which could but often do not refer to a single HOLC A-D polygon, but instead a collection of them.

The key identifying assumption of our approach is that within-CBSA selection into a HOLC A-D grade assignment for schools or districts is uncorrelated with (1) uncontrolled for time-variant determinants of our outcomes and (2) unobservable determinants of our outcomes. Threats of validity arise if $E(\epsilon_i | HOLC_A, HOLC_B, HOLC_C, \gamma_j) \neq 0$, which is true in the presence of school or district nonrandom “selection” into HOLC A-D grade assignment such that said assignment correlates with unobservable determinants of our outcome variables. While modern-day schools and districts were not historically assigned HOLC A-D grades and, in many cases, did not even exist yet, selection bias remains a threat if past HOLC A-D neighborhood grade

assignments captured pre-existing inequality between neighborhoods and is a determinant of our modern-day educational outcomes. For example, suppose the original 1935-1940 HOLC A-D grade assignment reflected already existing neighborhood variation (e.g., homeownership rates, home values, access to public services, crime) that favored higher-rated HOLC security ratings (i.e., A, B) relative to lower ones (i.e., C, D). In this case, these differences may be driving, in part, the differences in educational outcomes by HOLC A-D grades we see today. If true, our results are not solely attributable to the HOLC A-D security ratings themselves but differences in other pre-existing neighborhood characteristics. Notably, Aaronson et al. (2020) find evidence that this is indeed the case while also presenting causal evidence in Aaronson et al. (2021). Given these concerns, we avoid framing our results as causal and instead focus on describing the historical long-term correlational relationships between HOLC A-D grades and their conceptions of neighborhood quality and modern-day educational outcomes.

2.6 Results and Discussion

To facilitate a more meaningful discussion of our results, we partition them by the three overarching outcomes addressed in this paper: (1) district and school-level finance, (2) school-level racial diversity, and (3) school-level student performance.⁵¹ In each section, we provide and discuss the regression results from our CBSA FE estimation model described in the “Empirical Strategy” section above. For the district finance and racial diversity outcomes, we also provide a time series analysis that looks at how, if at all, the relationship between said outcomes and HOLC A-D security ratings changed over the past three decades.⁵²

2.6.1 District and School Finance

⁵¹ There is notably a mismatch in school years between the district and school spending data and SEDA data. However, the consequences of having test score data on 4th graders in 2009 and spending data for schools and districts nearly a decade later are immaterial. We find similar results using the SY2009-10 F-33 finance data (adjusted to USD 2018) in the magnitude of estimates and statistical significance.

⁵² Our time series analysis does not include SEDA school-level student achievement outcomes or NERDS school-level finance outcomes, as neither contains data from the 1980s or 1990s.

2.6.1.1 Current Outcomes – Districts

In Table 2.2, we provide nationwide results of our district-finance outcomes. We see that, on average, total per-pupil revenues are almost always lower for districts mapped to HOLC D grades than those mapped to A, B, and C grades. Nationwide, districts mapped to HOLC D grades have, on average, \$14,402 per-pupil total revenues with differences greatest between A vs. D districts (\$1,546, SE = \$822) and least between B vs. D districts (\$992, SE = \$536). All these differences are statistically significant. At the federal level, districts mapped to HOLC D grades have, on average, \$1,201 in per-pupil federal revenues, with A (-\$722, S.E. = \$139) and B (-\$391, S.E. = \$87) districts being outstripped in federal funding by D districts, and C districts (\$253, S.E. = \$76) just edging out D districts. Once again, all these differences are statistically significant. Moving to the state revenues, we find that D districts have, on average, \$7,960 in per-pupil state funding, with differences largest between A vs. D districts (\$3,670, S.E. = \$752) and smallest between C vs. D districts (-\$461, S.E. = \$495). All but the C vs. D average comparisons are statistically significant.

Table 2.2: District Per-Pupil Revenues and 1935-1940 HOLC A-D, SY2017-18 (USD 2018)

Nationwide	District PPR Total	District PPR Federal	District PPR State	District PPR Local
A	\$1,546* [\$822]	-\$722*** [\$139]	-\$3,670*** [\$752]	\$5,937*** [\$1,228]
B	\$992* [\$536]	-\$391*** [\$87]	-\$3,210*** [\$776]	\$4,592*** [\$892]
C	\$1,504*** [\$514]	\$253*** [\$76]	-\$461 [\$495]	\$1,713*** [\$468]
Constant	\$14,402*** [\$396]	\$1,201*** [\$62]	\$7,960*** [\$438]	\$5,240*** [\$425]
FE	Y	Y	Y	Y
R2	0.64	0.32	0.52	0.49
F	4.08	20.16	8.75	11.56
N	1,760	1,760	1,760	1,760

Notes: Cluster-robust standard errors are in parentheses with clustering done at the city-level. For each model we regress the outcome on HOLC grade indicators with the “D” security rating as the reference category. No controls are included. All regressions are weighted by 2017-18 total students. All models use city-level fixed effects to account for any differences that are fixed at the local level and that differ between CBSAs. All dollars denominated in USD

2018. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Finally, at the local level, we see that across the U.S., districts mapped to HOLC D grades have, on average, \$5,240 in per-pupil local revenues, which is less than those districts mapped to HOLC A, B, and C grades by -\$5,937, -\$4,592, and -\$1,713, respectively. All differences are statistically significant. A cohesive narrative emerges by combining these results with our above state, federal and total results. Districts mapped to HOLC D grades have significantly less per-pupil local funding than higher-rated HOLC A, B, and C districts in the sample. This gap in local funding is abated by redistributive federal and state funds that favor D mapped districts relative to A, B, and C districts; however, non-local educational funding is not enough to overcome the sizeable initial gap driven by differences at the local level.

2.6.1.2 Current Outcomes – Schools

In contrast to the district finance results, at the school level, we find that, on average, schools mapped to HOLC D schools have more per-pupil total expenditures than schools mapped to A, B, and C HOLC grades (Table 2.3). Differences between D schools and their higher HOLC A, B, and C schools are monotonically decreasing in HOLC security rating, such that gaps in per-pupil total expenditures are widest between schools mapped to A and D HOLC grades (-\$1,539, S.E. = \$852) and smallest between schools mapped to C and D HOLC grades (-\$882, S.E. = \$334). These differences are all statistically significant.

While the 2018-19 NERDS data affords us school-level finance data, it is more coarsely disaggregated relative to the NCES F-33 district finance data, limiting our ability to deconstruct the underlying variation per-pupil total expenditures into cleanly delineated federal, state, and local buckets. That said, we can isolate per-pupil federal expenditures and investigate the combined sum of state and local expenditures. Like the district-level finance results, we find that,

on average, schools mapped to D schools have more per-pupil federal expenditures than A, B, and C schools. These differences are all statistically significant and once again are largest between A and D schools (-\$602, S.E. = \$39) and smallest between A and C schools (\$141, S.E. = \$31).

Table 2.3: School Per-Pupil Expenditures and 1935-1940 HOLC A-D, SY2018-19 (USD 2018)

Nationwide	School PPE Total	School PPE Federal	School PPE State & Local
A	-\$1,539* [\$852]	-\$602*** [\$39]	-\$1,131 [\$860]
B	-\$1,313** [\$581]	-\$351*** [\$29]	-\$1,010* [\$590]
C	-\$882*** [\$334]	-\$141*** [\$31]	-\$745** [\$323]
Constant	\$16,890*** [\$304]	\$1,379*** [\$18]	\$15,658*** [\$297]
FE	Y	Y	Y
R2	0.39	0.14	0.41
F	3.67	96.49	4.15
N	8,573	8,573	8,573

Notes: Cluster-robust standard errors are in parentheses with clustering done at the city-level. For each model we regress the outcome on HOLC grade indicators with the “D” security rating as the reference category. No controls are included. All regressions are weighted by 2017-18 total students. All models use city-level fixed effects to account for any differences that are fixed at the local level and that differ between CBSAs. All dollars denominated in USD 2018. *** p< 0.01, ** p< 0.05, * p< 0.1.

For the combined state and local per-pupil expenditure outcome, we observe similar patterns in the average differences between D vs. A, B, C schools as we do for total and federal per-pupil expenditures; however, not all differences are statistically significant, namely, A vs. D differences (-\$1,131, S.E. = \$860). That said, average differences between D vs. B (-\$1,010, SE = \$590) and C (-\$745, SE = \$323) schools are statistically significant. Negative per-pupil state and local averages mean one of two things – per-pupil local expenditures outpace per-pupil state expenditures, or per-pupil state expenditures outpace per-pupil local ones. Incorporating what we know from the F-33 district finance results, where per-pupil state revenues offset large and

statistically significant differences in local per-pupil revenues that favor those districts mapped to higher HOLC A, B, or C grades relative to those mapped to redlined HOLC D grades, we argue the latter is most realistic.

2.6.1.3 Reconciliation of District and School Finance Results

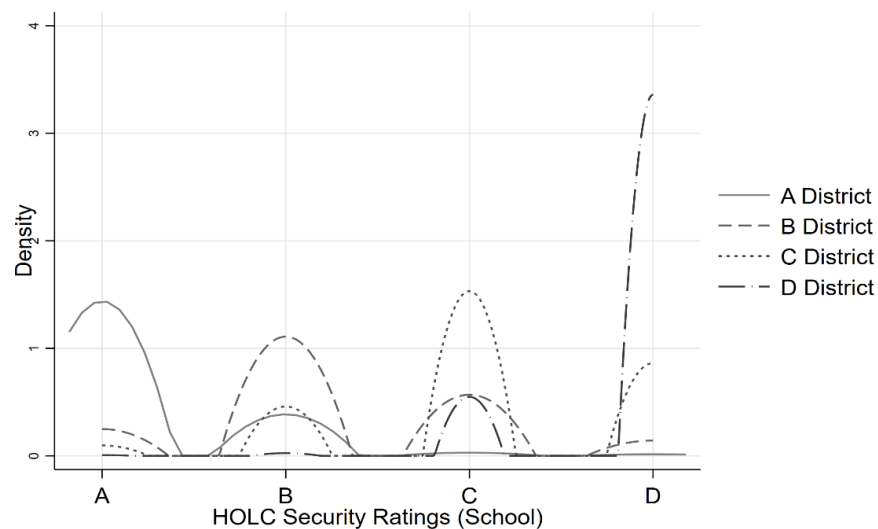
In this section, we attempt to reconcile the differences in the district and school-level results by purposing a few hypotheses and discussing their likelihood. Starting with the first, the differences in the district and school-level results may simply be an artifact of the outcomes chosen since the district-level analysis uses per-pupil revenues and the school-level analysis uses per-pupil expenditures.⁵³ If notable variation exists between funding allocations versus reported spending in non-uniform ways by HOLC A-D security rating, one would expect to see these differences play out in our regression results. When we substitute district per-pupil total expenditures for district per-pupil total revenues, our district-level results are nearly indistinguishable from another, indicating that measurement differences are an unlikely driver of our varying school and district finance results.

Another explanation is the notable differences in the school and district mapping strategies we use to assign HOLC grades. Recall, for districts, HOLC A-D mapping is based on the weighted average of the portions of HOLC A-D polygons contained within that district; however, for schools, HOLC A-D mapping is based on a one-to-one match of an individual school to a HOLC grade. By construction, a district with a given HOLC grade assignment will contain the largest share of that HOLC polygon type, in square miles, relative to other HOLC polygons; however, this does not preclude other HOLC graded polygons from being present in that district. If schools are uniformly distributed throughout a district, we would expect within-

⁵³ We use NCES F-33 revenues because it allows us to explore total revenues and the constituent local, state, and federal revenues that roll up to it. NCES F-33 expenditures do not provide this level of detail.

district school mappings to reflect the within-district distributions of HOLC polygons, but if schools cluster in HOLC polygons different from the district HOLC A-D mapping, the distribution of HOLC school mappings will not accurately reflect the HOLC polygons that make up that district.⁵⁴

Figure 2.2: HOLC A-D School Distributions by HOLC A-D District Grades



To check this, we construct the underlying HOLC A-D school distributions of each HOLC A-D district. In Figure 2.2, a few notable patterns emerge. First, higher (lower) rated HOLC districts have larger shares of higher (lower) rated HOLC schools, with the greatest being that of the HOLC A-D school rating congruent with the HOLC A-D district rating. For example, districts mapped to HOLC A security ratings have the greatest share of HOLC A schools in that district. Alternatively, districts mapped to HOLC D security ratings have the greatest share of HOLC D schools in that district. These findings hold for B and C districts as well. Second,

⁵⁴ For example, suppose a district is composed of mostly HOLC A polygons but has several small HOLC B-D polygons where schools cluster. Based on our district and school mapping strategies, the district will likely be assigned a HOLC A grade, whereas schools will be assigned B-D grades. A scenario such as this could explain our district and school-level finance results.

shares of HOLC A-D schools in HOLC A-D districts follow a strict rank order, such that HOLC D districts have the largest share of D mapped schools, second of C, third of B, and the least of A. Conversely, HOLC A districts have the largest share of A mapped schools, second of B, third of C, and the least of D. This pattern holds for B and C districts. These results validate our school and district mapping strategies and suggest schools are not clustering in polygons discordant from the mapped district HOLC A-D grade.

A final explanation, and one we believe is most likely, is that districts are simply allocating resources to those most in need (e.g., low-income students), which are those schools located today in what were once HOLC D neighborhoods. We check this by regressing the percentage of free and reduced-price lunch students (FRPL) in a school on HOLC grade indicators using the same CBSA FE model as before. We find that nationally, schools mapped to HOLC D grades have, on average, 79.3 percent of students qualifying for free and reduced-price lunch, where percentage point differences are greatest between A vs. D schools (-0.37 , S.E. = 0.02) and smallest between A vs. C schools (-0.06 , S.E. = 0.01). We observe a similar pattern across all regions. All differences between D vs. A, B, and C schools are statistically significant nationwide and by region. Combined with the district and school finance results, these findings suggest that districts, regardless of their A-D assigned grade, systematically target and allocate more funding to schools located in historically HOLC D neighborhoods because those schools today serve the largest shares of students from low-income households. These results can be seen as an extension of recent research that finds redlined HOLC D neighborhoods are worse off, both in terms of homeownership rates and home values, relative to their higher-rated peers (Aaronson et al., 2020) and provide some of the first evidence of secondary ripple effects that stem from the adverse impacts HOLC maps had on neighborhood quality and development.

Finally, considering our school-level findings, what level of importance should one place on the district-level results? We argue these results are still pertinent, as they represent an upper limit on how much a district can allocate to their most-in-need schools (i.e., D schools) while still maintaining adequate funding levels for all other schools they serve. Given that we find that districts' HOLC A-D mapping is predictive of their total per-pupil funding, such that districts mapped to higher-rated HOLC grades have, on average, more per-pupil total funding than those mapped to lower-rated HOLC grades, one might expect lower-rated HOLC schools in higher-rated HOLC districts to receive and spend more money relative to those in lower-rated HOLC districts. We test this theory by looking at kernel density plots (Figure B.4) and regressing per-pupil total funding HOLC indicators for a subset of our analysis sample containing only D schools.⁵⁵ While the kernel density analysis does provide evidence to support this theory, the CBSA FE regression results do not, which is likely an artifact of our smaller subsample and less within-CBSA variation in HOLC A-D districts. Zooming out to the state and regional levels, our results from the state F.E. and region F.E. support this narrative and align with the distributional patterns in Figure 3.⁵⁶

This evidence suggests that districts with more resources in our sample (i.e., A, B, C), relative to those with less (i.e., D), distribute more to their most-in-need schools (i.e., D schools). Thus, while D schools have the largest per-pupil total expenditures of all HOLC A-D schools regardless of HOLC A-D district type, our results show that variation in district resources yields variation in D school resources that favor higher-rated HOLC districts relative to lower-graded

⁵⁵ While there is considerable overlap between the F-33 district finance sample and the NERDS 2018-19 school-level sample, up to this point, our school and district finance samples are not perfectly congruous; that is, every district in the NCES F-33 sample is not in the NERDS 2018-19 sample, and vice versa. We link our NERDS school and NCES F33 district-level samples to facilitate a complete district-school comparison. Overall, these datasets share 95% of the same districts; however, each is plagued with missing outcome data, which after removing, leaves us with $s = 6,670$ schools that roll up to $d = 1,025$ unique districts. These samples represent just under 60% and 70% of the original F33 district-level analysis and NERDS school-level analysis samples, respectively.

⁵⁶ Results are statistically significant when clustering at the city or state level. Results become statistically insignificant ($p = 0.15$) for D vs. B when clustering at the regional level but remain statistically significant for D vs. C differences.

HOLC districts. Equalization of district per-pupil funding across HOLC A-D district ratings through targeted local, state, and federal programs could help equalize funding at the school level for those located today in historically redlined HOLC D neighborhoods that serve disproportionately low-income students.

2.6.1.4 Longitudinal Outcomes

In this section, we expand our analysis to look at how, if at all, the relationship between HOLC A-D grades and the finance outcomes changed over time. This exercise reduces our original cross-sectional district-level analysis sample ($d = 1,760$) by a third ($d_p = 1,109$). To ensure those districts that remain in our panel dataset are representative of those contained in the original 2017-18 cross-sectional sample, we perform a few robustness checks across samples. While the distribution of HOLC A-D grades across districts is slightly different between the 2017-18 cross-sectional and panel samples, we find no evidence suggesting that average outcomes by A-D HOLC grades vary (Figure B.5, Table B.4).⁵⁷ With these details in mind, let us discuss our results.

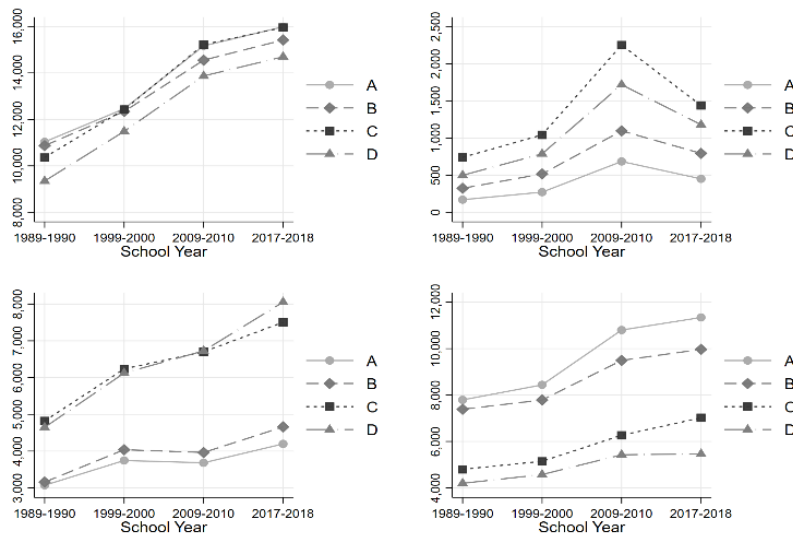
First, to better understand how district funding changed over time by HOLC A-D grade, in Figure 2.3, we show A-D averages across time for each educational finance outcome, weighted by district enrollment. Each outcome is adjusted for inflation and denominated in 2018 USD. Starting with average per-pupil total revenue, one can see near-parallel lines across time with only marginal differences in slopes by HOLC A-D grade. In addition, compound annual growth rates (CAGRs), calculated from 1989-90 to 2017-18, differ little by HOLC security rating and hover around four percent regardless of HOLC grade. While equality in growth rates

⁵⁷ For each F-33 finance outcome, we consider in this paper, we fail to reject the null hypothesis of equality of common coefficients across models using (1) the full cross-sectional 2017-18 sample, and (2) the partial panel 2017-18 sample, where each regression model consists of a given finance outcome regressed on an A-D HOLC indicator variable. This result suggests no statistically significant differences between the weighted average A-D HOLC grades across these two samples for our F-33 educational finance variables.

across time is encouraging, it is less so after considering level differences in average per-pupil total revenue between D districts and their A, B, and C counterparts (Figure 2.4, Table 2.4). Here we see an initial and subsequently increasing gap in per-pupil total revenue that favors districts located in historically non-redlined neighborhoods (i.e., A, B, C). Thus, while growth rates are similar across HOLC A-D security ratings, per-pupil funding gaps are not. This result is a direct consequence of initial per-pupil total revenue gaps by HOLC A-D grade in the late 1980s paired with near-identical A-D growth rates across time. Findings for per-pupil local revenue mirror those for per-pupil total revenue but are even more pronounced, with larger initial funding gaps in the late 1980s and smaller growth rates.

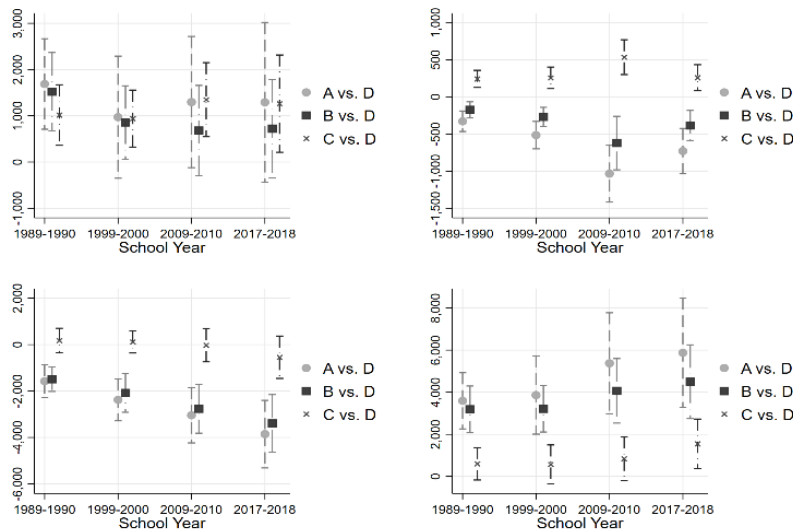
In contrast to per-pupil total and local revenue trends, per-pupil state and federal revenues favor those districts with lower HOLC security ratings (i.e., C, D) than those with higher (i.e., A, B). Like our 2017-18 cross-sectional findings, state and federal redistributive policies appear to benefit those districts most in need and have consistently done so over the three decades we consider here. For per-pupil state revenues, funding gaps in the late 1980s favor C and D districts relative to A and B districts. Growth rates lead to C and D district convergence, with B districts remaining mostly in parallel; however, there is a distinct increase in the gaps between B, C, and D districts and their highest-rated A counterparts. These results are further confirmed in Figure 2.4, where there is limited variation across years for D vs. C and D vs. B comparisons but a clear monotonic increasing relationship for the D vs. A group. Finally, per-pupil federal revenues match much of the findings for per-pupil state revenues, with a few notable differences, including the large positive spike in per-pupil federal funding for SY2009-10 consistent with the surge in education funding from the American Recovery and Reinvestment Act of 2009 (ARRA) (U.S. Department of Education, 2009).

Figure 2.3: HOLC Averages, Finance Outcomes, 1989-2018 (USD 2018)



Notes: [top left] Per-Pupil Total Revenue; [top right] Per-Pupil Federal Revenue; [bottom left] Per-Pupil State Revenue; [bottom right] Per-Pupil Local Revenue. All values represented in USD 2018. Weighted averages are from a regression of a given finance outcome on HOLC A-C indicators with student enrollment as an analytic weight. Standard errors are heteroskedasticity-consistent and clustered at the city-level. All dollars denominated in USD 2018.

Figure 2.4: HOLC Pairwise Comparisons, Finance Outcomes, 1989-2018 (USD 2018)



Notes: [top left] Per-Pupil Total Revenue; [top right] Per-Pupil Federal Revenue; [bottom left] Per-Pupil State Revenue; [bottom right] Per-Pupil Local Revenue. All values represented in USD 2018. Weighted averages are from a regression of a given finance outcome on HOLC A-C indicators with student enrollment as an analytic weight. Brackets represent 95% CIs. Standard errors are heteroskedasticity-consistent and clustered at the city-level. All dollars denominated in USD 2018.

Table 2.4: Margins & Differences Over Time, Finance Outcomes, 1989-2018 (USD 2018)

		Margins				D vs. A, B, and C		
		A	B	C	D	A vs. D	B vs. D	C vs. D
District PPR – Total (USD 2018)	1989-90	\$11,035 [\$439]	\$10,870 [\$445]	\$10,364 [\$95]	\$9,347 [\$265]	\$1,688*** [\$495]	\$1,524*** [\$427]	\$1,017*** [\$331]
	1999-00	\$12,457 [\$611]	\$12,343 [\$380]	\$12,426 [\$83]	\$11,486 [\$259]	\$971 [\$668]	\$857** [\$401]	\$940*** [\$311]
	2009-10	\$15,168 [\$693]	\$14,552 [\$579]	\$15,221 [\$136]	\$13,871 [\$298]	\$1,297* [\$718]	\$681 [\$493]	\$1,349*** [\$404]
	2017-18	\$15,989 [\$748]	\$15,419 [\$630]	\$15,961 [\$156]	\$14,698 [\$411]	\$1,291 [\$873]	\$721 [\$539]	\$1,263** [\$532]
District PPR – Federal (USD 2018)	1989-90	\$170 [\$58]	\$324 [\$43]	\$740 [\$11]	\$499 [\$50]	-\$329*** [\$70]	-\$175*** [\$55]	\$241*** [\$59]
	1999-00	\$272 [\$85]	\$516 [\$60]	\$1,042 [\$16]	\$786 [\$61]	-\$515*** [\$93]	-\$270*** [\$65]	\$256*** [\$73]
	2009-10	\$684 [\$169]	\$1,096 [\$143]	\$2,251 [\$32]	\$1,717 [\$104]	-\$1,033*** [\$194]	-\$622*** [\$182]	\$534*** [\$119]
	2017-18	\$450 [\$143]	\$793 [\$85]	\$1,437 [\$22]	\$1,179 [\$74]	-\$729*** [\$154]	-\$386*** [\$103]	\$258*** [\$88]
District PPR – State (USD 2018)	1989-90	\$3,072 [\$289]	\$3,159 [\$181]	\$4,822 [\$46]	\$4,648 [\$233]	-\$1,575*** [\$359]	-\$1,489*** [\$268]	\$174 [\$267]
	1999-00	\$3,746 [\$413]	\$4,039 [\$325]	\$6,237 [\$66]	\$6,123 [\$216]	-\$2,377*** [\$456]	-\$2,084*** [\$419]	\$114 [\$239]
	2009-10	\$3,684 [\$574]	\$3,964 [\$437]	\$6,701 [\$103]	\$6,727 [\$308]	-\$3,043*** [\$601]	-\$2,763*** [\$531]	-\$26 [\$361]
	2017-18	\$4,197 [\$686]	\$4,659 [\$567]	\$7,505 [\$141]	\$8,051 [\$379]	-\$3,854*** [\$735]	-\$3,392*** [\$629]	-\$546 [\$461]
District PPR – Local (USD 2018)	1989-90	\$7,793 [\$621]	\$7,388 [\$556]	\$4,802 [\$116]	\$4,201 [\$316]	\$3,592*** [\$675]	\$3,187*** [\$561]	\$601 [\$390]
	1999-00	\$8,439 [\$890]	\$7,787 [\$505]	\$5,147 [\$116]	\$4,576 [\$392]	\$3,863*** [\$939]	\$3,211*** [\$556]	\$571 [\$469]
	2009-10	\$10,800 [\$1,131]	\$9,492 [\$714]	\$6,269 [\$166]	\$5,427 [\$431]	\$5,374*** [\$1,215]	\$4,066*** [\$775]	\$842 [\$523]
	2017-18	\$11,341 [\$1,146]	\$9,967 [\$713]	\$7,019 [\$165]	\$5,468 [\$520]	\$5,874*** [\$1,312]	\$4,499*** [\$887]	\$1,551** [\$595]

Notes: Cluster-robust standard errors are in parentheses with clustering done at the city-level. For each model we regress the outcome on HOLC grade indicators with the “D” security rating as the reference category. No controls are included. All regressions are weighted by 2017-18 total students. All models use city-level fixed effects to account for any differences that are fixed at the local level and that differ between CBSAs. All dollars denominated in USD 2018. *** p< 0.01, ** p< 0.05, * p< 0.1.

2.6.2 School Racial Diversity

2.6.2.1 Current Outcomes

In Table 2.5, we see that nationwide schools with HOLC D grades relative to those schools with A, B, and C ratings have larger shares of Black and Non-White students in their schools. Overall, for those schools mapped to HOLC D grades in our sample, 36 percent (S.E. = 0.011) of students are Black, with differences greatest between A vs. D schools (-0.14, S.E. = 0.027) and smallest between C vs. D schools (-0.08, S.E. = 0.016). All these results are statistically significant. For percent non-White, we observe identical but inverse patterns to our findings above for percent White. Our final school diversity outcome variable is the Simpson's Diversity Index. Nationwide, the Simpson's Diversity Index monotonically decreases in HOLC A-D grade, such that A schools have, on average, the most diverse student populations and D schools have the least (0.42, S.E. = 0.009). Differences between D schools and their higher HOLC A (0.07, SE = 0.024), B (0.05, SE = 0.019) and C (0.03, SE = 0.011) are all statistically significant.

Table 2.5: School Racial Diversity and 1935-1940 HOLC A-D Grades, 2018-2019

Nationwide	% Black	% White	% Non-White	Simpson's Diversity (1-D)
A	-0.14*** [0.027]	0.3*** [0.022]	-0.3*** [0.022]	0.07*** [0.024]
B	-0.11*** [0.017]	0.18*** [0.013]	-0.18*** [0.013]	0.05*** [0.019]
C	-0.08*** [0.016]	0.06*** [0.01]	-0.06*** [0.01]	0.03*** [0.011]
Constant	0.36*** [0.011]	0.14*** [0.007]	0.86*** [0.007]	0.42*** [0.009]
FE	Y	Y	Y	Y
R2	0.33	0.34	0.34	0.29
F	13.74	86.13	86.13	3.8
N	9,709	9,709	9,709	9,709

Notes: Cluster-robust SEs are in parentheses with clustering at the city-level. For each model we regress the outcome on HOLC grade indicators with the "D" security rating as the reference category. No controls are included. All regressions are weighted by 2018-19 total students. All models use city-level fixed effects to account for any differences that are fixed at the local level and that differ between CBSAs. *** p < 0.01, ** p < 0.05, * p < 0.1.

These results point to a more nuanced narrative than, for example, D schools have predominantly Black students, whereas A and B schools have primarily White. While it is true that schools with higher-rated HOLC grades have far fewer shares of Black students than D schools, they also have, on average, greater diversity. Pairing these findings with our above district and school-finance results shows that those schools mapped to higher HOLC security ratings have more student diversity, smaller shares of Black students, and serve more affluent households. These takeaways are consistent with research on the school and neighborhood diversity in urban areas (e.g., Filardo et al., 2008; Candipan, 2019) and recent research on diversity and family income in U.S. K-12 public schools (U.S. Government Accountability Office, 2016).

What might be driving these results? One explanation is that more money provides more opportunities to enact diversity mandates set by a district or city. While there are many examples of schools and districts announcing diversity mandates, their ability to reach diversity targets may depend on their available resources.⁵⁸ Student recruitment, student assignment plans, and targeted integration strategies require time, staffing, and financial commitments. As such, holding all else equal, one would predict that schools with less funding will have less diversity. However, our results indicate that HOLC D schools have the greatest shares of FRPL students, the most per-pupil expenditures, and the least diversity compared to HOLC A-C schools. These findings do not support the above narrative.

An alternative explanation, and one we think is more probable, is rooted in the understanding that wealthy families are less restricted spatially and less likely to be priced out of

⁵⁸ For example, see the New York State Board of Regents' recent announcement of New York public schools' diversity, equity, and inclusion initiatives (NYSED, 2021). In addition, many charters are adopting "diverse-by-design" approaches where diversity is explicitly stated as a core value or part of the school design (Potter & Quick, 2018).

neighborhoods that act as gateways to high-performing schools. Thus, across the racial spectrum, high-performing districts attract and retain more affluent families that can afford the price of admission, namely a residential property located in the school district. In addition, if diversity mandates are in place, high-performing schools and districts can more easily recruit and retain families from diverse backgrounds relative to their low-performing counterparts. Under this model, enrollment patterns in high-performing districts would be more responsive to changes in diversity goals. In contrast, enrollment patterns in low-performing districts would be less responsive and more reflective of the neighborhood's status quo demographics. What should one expect the status quo demographic composition to be, on average, in communities being served by low-performing schools? Research shows that historical federal housing policies often buttressed patterns of neighborhood segregation by income and race, such that low-income minority families are, by design, spatially concentrated into a select few locales and therefore clustered in neighborhood schools and districts with high rates of student poverty (Katz & Turner, 2008; Turner & Berube, 2009). Given this, we should expect low-performing schools to be less diverse, have higher concentrations of racial minorities, and have more low-income students. Consequently, our results are consistent with this theory and lend further evidence to the long and harmful half-life of historically discriminatory neighborhood policies on educational outcomes.

2.6.2.2 Longitudinal Outcomes

In this section, we expand our analysis to look at how, if at all, the relationship between HOLC A-D grades and the racial diversity outcomes changed over time. Like the educational finance time series analysis, we first create a panel dataset that spans nearly three decades and includes school-level student demographic data from the late 1980s to the late 2010s. This

exercise reduces our original cross-sectional school-level analysis sample ($n = 9,709$) by nearly half ($n_p = 4,677$). To check if those schools that remain in our panel dataset are representative of those in the original 2018-19 cross-sectional sample, we perform a few robustness checks across samples. Overall, the distribution of HOLC A-D grades across schools varies little between the 2018-19 cross-sectional and panel samples. However, we find evidence that average outcomes by A-D HOLC grades differ by sample.⁵⁹ While often small, these differences are almost always statistically significant for A-D HOLC grades across each diversity outcome except for Simpson's Diversity Index, which has only statistically significant differences between samples for the B security rating (Figure B.6, Table B.5). With these details in mind, we discuss the time series results below.

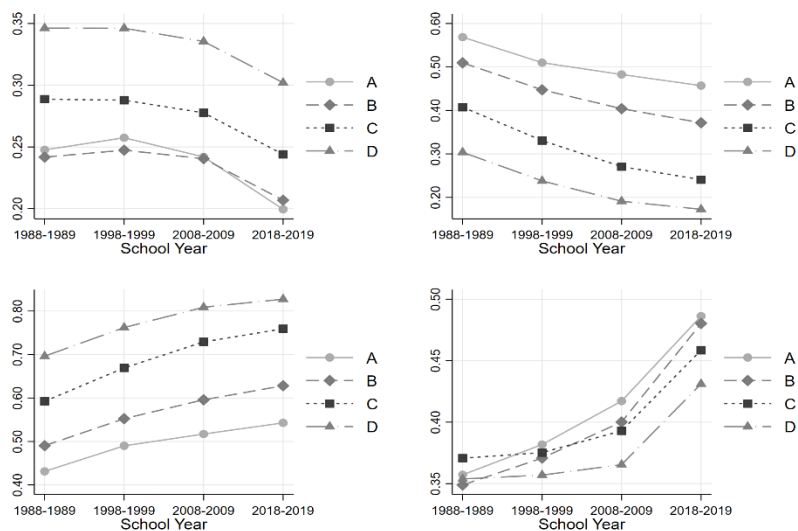
First, to better understand how school student demographics and racial diversity changed over time by HOLC A-D grade, in Figure 2.5 and Figure 2.6, we show A-D averages and differences across time for each racial diversity outcome, weighted by school enrollment. All mean outcomes and differences are captured in Table 2.6. Starting with the percent Black outcome, overall, we see negative downward sloping convex trends from 1988-89 through 2018-19. For those schools located in historically rated A and B neighborhoods, there is a small uptick in the percent Black from 1988-89 to 1998-99, reaching 24.8 and 26.5 percentage points, but this was overwhelmed by negative trends in the two decades that follow, so much so that both groups end up below their original 1988-89 shares. Positive gaps between D vs. A, B, and C schools exhibit an initial downward trend from 1988-89 to 2008-09 and flattens out over the final decade. The gap in average shares of Black students is largest between the D vs. A group beginning at

⁵⁹ For each diversity outcome we consider in this paper, we reject the null hypothesis of equality of common coefficients across models using 1) the full cross-sectional 2018-19 sample, and 2) the partial panel 2018-19 sample, where each regression model consists of a given diversity outcome regressed on an A-D HOLC indicator variable. This result suggests statistically significant differences between the weighted average A-D HOLC grades across these two samples for our diversity outcomes. The exception is the Simpson's Diversity outcome, which fails to reject the null hypothesis of equality of common coefficients for A, C, and D HOLC grades across samples.

13.4 percentage points in 1988-89 and shrinking to 10.3 percentage points by 2018-19. In comparison, the gap between D vs. B and D vs. C schools begins at 10.8 and 7.8 before decreasing by a few percentage points over the following decades to end at 7.4 and 6.5 percentage points, respectively.

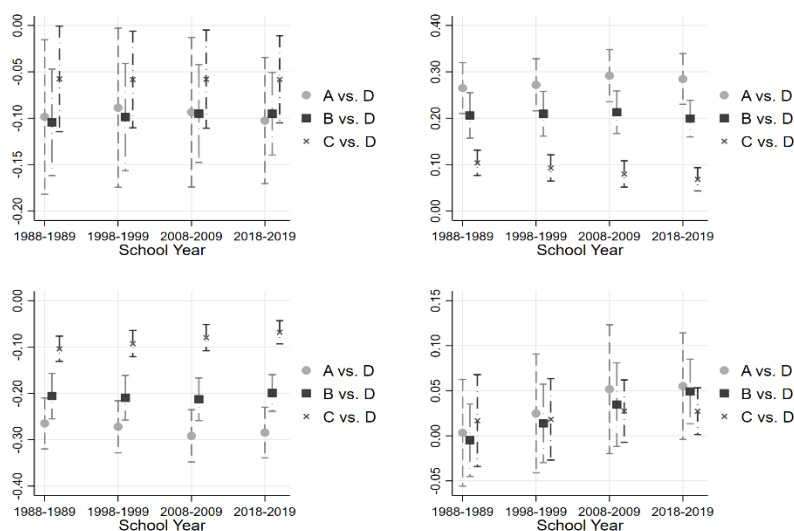
We see parallel downward sloping lines across time for the percent White outcome, with negative gaps between D vs. A, B, and C counterparts remaining mostly flat from 1988-89 through 2018-19. Thus, although student bodies have become less White over time across all HOLC A-D grades in our time series sample, they have done so at similar rates. There is a clear rank order by HOLC A-D grade to the lines in Figure 6, with A schools having, on average, the largest share of White students, B the second, C the third, and D the fourth and smallest. Thus, gaps are largest between A and D schools, with differences of around 35 percentage points. For the percent Non-White outcome, we see the same patterns as the percent White outcome, except in reverse. Finally, for the Simpson's Diversity (1-D) outcome, we see uniform increases across all HOLC A-D grades from 2008-09 to 2018-19, with nearly imperceptible gaps between A and B schools but notable negative differences between D vs. A, B, and C counterparts. However, these patterns are not consistent across time. For example, in 1988-89, average diversity levels hover around 0.35 across all HOLC A-D grades, with gaps near zero between D vs. A, B, and C schools. From this point forward, the diversity index steadily increases for those schools in the highest-rated HOLC neighborhoods (i.e., A, B) while remaining relatively flat for those schools in the lowest-rated HOLC neighborhoods (i.e., C, D). Thus, by 2008-09 there were notable negative gaps in the diversity index between D vs. A, B, and C schools that continue to grow into 2018-19. This pattern is especially true for D vs. A, B comparisons, where previously near-zero gaps in the diversity index in 1988-89 surpass -0.08 in 2018-19.

Figure 2.5: HOLC A-D Averages Over Time, Racial Diversity Outcomes, 1988-2019



Notes: [top left] Percent Black; [top right] Percent White; [bottom left] Percent Non-White; [bottom right] Simpson's Diversity Index (1-D). Weighted averages are from a regression of a given diversity outcome on HOLC A-C indicators with student enrollment used as analytic weights. Standard errors are heteroskedasticity-consistent and clustered at the city-level.

Figure 2.6: HOLC Pairwise Comparisons Over Time, Diversity Outcomes, 1988-2019



Notes: [top left] Percent Black; [top right] Percent White; [bottom left] Percent Non-White; [bottom right] Simpson's Diversity Index (1-D). Weighted averages are from a regression of a given diversity outcome on HOLC A-C indicators with student enrollment used as analytic weights. Brackets represent 95% confidence intervals. Standard errors are heteroskedasticity-consistent and clustered at the city-level.

Table 2.6: Margins & Differences Over Time, Racial Diversity Outcomes, 1988-2019

		Margins				D vs. A, B, and C HOLC		
		A	B	C	D	A vs. D	B vs. D	C vs. D
% Black	1988-89	0.25 [0.027]	0.24 [0.009]	0.29 [0.008]	0.35 [0.021]	-0.1** [0.042]	-0.1*** [0.029]	-0.06** [0.029]
	1998-99	0.26 [0.029]	0.25 [0.01]	0.29 [0.007]	0.35 [0.021]	-0.09** [0.043]	-0.1*** [0.029]	-0.06** [0.026]
	2008-09	0.24 [0.026]	0.24 [0.009]	0.28 [0.008]	0.34 [0.02]	-0.09** [0.041]	-0.1*** [0.027]	-0.06** [0.027]
	2018-19	0.20 [0.022]	0.21 [0.007]	0.24 [0.007]	0.30 [0.018]	-0.1*** [0.034]	-0.1*** [0.023]	-0.06** [0.024]
% White	1988-89	0.57 [0.027]	0.51 [0.015]	0.41 [0.005]	0.3 [0.011]	0.26*** [0.028]	0.21*** [0.025]	0.1*** [0.014]
	1998-99	0.51 [0.028]	0.45 [0.016]	0.33 [0.007]	0.24 [0.011]	0.27*** [0.028]	0.21*** [0.024]	0.09*** [0.014]
	2008-09	0.48 [0.027]	0.4 [0.016]	0.27 [0.007]	0.19 [0.011]	0.29*** [0.029]	0.21*** [0.023]	0.08*** [0.014]
	2018-19	0.46 [0.024]	0.37 [0.013]	0.24 [0.006]	0.17 [0.01]	0.28*** [0.028]	0.2*** [0.02]	0.07*** [0.013]
% Non-White	1988-89	0.43 [0.027]	0.49 [0.015]	0.59 [0.005]	0.7 [0.011]	-0.26*** [0.028]	-0.21*** [0.025]	-0.1*** [0.014]
	1998-99	0.49 [0.028]	0.55 [0.016]	0.67 [0.007]	0.76 [0.011]	-0.27*** [0.028]	-0.21*** [0.024]	-0.09*** [0.014]
	2008-09	0.52 [0.027]	0.6 [0.016]	0.73 [0.007]	0.81 [0.011]	-0.29*** [0.029]	-0.21*** [0.023]	-0.08*** [0.014]
	2018-19	0.54 [0.024]	0.63 [0.013]	0.76 [0.006]	0.83 [0.01]	-0.28*** [0.028]	-0.2*** [0.02]	-0.07*** [0.013]
Simpson's Diversity (1-D)	1988-89	0.36 [0.019]	0.35 [0.006]	0.37 [0.009]	0.35 [0.017]	<0.01 [0.03]	<0.01 [0.02]	0.02 [0.026]
	1998-99	0.38 [0.02]	0.37 [0.008]	0.38 [0.007]	0.36 [0.017]	0.02 [0.033]	0.01 [0.022]	0.02 [0.023]
	2008-09	0.42 [0.023]	0.4 [0.01]	0.39 [0.005]	0.37 [0.015]	0.05 [0.036]	0.03 [0.023]	0.03 [0.017]
	2018-19	0.49 [0.021]	0.48 [0.01]	0.46 [0.006]	0.43 [0.011]	0.06* [0.03]	0.05*** [0.018]	0.03** [0.013]

Notes: Cluster-robust standard errors are in parentheses with clustering done at the city-level. For each model we regress the outcome on HOLC grade indicators with the “D” security rating as the reference category. No controls are included. All regressions are weighted by 2018-19 total students. All models use city-level fixed effects to account for any differences that are fixed at the local level and that differ between CBSAs. All dollars denominated in USD 2018. *** p< 0.01, ** p< 0.05, * p< 0.1.

This longitudinal analysis shows student racial diversity increasing over time for all HOLC A-D grades. Notably, we see little movement in Simpson's Diversity Index from the late 1980s through the late 2000s for those in historically HOLC D neighborhoods. This might once again reflect the waning influence of court-order desegregation plans starting in the 1990s, and standalone could be a harbinger of resegregation in the years to follow. However, this trend reverses and spikes upward in the last decade, joining already upward sloping trend lines for A, B, and C schools. While these patterns could reflect more recent efforts that target school racial diversity through new avenues such as SES integration instead of historical policies based on race integration (Wells et al., 2020), it may also be an artifact of our chosen diversity measure.⁶⁰ Finally, over the three decades we consider, we observe gaps between D vs. A, B, and C grades that are persistent and often growing over time. These inequalities between those schools located today in what were historically the best-rated neighborhoods and those located today in what were the worst-rated neighborhoods highlight the potentially stubborn historical legacy of HOLC A-D map grades.

2.6.3 School Student Performance

2.6.3.1 Current Outcomes

Nationwide and by region, there are virtually no statistically significant differences across HOLC A-D grades for the outcomes of average student learning rates and average student test score trends. When differences are statistically significant, they are small as a share of the total variation in student test scores, with the largest statistically significant difference equating to just

⁶⁰ To complement the Simpson's Diversity Index results, we alternatively look at the Exposure Index for various race-group pairings. This is a well-known segregation index frequently used across the social sciences and differs from Simpson's Diversity Index by looking at exposure to different races via race-pair groupings. Looking at Figure B.7 and Table B.6, one can see an increase in White vs. Non-White exposure over time, driven largely by the increasing White-Hispanic Exposure Index. However, the White-Asian Exposure Index does increase moderately over time, too. This further corroborates that U.S. schools, at least those in urban areas we consider in our sample, are beneficiaries of shifting U.S. demographics. Also, the decreasing White-Black Exposure Index closely mirrors nationwide trends and lends evidence to the White-Black resegregation narrative resulting from a reversal of desegregation efforts in the 1960s and 1970s.

over 1/20th of a grade level (Table 2.7). Each of their respective HOLC A-D grade averages, both nationwide and by region, is not statistically different from their respective grand means. That said, given that the between-school standard deviation average student learning rates (i.e., ~0.07) and average student test score trends (i.e., ~0.04) are also small, the point estimates for these measures likely constitute a large share of the between-school standard deviation. Looking at these two outcomes, we see this is particularly the case for average student test score trends where differences between HOLC D vs. A (0.010, S.E. = 0.003), B (0.007, S.E. = 0.002), and C (0.004, S.E. = 0.001) make up 25%, 18% and 10% of the between-school standard deviation, respectively.

Table 2.7: School Student Performance and 1935-1940 HOLC A-D Grades, Pooled 2008-2019

	Average Student Math & ELA Scores	Average Student Learning Rate (Annual)	Average Student Trend in Test Scores
A	0.643*** [0.04]	0.006 [0.006]	0.010*** [0.003]
B	0.342*** [0.028]	0.008* [0.004]	0.007*** [0.002]
C	0.127*** [0.015]	0.005* [0.003]	0.004*** [0.001]
Constant	-0.365*** [0.013]	0.007*** [0.002]	-0.003*** [0.001]
FE	Y	Y	Y
R2	0.27	0.14	0.19
F	91.85	1.51	5.1
N	5,124	5,124	5,124

Notes: Cluster-robust standard errors are in parentheses with clustering done at the city-level. For each model we regress the outcome on HOLC grade indicators with the “D” security rating as the reference category. No controls are included. All regressions are weighted by 2009-2018 total number of math and ELA tests for pooled SEDA estimates. All models use city-level fixed effects to account for any differences that are fixed at the local level and that differ between CBSAs. *** p< 0.01, ** p< 0.05, * p< 0.1.

In contrast, the average student Math and ELA score outcome exhibits statistically significant differences between HOLC A-D grades nationwide and across all regions. This is true for all HOLC D schools versus their higher A, B, and C counterparts. Also, moving from A to B, B to C, and C to D, average student test scores decrease monotonically such that the gap between

A and D is the widest among all D vs. A, B, and C differences. This is once again true both nationwide and across each region.

These results tell us that while learning rates (i.e., how much students' scores improve each school year) and changes in educational opportunity (i.e., trends in test scores within grades across cohorts) are, on average, the same across all HOLC A-D grades, overall educational opportunity (i.e., average students' test scores) is not. Specifically, those schools located in historically D-assigned neighborhoods have less educational opportunity than those in A, B, and C neighborhoods. For example, in Table 7, nationwide A and D schools are separated by 0.64 SD units (S.E. = 0.04) or just over 1.9-grade levels. These gaps are present across all regions and widen to as much as 0.75 SD units (S.E. = 0.05) or about 2.25-grade levels in the West region and shrink to 0.54 SD units (S.E. = 0.10) or around 1.6-grade levels in the Northeast (Table S0.4). Comparing B and D schools paints a similar picture as above, albeit somewhat muted, with schools separated by 0.34 SD units (S.E. = 0.03) or just above one grade level. These gaps favoring B versus D schools are also exhibited within each region, increasing to as much as 0.46 SD units (S.E. = 0.07) in the West and decreasing to as little as 0.30 SD units (S.E. = 0.06) in the South (Table S0.4). Finally, C and D schools show the greatest similarity of all D vs. A, B, and C comparisons, with gaps shrinking to nearly single digits nationwide and across regions. Overall, C and D schools are separated by 0.13 SD units (S.E. = 0.02) or just under one-half grade level. Differences in educational opportunity are largest for the South region with a gap of 0.15 SD units (S.E. = 0.05), while the Northeast gap is smallest at 0.10 SD units (S.E. = 0.02) (Table S0.4). Both nationwide and by region, these differences in D vs. A, B, and C educational opportunity are nearly all statistically significant.

What do these findings mean? First and most importantly, they are only a snapshot of

present-day differences in educational performance measures by HOLC A-D grade. They thus cannot speak to whether gaps have risen or fallen over the past several decades. Without a complete accounting of these trends over time, we cannot measure progress nor bring historical context to bear. For example, suppose trends in educational opportunity favored D vs. A, B, and C schools over the past half-century. In that case, one might view the current gaps in this outcome as a historical lower bound and vice versa for a historical upper bound. Unfortunately, apart from the current 2009-2018 SEDA panel, we lack historical data on educational performance measures. Even so, we can make prognostications on what might be if the status quo remains. Given the large gaps in educational opportunity by HOLC A-D security rating and the near-zero S.D. unit changes in it for each HOLC A-D grade, the educational opportunity gap is expected to remain unabated into the future. The equality exhibited in average learning rates and average educational opportunity changes by HOLC A-D grade, which standalone might be a positive finding, will lead to a continued inequality in average educational opportunity across them, given the large and existing gaps in educational opportunity by HOLC A-D grade.

Finally, this equilibrium could have a silver lining if school and later life outcomes only depend on meeting a minimum educational opportunity threshold. For example, while gaps in educational opportunity would remain constant over time, a positive average change in educational opportunity uniform across A-D HOLC grades could, in time, raise all schools to and above the minimum educational opportunity threshold. Unfortunately, our findings do little to support this claim, as changes in educational opportunity, while positive, are very small and often fall below 0.01 SD units.

2.7 Conclusion

Between 1935-1940, the Home Owners' Loan Corporation (HOLC) assigned A (minimal risk) to D (hazardous) grades that arguably had meaningful effects on how the FHA, private banks, and mortgage lenders evaluated the creditworthiness and risk of home loans and mortgage insurance within residential neighborhoods over the next several decades. With the release of newly digitized HOLC A-D maps from the University of Richmond lead "Mapping Inequality Project," there has been a recent surge in research quantifying the negative impacts of redlining on long-term social and economic outcomes (Appel & Nickerson, 2016; Krimmel, 2018; Anders, 2018; Aaronson et al., 2020; Aaronson et al., 2021). However, this effort has yet to extend to K-12 public school educational outcomes to the best of our knowledge.

This paper examines the relationship between historic HOLC A-D maps and modern-day district and school funding patterns, racial diversity, and student performance. We employ a novel mapping strategy that links 1935-1940 HOLC A-D neighborhood grades to present-day districts and schools. At the district level, we find those mapped to historic HOLC D grades have the least favorable overall and local district-finance outcomes relative to those mapped to higher-rated HOLC A, B, and C grades.⁶¹ These results show how inequality in local per-pupil funding drives inequality in total per-pupil funding at the district level. Our findings also highlight the mitigating effects of redistributive federal and state policies on funding gaps generated by local differences. For example, we find those districts mapped to historic HOLC D grades have the most favorable state and federal district-finance outcomes today. These findings show a redistributive system targeting districts with higher percentages of students eligible for free or reduced-price lunch. However, the results also suggest that past neighborhood inequality lingers

⁶¹ Recall, that discrete A-D HOLC grades for districts are based on A-D HOLC weighted averages where weights were derived from A-D HOLC grade polygon areas.

well into the future. For example, those districts that receive redistributive funding to equalize local funding inequities are also those that serve families in neighborhoods disproportionately composed of HOLC D grade polygons.

In contrast to our district finance findings, we find the inverse relationship between HOLC A-D grades and funding at the school level. Schools mapped to the worst HOLC grades (i.e., D) have the most favorable school-finance outcomes than their higher-rated counterparts (i.e., A, B, and C). In reconciling the district and school-level finance results, we find that D schools have the greatest share of students eligible for free or reduced-price lunch relative to A, B, and C schools. Together, the district and school-level findings suggest that districts, regardless of their HOLC A-D grade, systematically target and allocate more money to schools in historic HOLC D neighborhoods and do so because these schools serve the largest shares of students from low-income households. Finally, much of these results persist across time, with overall positive time trends in outcome measures regardless of HOLC A-D grade but widening gaps between D vs. A, B, and C districts.

Overall, these results align with recent research that finds HOLC negatively impacted the development of urban neighborhoods and led to lower homeownership rates, home values, and racial diversity decades later (Aaronson et al., 2020; Aaronson et al., 2021). Making out-of-sample predictions from Aaronson et al. (2021) suggest that districts composed primarily of HOLC D neighborhoods (and therefore mapped to HOLC D grades in our paper) should have lower assessed property values, lower property tax bases, and less local funding than those districts composed primarily of A, B, and C HOLC neighborhoods. In general, any local funding shortfalls are addressed through targeted state and federal redistributive funding programs (e.g., Title I, IDEA) that allocate dollars to low-income districts to reduce funding gaps and better

equalize financing across districts. Our district finance results support this narrative and highlight a state and federal funding apparatus that appears to be effectively targeting those districts most in need, albeit at insufficient amounts to equalize funding altogether. While our results are not causal, our hypothesized mechanism underlying them hints at a lingering historical legacy of redlining, where HOLC neighborhood grades assigned in 1935-1940 predict local funding gaps today and where federal and state funds are needed to equalize funding for districts composed of primarily D neighborhoods.

We also find those schools located today in historically redlined residential neighborhoods to have, on average, larger shares of Black and non-White student bodies and less diverse student populations. These differences are persistent and growing overtime for the racial diversity outcomes, albeit for a smaller, less representative sample. However, while A assigned public schools have the highest percent White student populations, they also exhibit the highest student racial diversity levels via Simpson's Diversity Index. That said, A schools have the lowest Exposure Index values across all HOLC A-D grades and race-group pairings (i.e., White-Asian, White-Black, White-Hispanic, White-Non-White).

These findings reflect broader trends in U.S. K-12 public school demographics that have led to today's more racially and ethnically diverse school-age population. For example, over the past two decades, shares of White 5-17-year-olds have decreased from 62 percent to just over 50 percent, whereas shares of Hispanic 5-17-year-olds have increased to 25 percent from 16 percent (NCES, 2019). These public-school demographic shifts mirror the increasing racial diversity of the U.S. population, driven in part by diversifying urban demographics resulting from nationwide migration patterns that brought White families into cities from the suburbs and Black, Hispanic, and Asian families out to them (Wells et al., 2020). Even so, more diverse populations

may not always translate to more diverse schools. Since the 1990s, court-ordered desegregation plans from the 1960s and 1970s have been gradually lifted, leading to increased school segregation (Lutz, 2011; Reardon et al., 2012; Reardon et al., 2019). Also, intergroup exposure between Whites and Non-White students has decreased since the 1990s, with Non-White students attending schools with fewer shares of White students (Fiel, 2013).⁶² These countervailing forces could overwhelm, or at a minimum, limit some of the benefits that a more diverse U.S. population has on school racial diversity.

Finally, we also find that schools located today in what once were historically redlined areas have worse average ELA and math scores. However, there is virtually no difference in both average learning rates and trends in test scores across A-D schools. Notably, these findings only provide a snapshot of present-day differences in educational performance measures by HOLC A-D grade. They do not lend insight into whether gaps by HOLC A-D grade have abated over time. That said, looking forward, if the status quo continues, we would predict the gap in average test scores by HOLC A-D grade to remain unabated, given the limited to no difference in average learning rates and average test score trends by HOLC A-D grade.

Overall, our paper provides evidence that shows the stubborn association of HOLC A-D maps with modern educational outcomes and highlights the transmission of past neighborhood inequality to the present. In addition, these results suggest that education policymakers need to consider the historical implications of past neighborhood inequality on present-day neighborhoods when designing and implementing complex modern interventions that target inequitable outcomes between students of different socioeconomic and racial groups.

⁶² Importantly, Fiel (2013) finds this result was due to a growing share of the minority population relative to whites, not from increasing between-group segregation. This is reflected in the negative trends in percent White and positive trends in percent Hispanic and percent Non-White we present in this paper.

Chapter 3

The Impact of Lifting NYCDOE's Cellphone Ban on School Discipline and School Culture

3.1 Introduction

Debates about smartphone use in classrooms and schools are still ongoing and hotly contested among parents, teachers, and students alike. Over the past several decades, several large urban school districts, including Los Angeles, Chicago, and Miami, intermittently experimented with cellphone bans. However, most bans are now repealed due to their reported unpopularity with parents and students and concerns over equity (Rich & Taylor, 2014; NYC Press Office, 2015). In line with this trend, in March 2015, New York City (NYC) public schools lifted their longstanding districtwide cellphone ban and provided schools with significant discretion in designing and implementing school-level policies governing student cellphone use (NYC Press Office, 2015). As educators and policymakers act, there is still little research that quantifies the impact of smartphones on educational outcomes and whether cellphone bans are net negatives or positives to learning environments. Given the growing adoption and use of smartphones in today's 21st Century technology-facing society (Pew Research Center, 2018a), understanding their benefits and costs on learning and productivity remains a relatively nascent but important field of study.

In this paper, I estimate the impact of the New York City Department of Education (NYCDOE) mobile phone ban on per-pupil student suspensions and school culture. While the NYCDOE officially lifted its ban in March 2015, the monitoring and enforcement of the ban was effectively a school-by-school exercise given the large number of NYC public schools in the district and the onerous task of top-down monitoring of cellular phones by NYCDOE. Thus, while schools were granted limited autonomy in setting mobile phone policy prior to March

2015, implementation fidelity of NYCODE's ban likely varied among NYC public schools. However, some NYC public schools had full-time metal detectors, which reduced monitoring costs and improved ban enforcement (Matthews, 2014).⁶³ In addition, full-time metal detectors are often permanent fixtures in schools across time, with only a handful of schools that switch between full-time and non-full-time scanning status. I exploit these differences in metal-detector status between NYC public middle and high schools across time in pre-ban and post-ban periods using a differences-in-differences (DiD) estimation strategy with a single treatment period t such that one set of schools is treated at time t and another set of schools is not. Using this traditional 2x2 DiD framework, I compare the gains (deficits) in educational outcomes between those full-time scanning schools and non-full-time scanning schools before and after NYCDOE lifted their districtwide ban.

To do this, I created a unique dataset on the history of metal detectors and mobile phone policies from a survey of all NYC public secondary schools carried out in Summer 2020. I combined this survey data with NYCDOE student-level data, NYCDOE School Survey data, NYPD Student Safety Act data, and web-scraped InsideSchools.com metal detector data. Although cellphone ownership and use are not known for each student, research suggests that 78% of teenagers owned a cellphone in 2012 and 95% by 2018; therefore, cellphone bans and removals likely affected the large majority of NYC public middle and high school students in my sample (Pew, 2013, 2018b). Since some students in the panel may not have owned phones during the ban or at the time of repeal, this paper's estimates can be viewed as a lower-bound impact on smartphones. In addition, given the nature of smartphone distraction and peer effects, even those

⁶³ To such a degree that a cellphone truck cottage industry sprang up around NYC public schools with full-time metal detectors. Trucks acted as temporary holding places for students' cellphones during the school day and often charged daily fees (e.g., \$1/day) (e.g., Matthews, 2012). During my research, cellphone truck businesses that I found included Pure Loyalty, Safe 'N' Secure, Cell Secure, UCT, and Smart Dock.

students without phones could still be distracted by in-class smartphone usage.

I find that NYCDOE's cellphone ban removal positively impacted school discipline. On average, per-pupil suspensions (per 100 students) decreased by -0.99 following the removal of NYCDOE's cellphone ban policy. This finding is statistically significant. I also find evidence that the ban removal had a negative and statistically significant impact on various dimensions of students' perceptions of school culture, including student behavior and school safety. I do not find evidence that the ban impacted students' perceptions of teacher-to-student respect and peer-to-peer respect. Finally, this paper does not find evidence that the NYCDOE's cellphone ban impacted teachers' perceptions of school culture, including job satisfaction and school safety. These findings suggest that NYCDOE's cellphone ban, while active, may have led to more disciplinary actions because of strict and punitive cellphone ban policies. Given the negative relationship between school suspensions and academic outcomes (Craig & Martin, 2021), my findings suggest, all else equal, a net improvement in educational productivity. However, my results also highlight the trade-off of lifting the NYCDOE cellphone ban with costs to school culture, namely student perceptions of school safety (e.g., classroom, hallway, cafeterias) and student behavior (e.g., bullying).

This study adds to the literature on mobile phone bans by expanding the scope of outcomes considered, moving beyond test-score outcomes, and considering non-test score student and teacher outcomes, which may have secondary impacts on educational productivity. Research has provided evidence that cellphones directly impact educational productivity through distraction (Beland & Murphy, 2016); however, limited research documents how cellphone ban policies impact school discipline and school culture. As districts and schools consider similar ban removals or enforcements, they should consider these often-overlooked dimensions of schools

that also play a role in educational productivity (Craig & Martin, 2021).

The next sections of this paper are organized as follows: Section 3.2 provides an overview of the related literature, Section 3.3 describes my data, Section 3.4 outlines the empirical strategy, Section 3.5 provides and discusses the results. The paper concludes with Section 3.6.

3.2 Literature

3.2.1 Computers and Learning

To date, there is a limited selection of research related to the causal impact of smartphone use or other mobile device technology use on learning outcomes in school or non-school learning environments. Much of the extant research in the broader domain of technology use and learning explores the impact of computers or laptops on student outcomes, both within and outside the classroom environment. Research has found that increasing in-school computer access at the primary and secondary school level has no discernible impact on test scores and other educational outcomes (Angrist & Lavy, 2002; Barrera-Orsorio & Linden, 2009). In addition, a large swath of literature evaluating “one laptop per child” policies show little to no impact of laptop provision on student outcomes at the K-12 level (Shapley, Sheehan, Maloney & Caranikas-Walker, 2009; Suhr, Hernandez, Grimes & Warschauer, 2010; Cristia, Ibarra, Cueto, Santiago & Severin, 2012). Notably, much of this research focuses on the impact of the provision or access to computers or laptops. It does not necessarily extend to learning environments where these technologies are incorporated into the classroom in a structured, well-defined manner.

With that said, the use of computer-based instruction (CIT) or computer-aided instruction (CIA) within schools has shown mixed results, with RCTs providing both positive (Banerjee,

Cole, Duflo & Linden, 2007; Barrow, Markman & Rouse, 2009) and null effects of CIT on student learning and outcomes (Rouse & Krueger, 2004; Campuzano, Dynarski, Agodini & Rall, 2009). In addition, recent research from an RCT in migrant schools in Beijing suggests that out-of-school CIA could help improve student outcomes and interest in learning for remedial students (Lai, Renfu, Linxiu, Xinzhe & Rozelle, 2015). At the post-secondary level, results are less mixed. Although further evidence is needed, both RCTs and quasi-experiments have found that within-classroom computer use negatively impacts student achievement (Carter, Greenberg & Walker, 2017; Patterson & Patterson, 2017). Beyond within-school impacts, at-home computer provision for student use outside the classroom has adverse or null effects on academic and behavioral outcomes (Malamud & Pop-Eleches, 2011; Fairlie & Robinson, 2013). Apart from social science research, medical sciences research provides evidence that overall screen time is negatively associated with a battery of general health outcomes, including but not limited to obesity, body mass index, and cardiorespiratory fitness (Robinson, 1999; Crespo et al., 2001; Hancox et al., 2004; Marshall et al., 2004; Hardy 2009). However, questions of how total within-school computer screen time impacts students' short- and long-term outcomes and if these results are sensitive to age at exposure remain largely unanswered.

3.2.2 Smartphones and Learning

The extant literature related to the impact of within-classroom cellphone usage on student and school outcomes is even less populated than the research on CIT and CAI, with only a few high-quality causal analyses in either the United States or abroad. Although there are ample correlational studies that highlight a negative association between non-academic electronic device use in classrooms and student achievement (Jacobsen & Forste, 2011; Junco, 2012; Ravizza, Hambrick, & Fenn, 2014; Kraushaar & Novak, 2010; Ravizza, Uitvlugt, & Fenn,

2017), causal research is sparse. To date, there are only a few studies I am aware of that attempt to quantify the causal impact of mobile phones on school discipline and school culture. Beland & Murphy (2016), employing a two-way-fixed effects (TWFE) difference-in-differences empirical strategy, is the best current evidence on the impact of cellphone usage on student outcomes. The authors find that student performance on high-stakes exams in the UK increased due to within-school mobile phone bans. Notably, these results were primarily driven by the lowest-achieving students. More recently, using a within-student, within-item counterbalanced RCT, Glass & Kang (2018) measured the impact of off-task electronic device use on exam performance. The authors find that although off-task electronic device use did not impact short-term lecture comprehension, long-term retention was negatively impacted (Glass & Kang, 2018). Since electronic devices include both computers and cellphones, these results should be considered as the composite impact of off-task computer and cellphone use, not an isolated measure of the causal impact of cellphones in the classroom.

There are a variety of other studies related to cellphone use in educational settings. Fryer (2016), which provided free cellphones to students in Oklahoma City Public Schools to receive daily text messages related to the link between human capital and future outcomes, found no measurable impacts on attendance, behavior, and test scores. Bergman (2020) found that providing parents with information about their children's upcoming assignments via text message improved student test scores. Multitasking research in psychology has also looked at the impact of technology on student outcomes finding negative effects on learning and task completion (Ophir et al., 2009; Smith et al., 2011; Levine et al., 2013; Kuznekoff and Titsworth, 2013; Dietz and Henrich, 2014). More broadly, psychology research has found cellphones can be distracting and negatively impact reaction times, performance on and enjoyment of focal tasks, and

cognitive capacity (Caird, Willness, Steel & Scialfa, 2008; Stothart, Mitchum & Yehnert, 2015; Isikman, MacInnis, Ulkumen & Cavanaugh, 2016; Ward, Duke, Gneezy & Bos, 2017). Although these results are not directly linked to cellphone use in academic environments, they could be important harbingers of smartphones and other electronic devices' harmful effects on learning and performance at school and in the classroom.

Thus, the above literature overview provides evidence that technology can both enhance and hurt student learning and performance. For computers, despite their increased rates of adoption and use in and outside of the classroom, systematic evidence and robust data analyses on how this technological innovation impacts educational, social, and employment outcomes are limited with a lack of consensus on how computer provision, access, and overall screen time exposure impacts student outcomes. For smartphones, apart from a few notable studies, there is a dearth of rigorous empirical evidence directly addressing if and how they affect student performance and impact short- and long-term outcomes. As the adoptions and use of smartphones and other computing devices capture increasing shares of our time and proliferate in our lives, it is vital to understand if and how learning and in-school experiences are impacted and whether these effects vary by use, age, race, and gender. This research could help inform future cellphone policies at both the school and district level for NYC public schools and be one of only a handful of studies to provide systematic quantitative evidence of the causal impact of cellphones in schools.

3.3 Data

To research the connection between student discipline and mobile phone policies in NYC public schools, I combine NYCDOE administrative data on students and pedagogues with (1) school culture survey data from the NYCDOE, (2) cellphone and metal detector use data from

my author-led survey of NYC public school middle and high school principal and parent coordinators, and (3) school metal detector data from the NYPD and InsideSchools.com. In the next section below, I provide additional details on each of these datasets.

3.3.1 Administrative Data

The NYC student-level data includes NYCDOE and the New York State Education Department (NYSED) data on all students attending NYC public middle and high schools during 2006-07 to 2016-17 school years. It is a rich data set that captures student performance, student demographics, teacher turnover, and school climate. Student performance metrics include (1) NYDOE student-level information, such as attendance, course grades, GPA, ELA and math scores, graduation status, standardized test scores (e.g., ACT, SAT), and suspensions, and (2) NYSED student-level information, namely Regents Exam scores. The student demographic data contains a host of student-level characteristics, including sex, race, age, ethnicity, FRPL status, and special education status. NYCDOE assigns each NYC public school student an anonymized unique identifier, allowing researchers to track students over time. I use this data to create an NYC public middle and high school panel covering the 2012-2013 school year through the 2016-17 school year that includes information on school-level student demographics (e.g., race, sex, ELL status, FRPL status, SWD status) and school-level discipline. My primary measure of discipline is per-pupil total suspensions which includes both principal and superintendent suspensions. To construct this outcome, I use the NYCDOE student-level disciplinary data and aggregate total suspensions at the school level for each school year in my panel. I subsequently divide total suspensions by fall enrollment numbers for each respective school and school year combination.

3.3.2 School Characteristics

3.3.2.1 NYCDOE School Culture Survey

Each year, NYCDOE middle and high school students and their respective teachers and parents complete a school culture survey designed to evaluate NYC public schools' effectiveness in supporting student success (Craig & Martin, 2021). Following a similar procedure as Craig & Martin (2021), I use annual survey responses from teachers and students for the 2012-2013 through 2016-17 school years to construct measures of school culture for a variety of student-teacher dimensions, including: (1) respect, (2) behavior, (3) safety, and (4) job satisfaction (teachers only). I include all student and teacher dimensions, survey elements, and panel coverage in Table 3.1. For students, I have full panel coverage for all school culture dimensions sans job satisfaction, which I reserve for teachers. While many of my student-based school culture dimension indexes directly mirror Craig & Martin (2021), I use different survey elements for the dimension of respect due to incomplete panel coverage of certain survey elements with my SY2012-SY2016 panel. Thus, instead of having a student-to-teacher measure of respect, as seen in Craig & Martin (2021) (i.e., “Most students at my school treat teachers with respect”), I opt for a peer-to-peer (i.e., “Most students at this school treat each other with respect”) and teacher-to-student (i.e., “My teachers treat me with respect”) measure of respect, which both have full panel coverage.

For the respect and safety dimensions, each survey response is on a 4-point scale from "strongly disagree" to "strongly agree."⁶⁴ As in Craig & Martin (2021), I code student responses such that higher is better and normalize each outcome by school year to have a mean 0 and

⁶⁴ Recoding is required to ensure higher numeric values do indeed represent better school conditions when this occurs. For some survey questions, the ordering of the 4-point scale varies by survey year. For example, for the dimension of safety, all safety-related survey responses for SY2014-SY2016 are ordered “strongly disagree” to “strongly agree” but ordered “strongly agree” to “strongly disagree” for SY2012-SY2013.

standard deviation of 1. For student behavior, each survey response is on a 4-point scale from "none of the time" to "all of the time." Once again, I code responses so that higher is better and normalize each outcome by school year to have a mean of 0 and a standard deviation of 1. Finally, to construct indexes for each respect, behavior, and safety dimension, I create a new summary index that is the average of all standardized responses. I then re-standardized the combined average for each school culture dimension to create the final student-based summary indexes I use as outcomes in the analyses below.

Table 3.1: NYCDOE School Survey Questions on School Culture, Teachers & Students

Student Questions	Available
<i>Respect</i>	2007 - 2016
• Most students at this school treat each other with respect • My teachers treat me with respect	
<i>Behavior</i>	2006 - 2016
• At this school students harass, bully, or intimidate other students. • At this school students get into physical fights. • At this school students drink alcohol or use illegal drugs while at school. • At this school students there is gang activity	
<i>Safety</i>	2006 - 2016
• I feel safe in my classes at this school. • I feel safe in the hallways, bathrooms, locker rooms, and cafeteria of this school. • I feel safe outside around this school.	
Teacher Questions	Available
<i>Respect</i>	
• At my school, most students treat adults with respect.	2006 - 2013
<i>Behavior</i>	
• At my school, students are often harassed or bullied in school. • At my school, students use of alcohol or drugs. • At my school, there are conflicts based on differences (race, color, creed, ethnicity, national origin, citizenship/immigration status, religion, gender, gender identity, gender expression, sexual orientation, disability, or weight). • At my school, gang activity is a problem in school.	2006 - 2013
<i>Safety</i>	Varies
• Order and discipline are maintained at my school.	2006 - 2016
• At my school, I am safe.	2006 - 2013
• At my school, crime and violence are a problem.	2006 - 2013
<i>Job Satisfaction</i>	
• I usually look forward to each working day at this school.	2012 - 2016

Notes: Data compiled from NYC School Surveys SY2006-SY2016. A special thanks to Ashley Craig and David Martin for sharing replication code and data for this exercise.

For teachers, I exclude the respect and behavior dimensions entirely from my analysis, given that the survey elements underlying them only overlap my school year panel from 2012 to 2013. Given these panel limitations, most of the school safety elements are also excluded from my analysis. However, one element of school safety (i.e., “Order and disciplined are maintained”), in addition to a new element I include to measure teacher satisfaction (i.e., “I usually look forward to each working day”), is available across all SY2012-SY2016 panel years. I use these questions as a proxy for dimensions of school safety and job satisfaction in my analyses below and follow a similar procedure for the student summary indexes to create my teacher-based school culture measures.⁶⁵

3.3.2.2 Metal Detector & Cellphone Survey

NYCDOE had an official districtwide cellphone ban from 1988 through March 2015; however, variation in school-level enforcement mechanisms and inconsistent top-down enforcement by NYCDOE over time likely created variation in the *de facto* cellphone policies at the school level (Gewertz, 2006; Matthews, 2014; Rich & Taylor, 2014). I hypothesize that NYC public schools with full-time metal detectors had the most robust ban enforcement of all NYC public schools, given the ability of metal detectors to effectively identify cellphones each morning as students enter their respective school or educational campus. Since NYCDOE does not have a central repository for school-level cellphone policies and school-level metal detectors, I conducted a survey of all NYC public middle and high schools in Summer 2020 for pre-and post-ban periods. Before contacting school officials, I obtained permission from the NYCDOE.⁶⁶ Subsequently, I sent my approved survey to principals and parent coordinators for every active

⁶⁵ For the teacher-based school culture dimensions, there is only one survey element. As a result, I only standardize these responses to have 0 mean and 1 standard deviation. No average across standardized responses and re-standardization is required.

⁶⁶ Harvard IRB approval and NYCDOE IRB approval obtained for both student-level data and this principal/parent coordinator survey.

NYC public middle and high school in the district. I generated an NYC metal detector and cellphone policy panel from the survey covering pre-NYCDOE cellphone and post-NYCDOE cellphone ban school years. Questions related to NYCDOE's cellphone ban sought to understand better which schools still had active bans on mobile devices following NYCDOE's repeal of their districtwide ban in March 2015 and what school-level bans practically looked like within schools across time. Since a ban in one school may differ from another, mobile device survey responses helped capture and account for this between-school variation in ban definition and implementation across time.

The first part of the survey asked questions related to a school's metal detector status, including whether a school ever had a metal detector, and if so, the type of metal detector scanning (e.g., full-time, random), the start and end date of metal detector use, and what company the metal detectors were purchased from. The second part of the survey asked questions related to a school's pre-ban (i.e., SY 2005-06 to SY 2014-15) and post-ban (SY 2015-16 to 2019-20) cellphone policies, including whether cellphones were allowed in the school, classrooms, or during lunch and study hall periods, and the disciplinary practices if mobile phone policies were broken during the school day.

I received completed surveys from 63 schools, representing just over 5% of my sample's target NYC public middle and high schools across the 32 NYCDOE districts. This response rate is low relative to other non-governmental surveys in academic research (e.g., Beland & Murphy, 2016; Card et al., 2012; Hall & Krueger, 2012; Heffetz, 2011). As a result, the survey was relegated to supplemental data and used only to inform the identification strategy instead of drive it. Survey results regarding metal detector usage and presence over time also served as a helpful cross-reference when constructing the full-time metal detector panel. I report the survey results

in Appendix Table C.1, where responses are summarized by pre-policy and post-policy periods and broken out by those schools indicating they were full-time metal detector schools. Notably, all schools reporting as metal detectors pre-policy had full-time metal detectors during the post-policy period. All but two full-time schools had metal detector start dates prior to the 2010-11 school years. The fixed nature of full-time scanning schools across time is further corroborated by the 2014-15 through 2020-21 metal detector panel I constructed using NYC SSA and InsideSchools.com school-level data. I discuss this in further detail in the next section.

3.3.2.3 NYPD and InsideSchools.com

Given the poor response rate to the cellphone and metal detector survey, this paper required supplemental data to construct a more robust panel of metal detector use in NYC public middle and high schools. Separate from the metal detector survey, I construct a metal detector panel covering the 2013-14 through 2019-20 school years using InsideSchools.com and NYPD School Safety Act (SSA) data. Metal detector data scraped from Insideschools.com is part of a collaboration between InsideSchools.com, New York Civil Liberties Union (NYCLU), and WYNC New York Public Radio. Although not intended to track in-school metal detector use, NYPD SSA data indirectly does via NYPD's inclusion of "Scanning Type" when reporting in-school incidences.⁶⁷ The inclusion of "Scanning Type" allowed this project to distinguish between schools with permanent full-time metal detectors, schools' temporary scanning types (e.g., hand wands, portable metal detectors), and schools with neither full-time nor temporary metal detectors. NYPD SSA data is broken down by quarter, allowing one to map reporting to the appropriate school year (e.g., NYPD Q1 2014 mapped to SY 2013-2014; NYPD Q3 2014

⁶⁷ There is an ongoing lawsuit between the NYPD and NYCLU where NYCLU sued the NYPD in Summer 2020 to force the NYPD to comply with school-level metal-detector disclosure requirements. Ironically, albeit less exhaustive and usable than the disclosure requirement mandates, this data is already contained in NYPD SSA reports. With a bit of elbow grease, one can glean a school's metal detector status from the NYPD SSA data for the 2014-15 through 2020-2021 school years (i.e., normalize schools' names and educational campuses, link to NYSED BEDSCODE, link to NYCDOE DBN).

mapped to SY 2014-15).

3.3.3 Sample

I impose a set of restrictions on the data to obtain a balanced panel of public middle and high schools located within NYCDOE and restrict my sample from 2012-13 to 2016-17 school years to both (1) avoid the material change NYCDOE made to their school discipline policy enacted Fall 2012 and (2) have full panel coverage of my school culture outcomes.⁶⁸ Thus, my analysis sample includes only those NYC public middle and high schools across the 2012-13 through 2016-17 school years that have complete information on school discipline, metal detector status, and student demographics (e.g., sex, race, free and reduced-price lunch eligibility). Restricting my sample to SY2012-SY2016 removes about 52% of the school population from my original 2006-07 to 2016-17 school-level panel (i.e., 20,111 observations). Next, I keep only NYC middle and high schools, accounting for about 60% of the remaining sample. I also drop any schools in my panel that have missing metal detector data or did not merge with the NYPD and InsideSchools.com full-time scanning panel. This constitutes just over 13% of the remaining sample. Finally, I keep only those schools present in my 2012-13 to 2016-17 panel each school year. This restriction removes another 9% of the sample and gives me a final sample of 3,965 across my panel years 2012-13 to 2016-17, representing 793 unique NYC public middle and high schools. Please refer to Table 3.2 and Table 3.3 below for an overview of descriptive statistics for all NYC public middle and high schools spanning 2012-13 to 2016-17 school years, and for only those included in my analysis sample. In Table 3.3, summary statistics

⁶⁸ Beginning school year 2012-13, NYC public schools abolished suspensions for non-violent, disorderly behavior and replaced them with more minor punitive punishments such as classroom removal. The new districtwide policy encouraged schools to embrace non-punitive interventions in place of principal and superintendent suspensions (Craig & Martin, 2021). For my school culture outcomes, some survey elements are unavailable in earlier years of the survey. To ensure a complete panel, I restrict my sample to only those years where I have full access to each survey question.

are split by full-time scanning and non-full-time status, which plays a vital role in this paper’s identification strategy described in the next section.

Table 3.2: Descriptive Statistics for NYC Public Middle and High Schools

	NYC Public Middle and High Schools				
	2012	2013	2014	2015	2016
Race					
% Asian	0.148 [0.174]	0.149 [0.174]	0.149 [0.173]	0.15 [0.173]	0.152 [0.175]
% Black	0.312 [0.262]	0.309 [0.262]	0.304 [0.261]	0.298 [0.258]	0.291 [0.254]
% Hispanic	0.389 [0.236]	0.389 [0.235]	0.391 [0.236]	0.39 [0.236]	0.391 [0.235]
%White	0.139 [0.179]	0.14 [0.179]	0.141 [0.178]	0.144 [0.178]	0.145 [0.177]
Sex					
% Female	0.482 [0.103]	0.482 [0.104]	0.483 [0.106]	0.482 [0.106]	0.482 [0.105]
School Characteristic					
% ELL	0.127 [0.14]	0.125 [0.14]	0.12 [0.134]	0.117 [0.13]	0.125 [0.136]
% FRPL	0.727 [0.178]	0.735 [0.171]	0.724 [0.171]	0.711 [0.176]	0.713 [0.182]
% SWD	0.185 [0.161]	0.192 [0.164]	0.198 [0.163]	0.203 [0.165]	0.207 [0.165]
Number of Schools	1,134	1,183	1,195	1,191	1,176

Notes: This table presents descriptive statistics for select covariates for all NYC public middle and high schools across SY2012-SY2016. This excludes schools serving only K-5 grade levels. Means are presented first with standard deviations shown below in squared parentheses. FRPL stands for “free and reduced-price lunch”, ELL stands for “English language learners” and SWD stands for “students with disabilities”. All school-level characteristics taken from NYCDOE student-level data and aggregated at the school level. Means and standard deviations weighted by student enrollment. *Sources:* NYCDOE, NYPD SSA, InsideSchools.com

Table 3.3: Descriptive Statistics Pre- and Post-Policy Change

	No Fulltime Scanning (1)					Fulltime Scanning (2)					Pre Diff. (SD Units)		Post Diff. (SD Units)	
	2012	2013	2014	2015	2016	2012	2013	2014	2015	2016	(1) vs. (2)	(1) vs. (2)	(1) vs. (2)	(1) vs. (2)
Race														
% Asian	0.19 [0.19]	0.19 [0.19]	0.19 [0.19]	0.19 [0.19]	0.20 [0.19]	0.10 [0.12]	0.10 [0.12]	0.10 [0.12]	0.10 [0.12]	0.10 [0.12]	0.75	0.83		
% Black	0.25 [0.25]	0.24 [0.25]	0.24 [0.24]	0.23 [0.23]	0.22 [0.23]	0.41 [0.25]	0.4 [0.24]	0.39 [0.24]	0.38 [0.24]	0.37 [0.23]	-0.64	-0.65		
% Hispanic	0.38 [0.24]	0.38 [0.24]	0.38 [0.24]	0.38 [0.24]	0.38 [0.24]	0.43 [0.24]	0.43 [0.24]	0.43 [0.24]	0.44 [0.24]	0.44 [0.24]	-0.21	-0.25		
% White	0.18 [0.2]	0.18 [0.2]	0.18 [0.2]	0.18 [0.2]	0.18 [0.19]	0.06 [0.1]	0.06 [0.1]	0.06 [0.09]	0.07 [0.09]	0.07 [0.1]	1.20	1.10		
Sex														
% Female	0.50 [0.09]	0.50 [0.09]	0.50 [0.09]	0.50 [0.09]	0.50 [0.09]	0.46 [0.12]	0.46 [0.12]	0.46 [0.12]	0.46 [0.12]	0.46 [0.11]	0.33	0.36		
School Characteristic														
% ELL	0.12 [0.13]	0.12 [0.13]	0.11 [0.13]	0.11 [0.12]	0.12 [0.13]	0.17 [0.19]	0.17 [0.2]	0.16 [0.19]	0.16 [0.19]	0.18 [0.2]	-0.26	-0.30		
% FRPL	0.71 [0.19]	0.72 [0.18]	0.71 [0.18]	0.69 [0.19]	0.69 [0.19]	0.81 [0.1]	0.8 [0.09]	0.79 [0.1]	0.78 [0.1]	0.79 [0.1]	-1.00	-1.00		
% SWD	0.15 [0.07]	0.16 [0.07]	0.17 [0.07]	0.17 [0.07]	0.18 [0.07]	0.16 [0.06]	0.17 [0.06]	0.18 [0.06]	0.18 [0.07]	0.18 [0.06]	-0.17	0.00		
Number of Schools	603	603	603	603	603	190	190	190	190	190				

Notes: This table presents descriptive statistics for select covariates for those NYC public middle and high schools in my analysis sample, including, full-time scanning schools, non-full-time scanning schools, and differences between the two. The analysis sample is balanced across SY2012-SY2016 and contains no PK-5 schools. Any schools with missing metal detector data were also removed. Means are presented first with standard deviations shown below in squared parentheses. The difference between full-time scanning and non-full-time scanning schools over the entire panel is provided in the final column in SD units. Given the similarity in means across time for each covariate, differences are taken only for one school year. In this case SY2016. FRPL stands for “free and reduced-price lunch”, ELL stands for “English language learners” and SWD stands for “students with disabilities” *Sources:* NYCDOE, NYPD SSA, InsideSchools.com.

3.4 Empirical Strategy

3.4.1 Traditional 2x2 DiD Models

The ideal research design to answer this question is a randomized control trial where randomization occurs at the school level within a given district. In the simplest version, schools are randomly assigned to allow cellphones or not. To explore if and how treatment effects vary by aggregate screen time, randomization could be done such that schools are randomly assigned to varying levels of cellphone ban strictness ranging from zero tolerance to partial bans to total allowance. In the absence of an RCT, this study proposes to use a differences-in-differences (DiD) empirical design. Although NYCDOE had an official districtwide cellphone ban until March 2, 2015, ban enforcement was inconsistent across schools, particularly between those with full-time metal detectors and those without. This study relies on the idea that those schools with full-time metal detectors throughout NYCDOE's districtwide ban had the strictest enforcement of the NYCDOE ban since mobile phones are easily detectable by scanners.⁶⁹ Thus, I estimate the impact of NYCDOE's cellphone ban on student outcomes by exploiting variation in (1) NYCDOE's cellphone ban across time and (2) NYC public schools' full-time scanning status. The baseline model specification is:

$$Y_{st} = \beta_0 + \beta_1 Ban_t \times Metal_{s,2014} + \theta_t + \nu_s + \epsilon_{st}$$

where Y_{st} is the outcome for school s in school year t . My primary measure of school discipline is total suspensions (i.e., principal and superintendent) which I transform into per-pupil terms and scale by 100. I also use a variety of teacher and student school culture measures described

⁶⁹ This is not merely speculation. There is evidence of a small but lucrative cottage industry of cellphone trucks that sprang up around schools with full-time metal detector scanning during the decade of NYCDOE's active cellphone ban (e.g., Harris & Schweber, 2015; Matthews, K., 2012). For a daily fee, cellphone trucks and local bodegas held students' mobile phones (e.g., \$1/day). Interviews with cellphone truck owners and survey responses from NYCDOE principals and parent coordinators further corroborate this narrative.

above in Section 3.3. Ban_t is an indicator for whether NYCDOE had an active districtwide cellphone ban at the beginning of the school year t , and $Metal_{s,2014}$ is an indicator for whether school s had full-time scanning during the 2014-2015 school year.⁷⁰ Since NYCDOE formally lifted their districtwide cellphone ban at the end of the 2014-15 school year, Ban_t is simply an indicator variable for whether an observation occurs after the 2014-15 school year. The time-specific intercept θ_t , captures any shared shocks by all students within a given school year, the school-specific intercept ν_s accounts for any fixed school specific effects, and ϵ_{st} is the idiosyncratic random error for a given s school in year t . The interaction term $Ban_t \times Metal_{s,2014}$ can be interpreted as *a de facto* cellphone ban determined by enforceability, not the *de jure* districtwide ban, which I assume to be non-binding for a school in the absence of an effective enforcement mechanism.⁷¹ The interaction term, $Ban_t \times Metal_{s,2014}$, is equivalent to the traditional 2x2 DiD treatment indicator, D_{st} , but is partitioned in model specification to clarify the embedded assumption that the districtwide ban is binding for all schools with full-time metal detectors and non-binding for all schools without full-time scanning. I use the traditional 2x2 DiD framework for all model specifications, where one set of schools is treated at time t , and another set is not. Thus, compared to the often-used two-way-fixed-effect (TWFE) DiD approach, which requires a more stringent set of assumptions to retrieve the ATT, selection into treatment does not occur at different time intervals across the panel (Goodman-Bacon, 2021).⁷²

⁷⁰ The full-time scanning status is relatively stable over time. For the SY2014-SY2019 metal detector panel, very few schools switch between full-time and non-full-time scanning status. In addition, the 2x2 DiD results are not sensitive to which school year (i.e., SY2014, SY2015, SY2016, SY2017, SY2018, SY2019) is used to determine full-time scanning status.

⁷¹ Once again, this is not simply an assertion. Both (1) NYCDOE principal and parent coordinator survey responses and (2) interviews with cellphone truck owners confirm this assertion for many school buildings and campuses with full-time metal detector scanning.

⁷² For issues with the TWFE DiD approach, refer to Goodman-Bacon (2021), who finds that a causal interpretation of TWFE DiD estimates require both (1) the traditional parallel trends assumption found in 2x2 DiD and (2) constant treatment effects. See Calloway and Sant'Anna (2021) for a newly developed DiD approach that effectively handles the problems with TWFE DiD and differential timing put forth by Goodman-Bacon (2021).

Next, to account for student sorting into schools based on observable characteristics, in the model specification (2), I insert a vector of control variables, $\mathbf{X}_{st}$, which includes the percentage of students eligible for free and reduced-price lunch (FRPL), percentage of English language learners (ELL), percentage of students with a disability (SWD), percentage female, fall enrollment, and percent breakdowns across Asian, Black, Hispanic, and White race categories. This should help account for any non-random sorting into schools at the observable level and improve the precision of the 2x2 DiD point estimate. With these controls alone, unobservable selection into schools based on metal detector status is likely still problematic; however, the DiD model design should help mitigate biases brought about by any remaining unobservable differences.

$$Y_{st} = \beta_0 + \beta_1 Ban_t \times Metal_{s,2014} + \beta_2 \mathbf{X}_{st} + \theta_t + \nu_s + \epsilon_{st}$$

Another potential threat to identification when using 2x2 DiD is the existence of contemporaneous policy changes at the time of treatment that also relates to the treatment. For example, if there are other positive changes to a school at the time of NYCDOE's ban repeal, which are correlated with the positive (negative) impact of NYCDOE's cellphone ban on the outcome of interest, I would overestimate (underestimate) the effect of the ban repeal. Given the influence school leadership has on school culture, academic and disciplinary norms, any changes in school leadership should be accounted for to avoid conflating the impact of school leadership and the effect of NYCDOE's cellphone ban. To do so, I use educator-level NYCDOE pedagogue data and create an indicator variable for school s at time t for principal turnover called $OtherPolicy_{st}$, which is coded as 1 for any school s in school year t that has a new entering principal at the start of the fall school year. I add this variable to model specification (2) to get

model specification (3), which now accounts for policy changes, formal or informal, that coincide with the addition of new principals in the year at the treatment cutoff. This is the final and preferred model specification used in this paper.

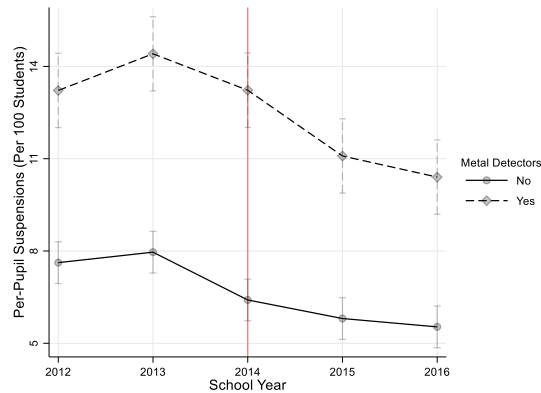
$$Y_{st} = \beta_0 + \beta_1 Ban_t \times Metal_{s,2014} + \beta_2 X_{st} + \beta_3 OtherPolicy_{st} + \theta_t + \nu_s + \epsilon_{st}$$

Barring any threats to identification, this specification should allow one to capture the overall causal impact of a cellphone ban on schools for the school culture (i.e., student and teacher dimensions) and school disciplinary (i.e., per-pupil suspensions) outcomes of interest.

3.4.2 Threats to Identification

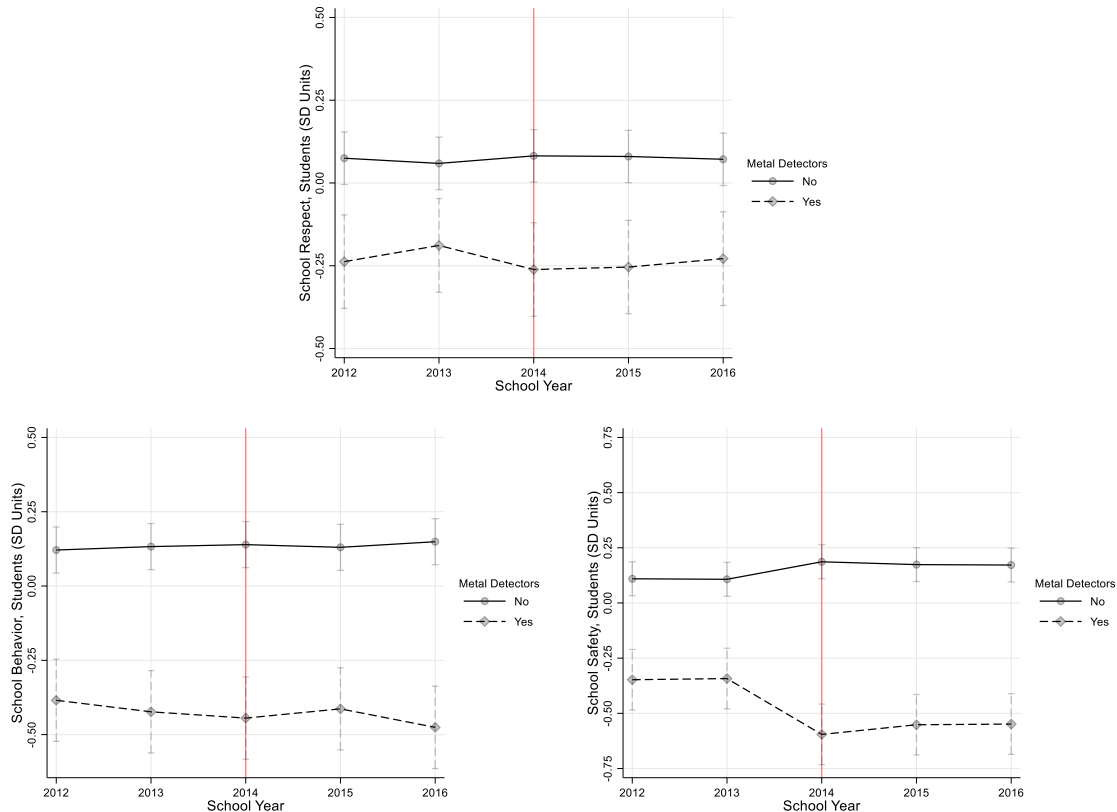
While this paper does account for the threat of other relevant policy changes specifically related to school leadership changes that occurred concurrently with NYCDOE's cellphone ban, there are other potential threats to identification that, if present, will bias my 2x2 DiD estimation results. Recall, for 2x2 DiD to produce unbiased ATET impact estimates, sampling units selecting into treatment must not have systematically different trends than those selecting into no treatment. Colloquially, this is known as the parallel trend assumption, and if violated, the 2x2 DiD point estimates will be biased. To check this assumption, I plot my outcomes over time by treatment status (Figure 3.1-Figure 3.4). Looking at Figure 3.1, the parallel trend assumption for per-pupil suspensions (per 100) looks to hold; there are similar increases in the outcome from SY2012 to SY2013 by treatment status and comparable decreases SY2013 to SY2014 by treatment status. However, slopes may be steeper from SY2012 to SY2013 and flatter from SY2013 to SY2014 for the treatment group, casting a sliver of doubt on the 2x2 DiD estimates for this outcome.

Figure 3.1: Per-Pupil Suspensions by Scanning Status, SY2012-2016



Notes: Per-pupil suspensions (per 100) is based on the sum of principal and superintendent suspensions for all NYC public middle and high schools in the analysis sample. Schools coded "Yes" had full-time metal detector scanning, those coded "No" did not have full-time metal detector scanning (i.e., none, random, hand wands). The vertical red line marks the SY2014-15, the last year before NYCDOE lifted their ban.

Figure 3.2: School Culture (Students) by Scanning Status, SY2012-2016



Notes: [top] Respect (Students) [left] Student Behavior (Students) [right] School Safety (Students). All measures calculated from student responses to annual questions on the NYCDOE School Survey. Schools coded "Yes" had full-time metal detector scanning, those coded "No" did not have full-time metal detector scanning (i.e., none, random, hand wands). The vertical red line marks the SY2014-15, the last year before NYCDOE lifted their ban.

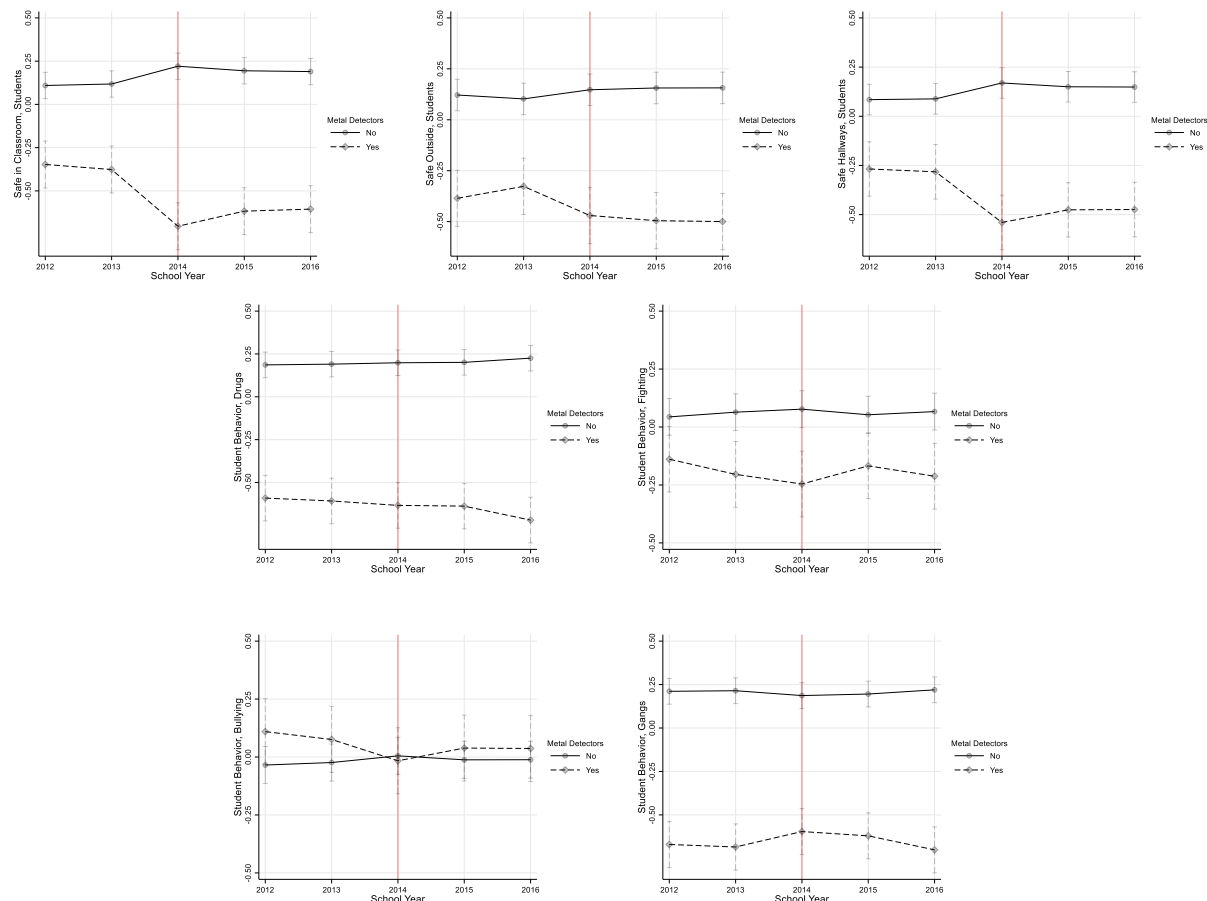
Moving to the school culture outcomes, one can see in Figure 3.2 that the parallel trends assumption for student perception of respect and student behavior indexes is likely met. However, for the final student-based school culture index, school safety, there is a clear downtrend in the treatment group one year prior to treatment in SY2014-15, which is not mirrored by the control group. Looking at Figure 3.3, there looks to be differential trend lines by treatment status for nearly every individual measure of school safety and student behavior. These findings provide evidence of a clear violation of the parallel trend assumption for many of the student-based school culture outcomes considered in this paper. Unless one thinks the assumption of parallel trends is conditionally met after adding the model controls, the 2x2 DiD estimate for the school safety index and individual school safety and student behavior outcomes will not produce the ATET of the treatment.

Importantly, since NYCDOE students and teachers typically submit responses to the NYC School Survey in late Spring of each academic year,⁷³ the responses to the SY2014-15 could be capturing early treatment impacts of NYCDOE's cellphone ban removal, which was announced in January 2015 and officially lifted in early March 2015 (Taylor, 2015). If true, the last school year before treatment is SY2013-14, not SY2014-15. Based on this alternative SY2013-14 pre-treatment cutoff, the trend lines of nearly all student-based school culture outcomes (indexes and disaggregated individual measures) appear to satisfy the parallel trend assumption.⁷⁴

⁷³ For example, the SY2021-22 NYC School Survey launched mid-February and is expected to close mid-April of 2022. The SY2020-21 NYC School Survey launched mid-May and was available until mid-June of 2021. The SY2019-20 NYC School Survey window was mid-Feb through end of March, although respondents were allowed to submit surveys until June 26, 2020 (NYCDOE, 2022).

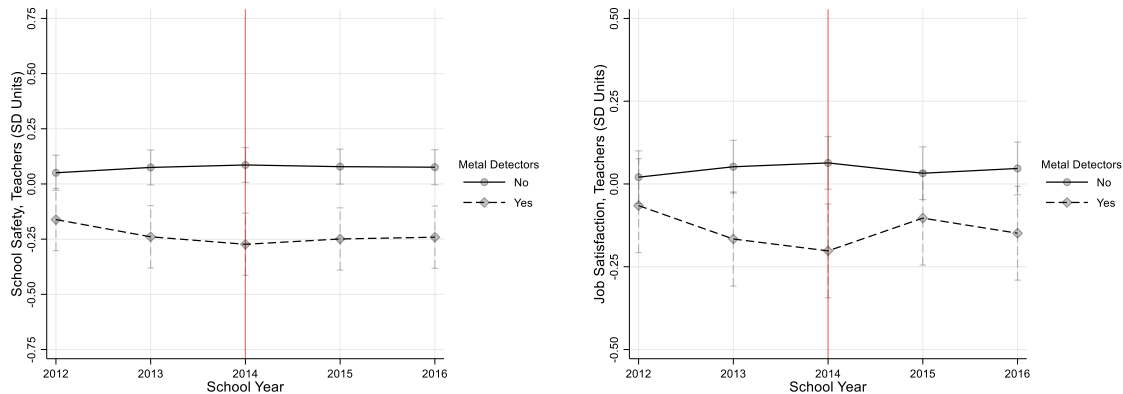
⁷⁴ In a brief sensitivity analysis following the main results in Section 3.5, I provide estimates for both treatment cutoffs and discuss implications. Overall, using the SY2013-14 cutoff relative to the SY2014-15 cutoff, the 2x2 DiD impact estimates for the school culture outcomes are larger in absolute magnitude, and I find more statistically significant results. Ideally, I would like to have the exact submission date for each respondent's survey to determine whether students and teacher submitted their surveys before or after the ban repeal. To the best of my knowledge, this data does not exist or is not publicly available.

Figure 3.3: School Culture (Students) by Scanning Status, Individual Outcomes, SY2012-2016



Notes: [top left] Safe in Classroom (Students) [top middle] Safe Outside (Students) [top right] Safe in Hallways (Students) [middle left] Student Behavior, Drugs [middle right] Student Behavior, Fights [bottom left] Student Behavior, Bullying [bottom right] All measures calculated from student and teacher responses to annual questions on the NYC School Survey. NYC School Survey data cleaned and linked to an author-constructed metal-detector panel from NYPD SSA incident data. Schools coded “Yes” had full-time metal detector scanning, those coded “No” did not have full-time metal detector scanning (i.e., none, random, hand wands). The vertical red line marks the SY2014-15, the last year before NYCDOE lifted their ban.

Figure 3.4: School Culture (Teacher) by Scanning Status, SY2012-2016



Notes: [left] School Safety (Teachers) [right] Job Satisfaction (Teachers). All measures calculated from teacher responses to annual questions on the NYCDOE School Survey. School Survey data cleaned and linked to an author-constructed metal-detector panel from NYPD SSA incident data. Schools coded “Yes” had full-time metal detector scanning, those coded “No” did not have full-time metal detector scanning (i.e., none, random, hand wands). The vertical red line marks the SY2014-15, the last year before NYCDOE lifted their ban.

Finally, for the two teacher-perceived school culture outcomes, one can see in Figure 3.4 that the parallel trend assumption for teacher job satisfaction is likely not met, even if we were to use the alternative SY2013-14 treatment cutoff as discussed above. There is a clear difference in the trend lines between the treatment and control groups prior to NYCDOE’s ban removal in March 2015. That said, trend lines appear to closely mirror one another for the teacher-based school safety outcome in the pre-treatment period.

Apart from violations in the parallel trend assumption for 2x2 DiD, there may be other relevant policy changes unrelated to school leadership changes that I do not control for and concurrently occurred with the NYCDOE cellphone ban removal. If these events positively or negatively impacted the school discipline and school culture outcomes, my estimates of the NYCDOE cellphone ban removal will be biased upward or downward, respectively. To this end, my results related to per-pupil suspensions should be viewed cautiously. In January 2014, Carmen Fariña was appointed as School Chancellor by Bill de Blasio, and in February 2015 made changes to the NYCDOE discipline code to further reduce school suspensions (Craig &

Martin, 2021). These changes and the public messaging around school discipline occurred concomitantly with NYCDOE's cellphone ban removal, and my results for per-pupil suspensions are likely biased upward in absolute magnitude. I am less concerned about the impact these policy changes had on measures of school culture, particularly when using the alternative (an likely more appropriate) SY2013-14 cutoff discussed above. While some of my school culture results may be magnified in later years, given these February 2015 policy changes, the original precipitous drop in my school culture measures occurred before this event.

3.5 Results and Discussion

3.5.1 Main Results

Table 3.4 and Table 3.5 show the impact estimates of the NYCDOE cellphone ban on student-level per-pupil suspensions per 100 students, student perceptions of school culture (i.e., respect, behavior, safety), and teacher perceptions of school culture (i.e., safety, job satisfaction). For each outcome, I provide 2x2 DiD regression point estimates and standard errors across each of the model specifications outlined above in Section 3.4. Starting with student-level per-pupil suspensions per 100 students, model specification (1) shows a -1.14 (SE = 0.58) reduction in the outcome after NYCDOE lifted its districtwide cellphone ban. The R² of the model is 0.09. Adding the model controls in the model specification (2) marginally reduces the point estimate to -1.17 (SE = 0.57) and gives an R² of 0.24. Finally, adding in the leadership change indicator variable, model specification (3) shows a -0.99 (SE = 0.57) reduction in per-pupil suspensions (per 100 students). The R² of this final model remains nearly identical to model specification (2). These impact estimates are statistically significant for all model specifications at $\alpha = 0.05$. These results show that, on average, per-pupil school suspensions per 100 students fell after NYCDOE lifted its cellphone ban and indicate that the ban, while in place, led to more

school suspensions.

Next, let us look school culture from students' perspectives, including the dimensions of respect, behavior, and school safety. Starting with the dimension of respect, all model specifications show a slightly negative and statistically insignificant impact on the outcome. Moving to the school culture dimension of behavior, specification (1) shows a statistically insignificant -0.036 (SE = 0.04) SD unit decrease in the composite measure of student-perceived behavior after NYCDOE lifted its districtwide cellphone ban. The R² of the model is 0.06. Adding the model controls in the model specification (2) reduces the 2x2 DiD point estimate to -0.08 (SE = 0.04) SD units. Notably, the R² of the model increases to 0.34. Finally, adding in the leadership change indicator variable, model specification (3) results are nearly identical to those produced by model specification (2). The impact estimate is statistically significant for the latter two model specifications at $\alpha = 0.10$ and indicate that lifting NYCDOE's cellphone ban had a statistically significant negative impact on student-perceived behavior in NYCDOE middle and high schools. Next, I evaluate the impact of a standardized measure of student-perceived safety. Model specification (1) provides a -0.16 (SE = 0.05) SD unit decrease in the outcome after NYCDOE lifted its districtwide cellphone ban. The R² of the model is 0.07. Adding the model controls in the model specification (2) further reduces the 2x2 DiD point estimate to -0.20 (SE = 0.05) and increases the model R² to 0.32. Adding the policy change indicator variable, model specification (3) provides nearly identical results to model specification (2). For all model specifications, these impact estimates are statistically significant at $\alpha = 0.01$ and indicate that removing the NYCDOE cellphone ban negatively impacted students' perceptions of school safety.

Table 3.4: Impact of NYCDOE Cellphone Ban on Suspensions and School Culture (Students)

	Per-Pupil Suspensions			School Respect (Students)			School Behavior (Students)			School Safety (Students)		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Cellphone Ban	-1.14** [0.577]	-1.17** [0.567]	-0.99* [0.568]	-0.012 [0.046]	-0.056 [0.048]	-0.054 [0.049]	-0.036 [0.043]	-0.078* [0.044]	-0.076* [0.044]	-0.16*** [0.046]	-0.20*** [0.048]	-0.20*** [0.049]
% FRPL Eligible		X	X		X	X		X	X		X	X
% ELL Students		X	X		X	X		X	X		X	X
% SWD Students		X	X		X	X		X	X		X	X
% Race Covariates		X	X		X	X		X	X		X	X
% Female		X	X		X	X		X	X		X	X
Enrollment		X	X		X	X		X	X		X	X
Leadership Changes		X				X			X			X
Sample Size	3,945	3,945	3,895	3,945	3,945	3,895	3,945	3,945	3,895	3,945	3,945	3,895
R2	0.086	0.235	0.235	0.017	0.246	0.254	0.057	0.341	0.344	0.072	0.324	0.331

Notes: Model (1) is the baseline DiD model specification, Model (2) includes school-level characteristics, and Model (3) adds an indicator variable for principal changes at the time of treatment. Per-Pupil suspensions are based on NYCDOE student suspension data aggregated at the school level and the sum of principal and superintendent suspensions for an entire school year. The school culture measures are derived from NYCDOE school-level School Survey data. I use robust standard errors clustered at the school-level. Race covariates include: (1) Asian, (2) Black, (3) Hispanic, (4) White. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 3.5: Impact of NYCDOE Cellphone Ban on School Culture (Teachers)

	School Safety (Teachers)			Job Satisfaction (Teachers)		
	(1)	(2)	(3)	(1)	(2)	(3)
Cellphone Ban	-0.038 [0.059]	-0.078 [0.060]	-0.100* [0.060]	0.013 [0.069]	-0.017 [0.069]	-0.039 [0.070]
% FRPL Eligible		X	X		X	X
% ELL Students		X	X		X	X
% SWD Students		X	X		X	X
% Race Covariates		X	X		X	X
% Female		X	X		X	X
Enrollment		X	X		X	X
Leadership Changes			X			X
Sample Size	3,945	3,945	3,895	3,945	3,945	3,895
R2	0.016	0.172	0.178	0.005	0.079	0.086

Notes: Model (1) is the baseline DiD model specification, Model (2) includes school-level characteristics, and Model (3) adds an indicator variable to account for principal changes at the time of treatment. The school culture measures are derived from NYCDOE school-level School Survey data. I use robust standard errors clustered at the school-level. Race covariates include: (1) Asian, (2) Black, (3) Hispanic, (4) White. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Now let us look at the impact the ban removal had on school culture through the lens of teachers (Table 3.5). Starting with teacher-perceived school safety, model specification (1) shows a -0.04 (SE = 0.06) SD unit reduction in the outcome after NYCDOE lifted its districtwide cellphone ban. The R² of the model is 0.02. Adding the model controls in the model specification (2) decreases the 2x2 DiD point estimate to -0.08 (SE = 0.06) SD units and gives an R² of 0.17. Finally, adding in the leadership change indicator variable, model specification (3) shows a -0.10 (SE = 0.06) SD unit decrease in teacher-perceived school safety. The R² of this final model is 0.18. Only the final model specification (3) produces statistically significant results at $\alpha = 0.10$ and indicates that the NYCDOE cellphone ban removal marginally impacted teacher-perceived school safety. Concluding with teacher-perceived job satisfaction, one can see small and statistically insignificant results for all model specifications. These findings suggest that the NYCDOE cellphone ban removal had no impact on how satisfied teachers were in their jobs in the preceding school years.

3.5.2 Discussion

Overall, these results show a mixed impact of removing NYCDOE's cellphone ban. On the one hand, the 2x2 DiD results indicate that lifting the districtwide ban had a notable and statistically significant impact on school discipline, student behavior (as perceived by students), and school safety (as perceived by students). On the other, I find null impacts of the ban removal on teacher satisfaction and student respect (as perceived by students) and marginal impact on school safety as perceived by teachers. How can these results be reconciled, and what mechanisms could be driving these findings? To begin, the school discipline results might be explained, in part, by the fact that teachers and school leaders were no longer obligated to enforce restrictive cellphone ban measures after NYCDOE lifted its districtwide ban in March

2015. Evidence from my researcher-led survey of NYCDOE parent coordinators and school leaders shows that multiple violations of a school's cellphone policy could lead to suspension. While this was one of the most punitive forms of discipline applied to students with multiple violations of their school's cellphone policy, it may be one driver of the school discipline results depending on how many schools had rigid and punitive school discipline policies cellphones.⁷⁵

Next, one might also predict an improvement in teacher job satisfaction, holding all else equal, in the face of improved working conditions stemming from reduced responsibilities surrounding cellphone policing and ban enforcement in those school years following the NYCDOE ban removal. While my results do not support this hypothesis, a different narrative emerges if we consider the general equilibrium effects of the ban removal on teacher satisfaction and school safety. For example, if lifting NYCDOE's cellphone ban also had adverse effects on school safety that led to counterbalancing negative impacts on teacher job satisfaction, one would predict null teacher satisfaction effects of the ban removal. This paper's null results for teacher job satisfaction and statistically significant negative results for student-perceived school safety support this latter hypothesis.

Finally, this paper's school discipline findings could alternatively be explained by a cumulative-effects model, where small but frequent negative teacher-student interactions over time manifest in student non-compliance and overly punitive disciplinary actions by school leaders. For example, consider a scenario where teachers' and principals' frequent policing of students' cellphones throughout an academic year erodes baseline levels of teacher-student and principal-student goodwill. Once goodwill crosses a lower limit, students act out more and are subsequently disciplined for violations of school policies related to an unrelated to cellphone use

⁷⁵ My researcher-led survey of NYC public middle and high school principals and parent coordinators had a meager response rate. I am hesitant to make broad-sweeping claims on such a small sample of respondents.

that ultimately stem from a school's aggressive and frequent cellphone ban enforcement. In this model, once the catalyst of the negative and frequent teacher-student interactions is removed, so will the negative outcomes that stem from those interactions. In the context of this paper, this cumulative-effects model predicts fewer negative teacher-student interactions and fewer disciplinary events after NYCDOE removed its districtwide cellphone ban. While this paper does not have direct measures of the quality of teacher-student or principal-student interactions, it does have direct measures of disciplinary outcomes in the form of suspensions, which supports the above hypothesis. However, one might expect a measure of teacher job satisfaction and student-perceived respect to increase, too. If so, this paper's null results for both of these outcomes work against this hypothesis and suggest an alternative mechanism.

Next, it is worth highlighting that the 2x2 DiD point estimates for school discipline, and student perceptions of student behavior and safety are directionally negative and statistically significant. At first glance, these results may seem counterintuitive, particularly if one views a reduction in school suspensions as a signal of a safer school environment. However, decreases in school suspensions could also be driven by systematic changes in schools' disciplinary code, where more leniency is granted, and once punishable offenses are no longer met with punitive measures. A deterrence model, where the threat of discipline deters a targeted action or set of actions, predicts an increase in an adverse action or set of actions if there is no longer a cost to that action. In the context of this paper, if that now unpunishable negative action (i.e., cellphone usage) directly impacts school safety, one will also expect a decrease in the reported safety of that environment. This paper's results support this hypothesis.

Absent from this discussion is exactly how in-school cellphone use influences school safety and student behavior and why one would expect these outcomes to be made worse (or

better) when moving from a school environment where cellphone access is restricted to one where cellphones are readily accessible. One explanation could be tied to cellphone facilitation and encouragement of bullying and cyberbullying, where negative student behavior is magnified by access to cellphones which in turn leads to unsafe school environments. To better understand what drives school safety and student behavior results, I provide 2x2 DiD estimates and standard errors for each school outcome included in the student-perceived school safety and student behavior indexes. Recall, these indexes are composed of normalized outcomes by school year with 0 mean and 1 standard deviation for the following school culture elements, including: (1) safety in classrooms, (2) safety outside near school grounds, (3) safety in hallways, bathrooms, locker rooms, and school cafeterias, (4) use of drugs and alcohol, (5) physical fights, (6) bullying, and (7) gang activity. To explore this, I use model specification (3) to estimate the impact of NYCDOE's cellphone ban removal on each of these outcomes. Table 3.6 captures the results from this analysis. Looking at the school safety outcomes, I find a notable decrease in the student-perceived school safety, including inside the classroom (-0.20 SD, SE = 0.053), outside on school grounds (-0.17, SE = 0.047), and in non-classroom school spaces (i.e., hallways, locker rooms, cafeterias) (-0.187, SE = 0.048). All estimates are statistically significant.

Next, I also found an increased frequency of troublesome student behavior after NYCDOE removed its districtwide ban. This is specifically for student perceptions of in-school bullying. Looking at Table 3.6, one can see that NYCDOE's removal of their cellphone ban worsened students' perception of the amount of harassment, bullying, and intimidation that occurred in their schools (-0.095 SD, SE = 0.044). This result is statistically significant at $\alpha = 0.05$. My findings also show impaired perceptions of in-school gang activity, fighting, and drug use. However, these results are not statistically significant.

Table 3.6: Impact of NYCDOE Cellphone Ban on School Safety and Behavior by Individual Outcome (Student)

	Safe Classroom (1)	Safe Outside (2)	Safe Hallways (3)	Drug Use (4)	Fights (5)	Bully (6)	Gangs (7)
Cellphone Ban	-0.201*** [0.053]	-0.174*** [0.047]	-0.187*** [0.048]	-0.07 [0.043]	-0.055 [0.039]	-0.095** [0.044]	-0.035 [0.047]
% FRPL Eligible	X	X	X	X	X	X	X
% ELL Students	X	X	X	X	X	X	X
% SWD Students	X	X	X	X	X	X	X
% Race Covariates	X	X	X	X	X	X	X
% Female	X	X	X	X	X	X	X
Enrollment	X	X	X	X	X	X	X
Leadership	X	X	X	X	X	X	X
Changes							
Sample Size	3,895	3,895	3,895	3,895	3,895	3,895	3,895
R ²	0.321	0.32	0.291	0.267	0.392	0.261	0.408

Notes: This table provides 2x2 DiD regression estimates for the individual standardized outcomes contained in the student-based school safety and school behavior indexes. Each model (1) – (7) uses the 2x2 DiD regression model specification (3) which contains school-level characteristics and an indicator for leadership changes. The school culture measures are derived from NYCDOE school-level School Survey data. I use robust standard errors clustered at the school-level. Race covariates include: (1) Asian, (2) Black, (3) Hispanic, (4) White. *** p < 0.01, ** p < 0.05, * p < 0.1.

3.5.3 Sensitivity Analysis

3.5.3.1 Pre-Treatment Cutoffs

As discussed in Section 3.4 above, there is a clear violation of the parallel trend assumption for some of the school culture indexes and many of the individual student-based school culture outcomes. However, the SY2014-15 pre-treatment cutoff I use to evaluate the 2x2 DiD parallel trend assumption and to derive the 2x2 DiD main results of this paper may be inappropriate given that students and teachers often have until late Spring to submit their survey responses to the NYCDOE (e.g., June 26, 2020, for SY2019-20 School Survey; June 11, 2021, for the SY2020-21 School Survey).⁷⁶ If the NYC School Survey results for SY2014-15 reflect, on the whole, students' and teachers' experiences in schools after NYCDOE removed their ban on March 2, 2015, the 2x2 DiD estimates and the evaluation of the 2x2 DiD parallel trend assumption using the SY2014-15 cutoff will be incorrect.

Before testing the sensitivity of the school culture results to the alternative SY2013-14 pre-treatment cutoff, I first reevaluate the 2x2 DiD parallel trend assumption using this new cutoff. I find that the trend lines of each student-based school culture index by treatment status are similar and appear to satisfy the parallel trend assumption. This is also the case for nearly all individual school culture measures that capture students' perceptions of student behavior and school safety. Finally, the parallel trend assumption looks to be met for the teacher-based school safety outcome but not for teacher job satisfaction which still shows pre-treatment trend lines that differ by treatment status. As before, this can be seen by looking at Figure 3.2-Figure 3.4 but instead of using SY2014-15 as the treatment marker, use SY2013-14. Given these findings, if

⁷⁶ I was able to find submission deadlines for a few earlier versions of the NYC School Survey as well. For example, the SY2006-07 NYC School Survey for teachers needed to be returned by 05/18/2007. Although I do not have submission deadlines for every school year in my panel, the evidence I do have strongly hints at a consistent late Spring deadline for the NYC School Survey regardless of the school year.

SY2013-14 is the correct pre-treatment cutoff, one can be more confident that the 2x2 DiD results (except for teacher job satisfaction) derived using this alternative SY2013-14 pre-treatment cutoff are correct.

Table 3.7. Impact of NYCDOE Cellphone Ban on School Culture, Cutoff Comparison

	Pre-treatment cutoff		DiD parallel trends?	
	SY2013-14	SY2014-15	SY2013-14	SY2014-15
<i>Indexes</i>				
School Respect (Students)	-0.085* [0.050]	-0.054 [0.049]	Y	Y
School Behavior (Students)	-0.092** [0.047]	-0.076* [0.044]	Y	Y
School Safety (Students)	-0.327*** [0.049]	-0.20*** [0.049]	Y	N
School Safety (Teachers)	-0.134** [0.064]	-0.100* [0.060]	Y	Y
Job Satisfaction (Teachers)	-0.093 [0.075]	-0.039 [0.070]	N	N
<i>School Safety Measures (Students)</i>				
Safe Classroom	-0.396*** [0.054]	-0.201*** [0.053]	Y	N
Safe Outside	-0.213*** [0.048]	-0.174*** [0.047]	Y	Y
Safe Hallways	-0.333*** [0.050]	-0.187*** [0.048]	Y	N
<i>Student Behavior Measures (Students)</i>				
Drug Use	-0.064 [0.046]	-0.07 [0.043]	Y	Y
Fights	-0.110*** [0.040]	-0.055 [0.039]	Y	N
Bully	-0.163*** [0.046]	-0.095** [0.044]	Y	N
Gangs	0.029 [0.048]	-0.035 [0.047]	Y	N

Notes: This table provides 2x2 DID regression estimates for the student and teacher-based school culture indexes in addition to the individual standardized outcomes contained in the student-based school safety and school behavior indexes. Each model uses the 2x2 DiD regression model specification (3) which contains school-level characteristics and an indicator for leadership changes. cellphone ban. The school culture measures are derived from NYCDOE school-level School Survey data. I use robust standard errors clustered at the school-level. Race covariates include: (1) Asian, (2) Black, (3) Hispanic, (4) White. *** p < 0.01, ** p < 0.05, * p < 0.1.

Table 3.7 shows the impact estimates of the NYCDOE cellphone ban on student perceptions of school culture (i.e., respect, behavior, safety), and teacher perceptions of school culture (i.e., safety, job satisfaction) using the updated SY2013-14 pre-treatment cutoff. I provide 2x2 DiD regression point estimates and standard errors for each outcome for my preferred model specification (3), as outlined above in Section 3.4. Starting with student perceptions of school culture, I find similar results; however, the 2x2 DiD point estimates are larger in absolute magnitude. For student perceptions of school respect (-0.085 SD, SE = 0.05), student behavior (-0.092 SD, SE = 0.047), and school safety (-0.327 SD, SE = 0.049), I find the NYCDOE cellphone ban removal had a negative impact. That is, students reported a school environment in which they felt less safe, there was less respect, and worse student behavior in the wake of NYCDOE lifting their districtwide cellphone ban. The result for the student respect index is statistically significant at $\alpha = 0.10$, for student behavior index at $\alpha = 0.05$, and for school safety index at $\alpha = 0.01$. Moving to teacher perceptions, one can see that the results for teacher job satisfaction remain the same as before where I find no evidence that NYCDOE's cellphone ban impacted the outcome. However, the results for teacher perceptions of school safety did change, where I now report negative and statistically significant impacts of the ban removal (-0.134, SE = 0.064). The sign and magnitude of the point estimate are similar to before, albeit slightly larger in absolute value, and the results using the alternative cutoff are still statistically significant.

Finally, Table 3.7 also provides updated 2x2 DiD estimates for the disaggregated student-perceived school safety index. As before, all school safety measures are negative and statically significant at $\alpha = 0.01$. The size of each 2x2 DiD estimate is noticeably larger in absolute terms relative to the original estimates. For the individual student-behavior outcomes, I still find that the ban removal had a negative and statistically significant impact on the outcome of school

bullying (-0.163, SE = 0.046); however, I now find a negative and statistically significant impact on school fights (-0.110, SE = 0.040) where before the result was null. Finally, the null results for drug use and gangs are nearly identical. Overall, changing the cutoff from SY2014-15 to SY2013-14 doesn't materially change the story of my findings, although impacts are notably larger using the latter cutoff compared to the former and the parallel trend assumption is more likely to hold.

3.5.3.2 Extended Panel

Up to this point, this paper has evaluated NYCDOE's policy to remove their districtwide cellphone ban using a relatively short panel covering SY2012 to SY2016. As discussed above in the earlier "Data" section, I limited my panel to these school years to: (1) avoid the material change NYCDOE made to their school discipline policy enacted in the Fall 2012 and (2) have full panel coverage of my chosen school culture outcomes. An additional undiscussed reason for my original SY2012 to SY2016 choice is the limitation of my school-level metal detector panel, which begins SY2013-14. Although full-time metal detector status is relatively stable across time in the short term, this may not be the case in the longer term. That said, given I have data from SY2006-SY2016 for most of my outcomes,⁷⁷ I run additional 2x2 DiD models to see how, if at all, my results change using a more extensive pre-treatment period. One note, if a school's full-time metal detector status is not stable across this extended panel, these results of this robustness check will be wrong and should be discarded.

First, I once again provide time series plots of each outcome across time by treatment status (Figure C.1-C.4). Most of the original outcomes I evaluated before are available, including per-pupil suspensions (per 100 students), indices for students' perceptions of student behavior

⁷⁷ Sans a few school culture outcomes (i.e., teacher job satisfaction, student perceptions of respect) that I exclude from this exercise due to missingness and lack of coverage across the extended SY2006-SY2016 panel.

and school safety, individual measures of students' perception of school safety (i.e., safe classroom, safe outside, safe hallway), and teachers' perception of school safety. Due to panel limitations, I omit teachers' perception of job satisfaction and the index for students' perception of respect. I also omit individual measures of student behavior given the lack of trend break by treatment status at the pre-treatment cutoff.⁷⁸ Evaluating the 2x2 DiD parallel trend assumption for each of the included outcomes, I see that most outcomes should satisfy it except for per-pupil suspensions (per 100 students) and possibly the index of student behavior. As before, violations of this assumption will lead to biased estimates of the ATET using 2x2 DiD, and estimates should be interpreted with caution and used with care, if at all.

Appendix Table C.2 shows the 2x2 DiD impact estimates of the NYCDOE cellphone ban on student perceptions of school culture (i.e., behavior, safety) and teacher perceptions of school culture (i.e., safety) using the extended panel. For the per-pupil suspension outcome, I use the SY2014-15 pre-treatment cutoff, and for the school culture outcomes I use the SY2013-14 pre-treatment cutoff. As before, for each outcome, I provide 2x2 DiD regression point estimates and standard errors across each of the model specifications outlined above in Section 3.4. The results for per-pupil total suspensions (per 100 students) and the index comprised of students' perception of student behavior are null for all model specifications. This contrasts with the statistically significant results I found in the main results using SY2012-SY2016 panel. Notably, the parallel trend assumption likely fails for both outcomes as well. Moving to student perceptions of school safety, I find similar results as before; however, the 2x2 DiD point estimates are slightly smaller in absolute magnitude when compared to those in Table 8 where SY2013-14 pre-treatment cutoff is used. All 2x2 DiD impact estimates are statistically

⁷⁸ Running the 2x2 DiD model for the index of student behavior and for each of the individual student-behavior outcomes yields non-statistically significant point estimates.

significant at $\alpha = 0.01$. Finally, for teachers' perception of school safety, I also find statistically significant results across all model specifications with point estimates of similar size and magnitude when compared to those in Table 3.7 where SY2013-14 pre-treatment cutoff is used. Finally, I break down the index of students' perception of school safety and report the 2x2 DiD impact estimates from model specification (3) in Table C.3. Compared to the original SY2012-SY2016 panel, I find similar negative statistically significant results; however, the point estimates are smaller in absolute magnitude when compared to those in Table 3.7 where SY2013-14 pre-treatment cutoff is used. Notably, for each of these outcomes (indexes and individual outcomes) the parallel trend assumption likely holds and provides more robust evidence of the negative impacts of NYCDOE's ban removal on students' and teachers' perception of school safety.

3.6 Conclusion

This paper investigates the impact of lifting cellphone bans in NYCDOE public middle and high schools on student discipline and school culture, both of which have been found to play an important role in educational productivity (Craig & Martin, 2021). I combine student-level administrative data with publicly available school-level data on school safety and an author-constructed panel documenting full-time metal detector schools in NYCDOE. Using schools' full-time metal detector status as a proxy for school-level cellphone ban enforcement and exploiting variation in NYCDOE's districtwide cellphone ban across time, my results indicate that lifting NYCDOE's cellphone ban positively impacted school discipline, driving down per-pupil total suspensions (per 100 students) in post-ban school years. This result could be driven by reduced disciplinary obligations for school staff in post-ban periods and fewer teacher-student enforcement events for cellphone violations in post-ban school years. That said, these results

should be taken with a grain of salt, given the concurrent changes in NYC's disciplinary code in February of 2015. In addition, I find the removal of NYCDOE's cellphone ban negatively impacts student perceptions of school culture, including school safety and student behavior. I find no evidence that the ban affected teacher perceptions of school culture (i.e., job satisfaction, school safety) or student perceptions of teacher-to-student respect and peer-to-peer respect. Robustness checks using an extended panel provide additional evidence of the negative impacts of the NYCDOE's removal of their districtwide ban on school culture outcomes, particularly outcomes related to student perceptions of school safety.

The present literature related to the causal effects of technology in the classroom provides evidence that technology can both enhance and hurt student learning and performance (e.g., Angrist & Lavy, 2002; Shapley et al., 2009; Banerjee et al., 2007; Rouse & Krueger, 2004; Campuzano et al., 2009). For cellphones, apart from a few notable studies that indicate negative impacts of usage on high-stakes exams (Beland & Murphy, 2016) and long-term retention (Glass & Kang, 2018), there is a dearth of rigorous empirical evidence directly addressing if and how to ban's impact student discipline and school culture. As the use of smartphones and other computing devices proliferates in our lives and schools, it is vital to understand if and how necessary measures of school and student well-being are impacted and whether the secondary effects of efforts to curb usage overwhelm or, at a minimum, act as a countervailing force to the benefits of such policies.

While this paper's findings do not definitively answer all these questions, they do indicate that strict enforcement policies surrounding cellphones could be counterproductive to school's goals of limiting school suspensions, mainly if school discipline policies linked to violations in cellphone usage are overly burdensome (i.e., classroom or school removal). In

addition, consideration should be given to the impact school policies may have on school culture and the type of interactions, positive and negative, said policies create between teachers and students and between students and their peers. Importantly, these findings do not dismiss the possibility of creative solutions to school cellphone policies and ban enforcement that effectively limit students' off-task cellphone use while limiting the regularity of negative teacher-student interactions due to code violations.⁷⁹

⁷⁹ Students also have a role in regulating their cellphone use and abiding by school-level policies. That said, students are likely to be more compliant with school policies, even those they believe to be suboptimal, if they have a role in policy design. While not realistic for many school policies, co-opting students to tackle the challenge of cellphone bans and usage could be fruitful.

Appendix A: Additional Tables and Figures from Chapter

Note: The “Online Appendix” (S1-S20; Table A1) included in the published version of this paper is provided as a supplemental document (i.e., LUKES-DISSERTATION-SUPPLEMENT-2022) to this dissertation.

Appendix B: Additional Tables and Figures from Chapter 2

Table B.1: Diversity Sample Overview: All 2018-19 U.S. Public Schools vs. HOLC Sample

	U.S. Census Bureau Region				
	Nationwide	Midwest	Northeast	South	West
<i>Race: Percent Black</i>	15.1%	13.6%	13.9%	22.8%	4.7%
HOLC A-D Sample	29.2%	37.2%	28.8%	40.0%	8.8%
Share	193%	274%	207%	175%	187%
<i>Race: Percent Hispanic</i>	27.1%	13.0%	22.0%	27.0%	42.7%
HOLC A-D Sample	37.3%	25.9%	36.4%	37.2%	57.8%
Share	138%	199%	165%	138%	135%
<i>Race: Percent White</i>	47.1%	64.4%	53.1%	42.0%	36.8%
HOLC A-D Sample	22.1%	27.9%	21.0%	18.1%	18.9%
Share	47%	43%	40%	43%	51%
<i>Urban</i>	26,070	5,333	3,791	9,092	7,854
HOLC A-D Sample	7,558	2,370	2,579	1,336	1,273
Share	29%	44%	68%	15%	16%
<i>Suburban</i>	30,355	6,712	7,081	9,278	7,284
HOLC A-D Sample	2,133	656	1,011	109	357
Share	7%	10%	14%	1%	5%
<i>Charter Schools</i>	7,340	1,473	746	2,420	2,701
HOLC A-D Sample	1,591	585	420	259	327
Share	22%	40%	56%	11%	12%
<i>Schools</i>	95,432	24,669	14,775	33,040	22,948
HOLC A-D Sample	9,709	3,031	3,595	1,453	1,630
Share	10%	12%	24%	4%	7%
<i>Districts</i>	17,741	5,913	3,640	3,826	4,362
HOLC A-D Sample	1,955	666	726	195	368
Share	11%	11%	20%	5%	8%
<i>CBSAs</i>	934	291	91	375	177
HOLC A-D Sample	141	51	29	46	15
Share	15%	18%	32%	12%	8%

Notes: First line of each variable includes all active 2018-19 U.S. public primary and secondary schools. Only those schools with non-missing and non-zero student enrollments included. The second line represents all 2018-19 U.S. public primary and secondary schools for the HOLC A-D “Student Diversity” analysis sample. Schools were only included in this sample if they were matched to 1935-1940 HOLC A-D maps. Shares represent match rates with HOLC A-D maps for “Schools”, “Districts” and “CBSAs” variables and simply sample representativeness for all other variables.

Table B.2: SEDA Sample Overview: All 2017-18 U.S. Public Districts vs. HOLC Sample

	U.S. Census Bureau Region				
	Nationwide	Midwest	Northeast	South	West
<i>Race: Percent Black</i>	15.2%	13.7%	13.9%	23.0%	4.8%
HOLC A-D Sample	28.9%	35.9%	28.6%	39.7%	7.7%
Share	190%	262%	206%	173%	160%
<i>Race: Percent Hispanic</i>	26.7%	12.7%	21.5%	26.5%	42.5%
HOLC A-D Sample	37.8%	27.9%	35.0%	38.1%	60.4%
Share	142%	220%	163%	144%	142%
<i>Race: Percent White</i>	47.6%	64.9%	53.9%	42.6%	37.1%
HOLC A-D Sample	22.2%	27.4%	22.3%	17.7%	17.1%
Share	47%	42%	41%	42%	46%
<i>Urban</i>	25,964	5,335	3,788	9,038	7,803
HOLC A-D Sample	3,847	1,432	999	799	617
Share	15%	27%	26%	9%	8%
<i>Suburban</i>	30,335	6,715	7,097	9,264	7,259
HOLC A-D Sample	1,272	390	590	64	228
Share	4%	6%	8%	1%	3%
<i>Charter Schools</i>	7,158	1,474	723	2,337	2,624
HOLC A-D Sample	505	219	155	101	30
Share	7%	15%	21%	4%	1%
<i>Schools</i>	95,242	24,643	14,812	32,940	22,847
HOLC A-D Sample	5,124	1,826	1,589	864	845
Share	5%	7%	11%	3%	4%
<i>Districts</i>	16,782	5,926	3,653	3,809	3,394
HOLC A-D Sample	1,006	393	397	127	89
Share	6%	7%	11%	3%	3%
<i>CBSAs</i>	934	291	91	375	177
HOLC A-D Sample	137	50	27	45	15
Share	15%	17%	30%	12%	8%

Notes: First line of each variable includes all active 2017-18 U.S. public primary and secondary schools. Only those schools with non-zero or non-missing total student enrollment data included in sample. The second line represents all 2017-18 U.S. public primary and secondary schools for the HOLC A-D analysis sample. Schools were only included in this sample if they were matched to 1935-1940 HOLC A-D maps. Shares represent match rates for “Schools” and “CBSAs” variables and sample representativeness for all others.

Table B.3: Finance Sample Overview: All 2017-18 U.S. Public Districts vs. HOLC Sample

	U.S. Census Bureau Region				
	Nationwide	Midwest	Northeast	South	West
<i>Race: Percent Black</i>	15.1%	13.7%	14.1%	23.0%	4.8%
HOLC A-D Sample	24.9%	28.5%	22.4%	34.3%	7.9%
Share	165%	208%	159%	149%	165%
<i>Race: Percent Hispanic</i>	26.7%	12.7%	21.6%	26.5%	42.4%
HOLC A-D Sample	33.0%	20.6%	27.0%	32.1%	53.3%
Share	124%	162%	125%	121%	126%
<i>Race: Percent White</i>	47.6%	64.8%	53.7%	42.5%	37.2%
HOLC A-D Sample	31.7%	41.1%	39.0%	26.8%	22.3%
Share	67%	63%	73%	63%	60%
<i>Urban</i>	2,796	824	636	655	681
HOLC A-D Sample	822	419	190	137	76
Share	29%	51%	30%	21%	11%
<i>Suburban</i>	3,877	1,223	1,503	488	663
HOLC A-D Sample	893	364	404	45	80
Share	23%	30%	27%	9%	12%
<i>Charter Schools</i>	3,762	1,090	704	788	1,180
HOLC A-D Sample	726	368	174	108	76
Share	19%	34%	25%	14%	6%
<i>Districts</i>	16,799	5,924	3,668	3,812	3,395
HOLC A-D Sample	1,760	806	608	187	159
Share	10%	14%	17%	5%	5%
<i>CBSAs</i>	931	291	91	375	174
HOLC A-D Sample	144	53	30	46	15
Share	15%	18%	33%	12%	9%

Notes: First line of each variable includes all active 2017-18 U.S. public primary and secondary districts. Only those schools with non-zero or non-missing total student enrollment data included in sample. The second line represents all 2017-18 U.S. public primary and secondary districts for the HOLC A-D analysis sample. Districts were only included in this sample if they were matched to 1935-1940 HOLC A-D maps. Shares represent match rates for “Districts” and “CBSAs” variables and sample representativeness for all others.

Table B.4: HOLC A-D Means, SEs and Differences, Full vs. Panel, Finance (USD 2018)

		Sample Comparisons			
		Full	Panel	Diff.	p-value
District PPR – Total, 2017-18 (\$USD)	A	\$16,938 [\$1,778]	\$17,059 [\$2,787]	-\$120	0.91
	B	\$17,662 [\$583]	\$17,738 [\$906]	-\$75	0.82
	C	\$15,435 [\$399]	\$15,526 [\$627]	-\$92	0.70
	D	\$14,453 [\$472]	\$14,269 [\$666]	\$184	0.63
District PPR – Federal, 2017- 18 (\$USD)	A	\$705 [\$133]	\$695 [\$208]	\$10	0.9
	B	\$942 [\$52]	\$932 [\$81]	\$10	0.74
	C	\$1,430 [\$49]	\$1,411 [\$74]	\$18	0.57
	D	\$1,136 [\$52]	\$1,088 [\$89]	\$47	0.24
District PPR – State, 2017-18 (\$USD)	A	\$4,498 [\$571]	\$4,481 [\$874]	\$17	0.95
	B	\$6,236 [\$424]	\$6,231 [\$655]	\$6	0.98
	C	\$7,272 [\$378]	\$7,281 [\$593]	-\$9	0.97
	D	\$7,514 [\$426]	\$7,315 [\$536]	\$199	0.59
District PPR – Local, 2017-18 (\$USD)	A	\$11,735 [\$1,610]	\$11,883 [\$2,535]	-\$147	0.87
	B	\$10,484 [\$604]	\$10,575 [\$937]	-\$91	0.79
	C	\$6,733 [\$296]	\$6,834 [\$465]	-\$101	0.57
	D	\$5,803 [\$300]	\$5,866 [\$519]	-\$63	0.79

Notes: Means are HOLC A-D weighted averages from regressions of respective finance outcomes on HOLC B-D indicators. Standard errors are Huber-White heteroskedasticity-consistent and are in brackets. Regressions are run using both the 2017-18 full sample and the 2017-18 data from the panel sample. The full sample is the original 2017-18 cross-sectional sample and includes all $d = 1,760$ districts. The panel sample is the time series sample spanning 1989-90 through 2017-18 school years and includes $d_p = 1,109$ districts. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.5: HOLC A-D Means, SEs and Differences, Full vs. Panel, Racial Diversity Outcomes

		Sample Comparisons			
		Full	Panel	Diff.	p-value
% Black	A	0.22 [0.013]	0.20 [0.015]	0.018	0.047**
	B	0.26 [0.008]	0.23 [0.009]	0.036	<0.001***
	C	0.27 [0.005]	0.23 [0.007]	0.037	<0.001***
	D	0.36 [0.007]	0.30 [0.01]	0.056	<0.001***
% White	A	0.46 [0.015]	0.48 [0.020]	-0.026	0.004***
	B	0.33 [0.008]	0.37 [0.011]	-0.042	<0.001***
	C	0.21 [0.005]	0.25 [0.008]	-0.041	<0.001***
	D	0.13 [0.004]	0.15 [0.008]	-0.023	<0.001***
% Non-White	A	0.54 [0.015]	0.52 [0.020]	0.026	0.004***
	B	0.67 [0.008]	0.63 [0.011]	0.042	<0.001***
	C	0.79 [0.005]	0.75 [0.008]	0.041	<0.001***
	D	0.87 [0.004]	0.85 [0.008]	0.023	<0.001***
Simpson's Diversity (1-D)	A	0.49 [0.011]	0.50 [0.014]	-0.006	0.329
	B	0.480 [0.006]	0.50 [0.008]	-0.018	<0.001***
	C	0.45 [0.004]	0.45 [0.007]	-0.005	0.246
	D	0.41 [0.005]	0.42 [0.009]	-0.007	0.295

Notes: Means are HOLC A-D weighted averages from regressions of respective diversity outcomes on HOLC B-D indicators. Standard errors are Huber-White heteroskedasticity-consistent and are in brackets. Regressions are run using both the 2017-18 full sample and the 2017-18 data from the panel sample. The full sample is the original 2017-18 cross-sectional sample and includes all $n = 9,709$ schools. The panel sample is the time series sample spanning 1988-89 through 2018-19 school years and includes $n_p = 4,677$ schools. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.6: HOLC A-D Averages Over Time, Full and Panel, Exposure Index, 1988-2019

		HOLC Security Rating				
		Overall	A	B	C	D
<i>White-Asian</i> ^a	1988-89	4.4%	4.1%	4.6%	4.4%	4.2%
	1998-99	5.4%	5.0%	5.7%	5.5%	4.8%
	2008-09	6.2%	5.6%	6.4%	6.5%	5.3%
	2018-19	6.8%	6.1%	7.1%	7.1%	6.0%
	2018-19 ^b	7.5%	6.5%	7.6%	8.0%	7.3%
<i>White-Black</i> ^a	1988-89	15.9%	13.0%	14.7%	14.9%	22.2%
	1998-99	16.3%	13.8%	15.3%	15.3%	23.1%
	2008-09	16.4%	13.9%	14.9%	16.3%	22.8%
	2018-19	13.8%	11.5%	12.5%	13.8%	19.5%
	2018-19 ^b	15.0%	11.5%	13.3%	14.9%	20.1%
<i>White-Hispanic</i> ^a	1988-89	8.9%	5.1%	6.4%	9.8%	13.3%
	1998-99	11.3%	6.5%	8.8%	12.6%	16.4%
	2008-09	14.5%	9.2%	11.7%	16.2%	21.4%
	2018-19	18.8%	12.3%	15.5%	21.0%	27.3%
	2018-19 ^b	19.9%	12.8%	16.3%	22.0%	25.8%
<i>White-Non-White</i> ^a	1988-89	29.8%	22.5%	26.3%	29.8%	40.5%
	1998-99	33.7%	25.8%	30.5%	34.1%	45.5%
	2008-09	38.4%	29.9%	34.3%	40.3%	51.1%
	2018-19	45.9%	36.0%	41.8%	48.5%	59.3%
	2018-19 ^b	48.7%	36.9%	43.6%	51.1%	59.3%

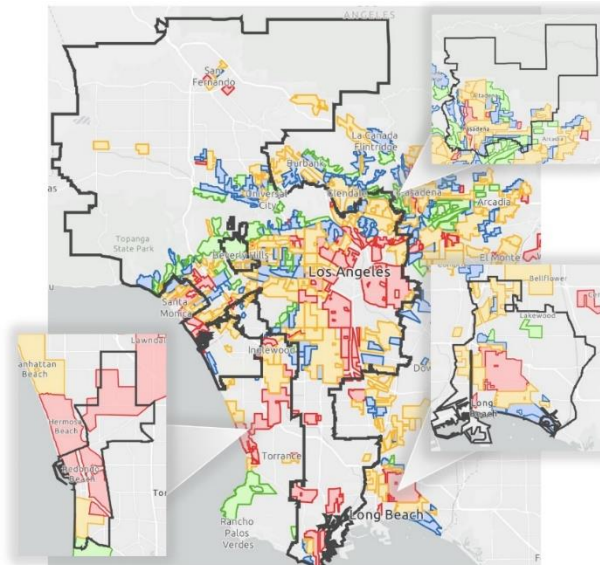
Notes: a – Signifies the student race pair used in the exposure index calculation. For example, White-Black provides the White-Black exposure index, while factoring in all other student race groups. Exposure index calculations do vary if all other race groups apart from the relevant race pair are ignored; b – Exposure index values derived using the full 2018-19 cross-sectional data set. All other within-group (i.e., White-Black) values across time derived from the 1988-89 through 2018-19 panel data set

Figure B.1: Los Angeles County Area Descriptions for Nos. A-1 and D-1

AREA DESCRIPTION																																																																																								
Security Map of Los Angeles County																																																																																								
1. POPULATION: a. Increasing <u>Rapidly</u> Decreasing <u>None</u> Static <u>Yes</u>																																																																																								
b. Class and Occupation <u>Motion picture stars, executives & technicians, professional and business men.</u> Income <u>\$3600-\$6000 & up.</u>																																																																																								
c. Foreign Families <u>0 %</u> Nationalities <u>---</u> d. Negro <u>0 %</u>																																																																																								
e. Shifting or Infiltration <u>None apparent</u>																																																																																								
2. BUILDINGS: <table border="1"> <thead> <tr> <th></th> <th>PREDOMINATING</th> <th>85 %</th> <th>OTHER TYPE</th> <th>15 %</th> </tr> </thead> <tbody> <tr> <td>a. Type and Size</td> <td><u>5, 6 & 7 rooms</u></td> <td></td> <td><u>Larger type</u></td> <td></td> </tr> <tr> <td>b. Construction</td> <td><u>Frame, stucco & masonry</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>c. Average Age</td> <td><u>2 years</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>d. Repair</td> <td><u>Good</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>e. Occupancy</td> <td><u>99%</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>f. Owner-occupied</td> <td><u>90%</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>g. 1935 Price Bracket</td> <td><u>\$ --</u></td> <td><u>% change</u></td> <td><u>\$</u></td> <td><u>% change</u></td> </tr> <tr> <td>h. 1937 Price Bracket</td> <td><u>\$ 5000-7500 & up</u></td> <td><u>%</u></td> <td><u>\$</u></td> <td><u>%</u></td> </tr> <tr> <td>i. 1939 Price Bracket</td> <td><u>\$ 5000-7500 & up</u></td> <td><u>%</u></td> <td><u>\$</u></td> <td><u>%</u></td> </tr> <tr> <td>j. Sales Demand</td> <td><u>Good</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>k. Predicted Price Trend (next 6-12 months)</td> <td><u>Static & up</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>l. 1935 Rent Bracket</td> <td><u>\$ No record</u></td> <td><u>% change</u></td> <td><u>\$</u></td> <td><u>% change</u></td> </tr> <tr> <td>m. 1937 Rent Bracket</td> <td><u>\$ of rentals</u></td> <td><u>%</u></td> <td><u>\$</u></td> <td><u>%</u></td> </tr> <tr> <td>n. 1939 Rent Bracket</td> <td><u>\$ --</u></td> <td><u>%</u></td> <td><u>\$</u></td> <td><u>%</u></td> </tr> <tr> <td>o. Rental Demand</td> <td><u>Good</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>p. Predicted Rent Trend (next 6-12 months)</td> <td><u>---</u></td> <td></td> <td></td> <td></td> </tr> </tbody> </table>					PREDOMINATING	85 %	OTHER TYPE	15 %	a. Type and Size	<u>5, 6 & 7 rooms</u>		<u>Larger type</u>		b. Construction	<u>Frame, stucco & masonry</u>				c. Average Age	<u>2 years</u>				d. Repair	<u>Good</u>				e. Occupancy	<u>99%</u>				f. Owner-occupied	<u>90%</u>				g. 1935 Price Bracket	<u>\$ --</u>	<u>% change</u>	<u>\$</u>	<u>% change</u>	h. 1937 Price Bracket	<u>\$ 5000-7500 & up</u>	<u>%</u>	<u>\$</u>	<u>%</u>	i. 1939 Price Bracket	<u>\$ 5000-7500 & up</u>	<u>%</u>	<u>\$</u>	<u>%</u>	j. Sales Demand	<u>Good</u>				k. Predicted Price Trend (next 6-12 months)	<u>Static & up</u>				l. 1935 Rent Bracket	<u>\$ No record</u>	<u>% change</u>	<u>\$</u>	<u>% change</u>	m. 1937 Rent Bracket	<u>\$ of rentals</u>	<u>%</u>	<u>\$</u>	<u>%</u>	n. 1939 Rent Bracket	<u>\$ --</u>	<u>%</u>	<u>\$</u>	<u>%</u>	o. Rental Demand	<u>Good</u>				p. Predicted Rent Trend (next 6-12 months)	<u>---</u>			
	PREDOMINATING	85 %	OTHER TYPE	15 %																																																																																				
a. Type and Size	<u>5, 6 & 7 rooms</u>		<u>Larger type</u>																																																																																					
b. Construction	<u>Frame, stucco & masonry</u>																																																																																							
c. Average Age	<u>2 years</u>																																																																																							
d. Repair	<u>Good</u>																																																																																							
e. Occupancy	<u>99%</u>																																																																																							
f. Owner-occupied	<u>90%</u>																																																																																							
g. 1935 Price Bracket	<u>\$ --</u>	<u>% change</u>	<u>\$</u>	<u>% change</u>																																																																																				
h. 1937 Price Bracket	<u>\$ 5000-7500 & up</u>	<u>%</u>	<u>\$</u>	<u>%</u>																																																																																				
i. 1939 Price Bracket	<u>\$ 5000-7500 & up</u>	<u>%</u>	<u>\$</u>	<u>%</u>																																																																																				
j. Sales Demand	<u>Good</u>																																																																																							
k. Predicted Price Trend (next 6-12 months)	<u>Static & up</u>																																																																																							
l. 1935 Rent Bracket	<u>\$ No record</u>	<u>% change</u>	<u>\$</u>	<u>% change</u>																																																																																				
m. 1937 Rent Bracket	<u>\$ of rentals</u>	<u>%</u>	<u>\$</u>	<u>%</u>																																																																																				
n. 1939 Rent Bracket	<u>\$ --</u>	<u>%</u>	<u>\$</u>	<u>%</u>																																																																																				
o. Rental Demand	<u>Good</u>																																																																																							
p. Predicted Rent Trend (next 6-12 months)	<u>---</u>																																																																																							
3. NEW CONSTRUCTION (next yr.) No. <u>250</u> Type & Price <u>5, 6 & 7 rms. \$5500-\$7500 & up</u> How Selling <u>Moderately</u>																																																																																								
4. OVERHANG OF HOME PROPERTIES: a. HOLC <u>1</u> b. Institutions <u>For</u>																																																																																								
5. SALE OF HOME PROPERTIES (3-yr.) a. HOLC <u>2</u> b. Institutions <u>For</u>																																																																																								
6. MORTGAGE FUNDS: <u>Ample</u> 7. TOTAL TAX RATE PER \$1000 (1937-) <u>\$54.25</u>																																																																																								
8. DESCRIPTION AND CHARACTERISTICS OF AREA:																																																																																								
Terrain: Level to rolling hillside. No construction hazard or flood threat. Land improved 30%. Highly deed restricted and protected from racial hazards. Conveniences are all readily available with exception of transportation which is as yet only fair. Street improvements are still in process of construction. This area, located on the southern rim of the San Fernando Valley, was subdivided some 15 years ago, and substantial street improvements were installed. The depression retarded development. Under the stimulus of well directed promotional effort and FHA financing a revival started some 4 years ago and it is now one of the "hot spots" in the Valley. Construction and maintenance are of high quality. Architectural designs are harmonious. Location is favorable and attractive. Population is homogeneous. Lots are generally of extra size and sold on "homesite" basis and at widely varying prices according to size and location. The pattern of the area is well established and it is accorded a "low green" grade.																																																																																								
9. LOCATION <u>Encino</u> SECURITY GRADE <u>1st</u> - AREA NO. <u>4-1</u> DATE <u>3/23/39</u>																																																																																								
CAUTION: This area is currently affected in whole or in part by an Ad valorem Tax District. Individual properties should be checked for this hazard.																																																																																								

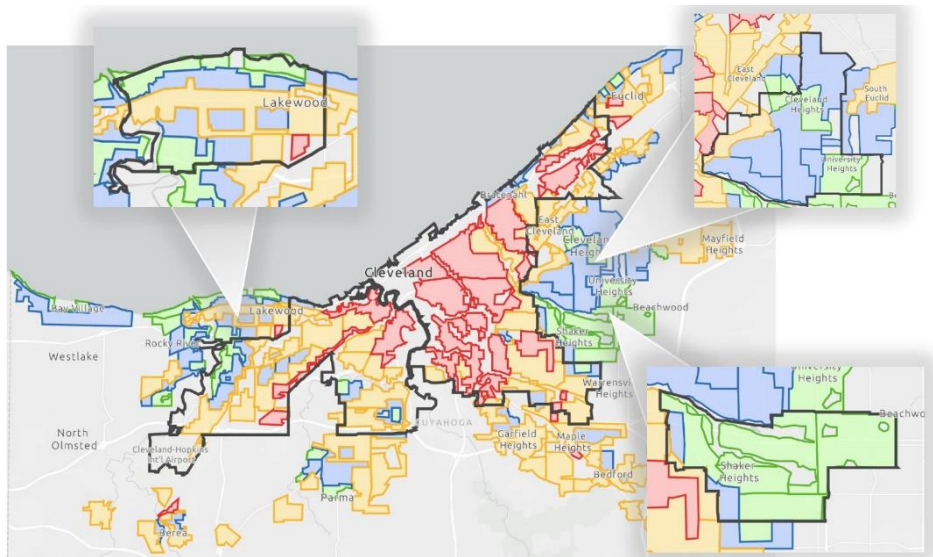
AREA DESCRIPTION																																																																																								
Security Map of Los Angeles County																																																																																								
1. POPULATION: a. Increasing <u>---</u> Decreasing <u>---</u> Static <u>Yes</u>																																																																																								
b. Class and Occupation <u>Agricultural, Cannery & Railroad laborers & WPA workers</u> Income <u>\$700-\$1200</u>																																																																																								
c. Foreign Families <u>80 %</u> Nationalities <u>Mexican predominating</u> d. Negro <u>0 %</u>																																																																																								
e. Shifting or Infiltration <u>None apparent</u>																																																																																								
2. BUILDINGS: <table border="1"> <thead> <tr> <th></th> <th>PREDOMINATING</th> <th>90 %</th> <th>OTHER TYPE</th> <th>%</th> </tr> </thead> <tbody> <tr> <td>a. Type and Size</td> <td><u>3-4 room shacks</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>b. Construction</td> <td><u>Shacks</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>c. Average Age</td> <td><u>Old</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>d. Repair</td> <td><u>Poor</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>e. Occupancy</td> <td><u>99%</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>f. Owner-occupied</td> <td><u>5%</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>g. 1935 Price Bracket</td> <td><u>\$500-1000</u></td> <td><u>% change</u></td> <td><u>\$</u></td> <td><u>% change</u></td> </tr> <tr> <td>h. 1937 Price Bracket</td> <td><u>\$700-1200</u></td> <td><u>%</u></td> <td><u>\$</u></td> <td><u>%</u></td> </tr> <tr> <td>i. 1939 Price Bracket</td> <td><u>\$700-1200</u></td> <td><u>%</u></td> <td><u>\$</u></td> <td><u>%</u></td> </tr> <tr> <td>j. Sales Demand</td> <td><u>Poor</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>k. Predicted Price Trend (next 6-12 months)</td> <td><u>Static</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>l. 1935 Rent Bracket</td> <td><u>\$5.00-12.50</u></td> <td><u>% change</u></td> <td><u>\$</u></td> <td><u>% change</u></td> </tr> <tr> <td>m. 1937 Rent Bracket</td> <td><u>\$7.50-15.00</u></td> <td><u>%</u></td> <td><u>\$</u></td> <td><u>%</u></td> </tr> <tr> <td>n. 1939 Rent Bracket</td> <td><u>\$7.50-15.00</u></td> <td><u>%</u></td> <td><u>\$</u></td> <td><u>%</u></td> </tr> <tr> <td>o. Rental Demand</td> <td><u>Good</u></td> <td></td> <td></td> <td></td> </tr> <tr> <td>p. Predicted Rent Trend (next 6-12 months)</td> <td><u>Static</u></td> <td></td> <td></td> <td></td> </tr> </tbody> </table>					PREDOMINATING	90 %	OTHER TYPE	%	a. Type and Size	<u>3-4 room shacks</u>				b. Construction	<u>Shacks</u>				c. Average Age	<u>Old</u>				d. Repair	<u>Poor</u>				e. Occupancy	<u>99%</u>				f. Owner-occupied	<u>5%</u>				g. 1935 Price Bracket	<u>\$500-1000</u>	<u>% change</u>	<u>\$</u>	<u>% change</u>	h. 1937 Price Bracket	<u>\$700-1200</u>	<u>%</u>	<u>\$</u>	<u>%</u>	i. 1939 Price Bracket	<u>\$700-1200</u>	<u>%</u>	<u>\$</u>	<u>%</u>	j. Sales Demand	<u>Poor</u>				k. Predicted Price Trend (next 6-12 months)	<u>Static</u>				l. 1935 Rent Bracket	<u>\$5.00-12.50</u>	<u>% change</u>	<u>\$</u>	<u>% change</u>	m. 1937 Rent Bracket	<u>\$7.50-15.00</u>	<u>%</u>	<u>\$</u>	<u>%</u>	n. 1939 Rent Bracket	<u>\$7.50-15.00</u>	<u>%</u>	<u>\$</u>	<u>%</u>	o. Rental Demand	<u>Good</u>				p. Predicted Rent Trend (next 6-12 months)	<u>Static</u>			
	PREDOMINATING	90 %	OTHER TYPE	%																																																																																				
a. Type and Size	<u>3-4 room shacks</u>																																																																																							
b. Construction	<u>Shacks</u>																																																																																							
c. Average Age	<u>Old</u>																																																																																							
d. Repair	<u>Poor</u>																																																																																							
e. Occupancy	<u>99%</u>																																																																																							
f. Owner-occupied	<u>5%</u>																																																																																							
g. 1935 Price Bracket	<u>\$500-1000</u>	<u>% change</u>	<u>\$</u>	<u>% change</u>																																																																																				
h. 1937 Price Bracket	<u>\$700-1200</u>	<u>%</u>	<u>\$</u>	<u>%</u>																																																																																				
i. 1939 Price Bracket	<u>\$700-1200</u>	<u>%</u>	<u>\$</u>	<u>%</u>																																																																																				
j. Sales Demand	<u>Poor</u>																																																																																							
k. Predicted Price Trend (next 6-12 months)	<u>Static</u>																																																																																							
l. 1935 Rent Bracket	<u>\$5.00-12.50</u>	<u>% change</u>	<u>\$</u>	<u>% change</u>																																																																																				
m. 1937 Rent Bracket	<u>\$7.50-15.00</u>	<u>%</u>	<u>\$</u>	<u>%</u>																																																																																				
n. 1939 Rent Bracket	<u>\$7.50-15.00</u>	<u>%</u>	<u>\$</u>	<u>%</u>																																																																																				
o. Rental Demand	<u>Good</u>																																																																																							
p. Predicted Rent Trend (next 6-12 months)	<u>Static</u>																																																																																							
3. NEW CONSTRUCTION (next yr.) No. <u>0</u> Type & Price <u>---</u> How Selling <u>---</u>																																																																																								
4. OVERHANG OF HOME PROPERTIES: a. HOLC <u>2</u> b. Institutions <u>For</u>																																																																																								
5. SALE OF HOME PROPERTIES (3-yr.) a. HOLC <u>2</u> b. Institutions <u>For</u>																																																																																								
6. MORTGAGE FUNDS: <u>None</u> 7. TOTAL TAX RATE PER \$1000 (1937-) <u>\$52.70</u>																																																																																								
8. DESCRIPTION AND CHARACTERISTICS OF AREA:																																																																																								
Terrain: Level. No construction or flood hazards. Land improved 60%. Conveniences are all readily available. This is an old blighted area populated very largely by Mexican laborers whose employment is highly seasonal. Improvements are uniformly nondescript - shacks, evidencing no pride of occupancy. The area has many characteristics of a slum section and is accorded a "low red" grade.																																																																																								
9. LOCATION <u>San Fernando</u> SECURITY GRADE <u>4th</u> - AREA NO. <u>7-1</u> DATE <u>3/25/39</u>																																																																																								
349																																																																																								

Figure B.2: Los Angeles and Surrounding Districts & 1935 – 1940 HOLC Maps



Notes: “Best” (A, outlined in green), “Still Desirable” (B, outlined in blue), “Definitely Declining” (C, outlined in yellow), to “Hazardous” (D, outlined in red) - [bottom left] Redondo Beach Unified; [center] Los Angeles Unified; [top right] Pasadena ISD; [bottom right] Long Beach Unified

Figure B.3: Cleveland and Surrounding School Districts & 1935 – 1940 HOLC Maps



Notes: “Best” (A, outlined in green), “Still Desirable” (B, outlined in blue), “Definitely Declining” (C, outlined in yellow), to “Hazardous” (D, outlined in red) [center] Cleveland Municipal; [top left] Lakewood City; [top right] Cleveland Heights-University Heights; [bottom right] Shaker Heights

Figure B.4: School Total PPE by HOLC A-D District Grades – Only “D” Schools

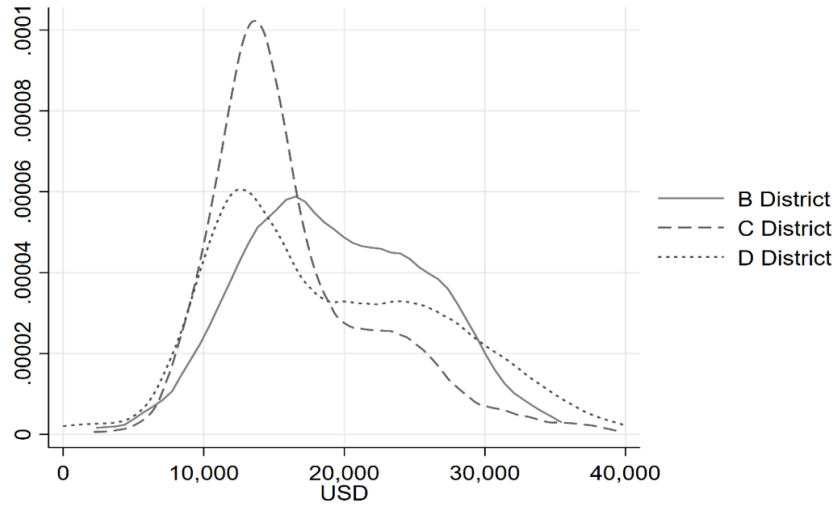
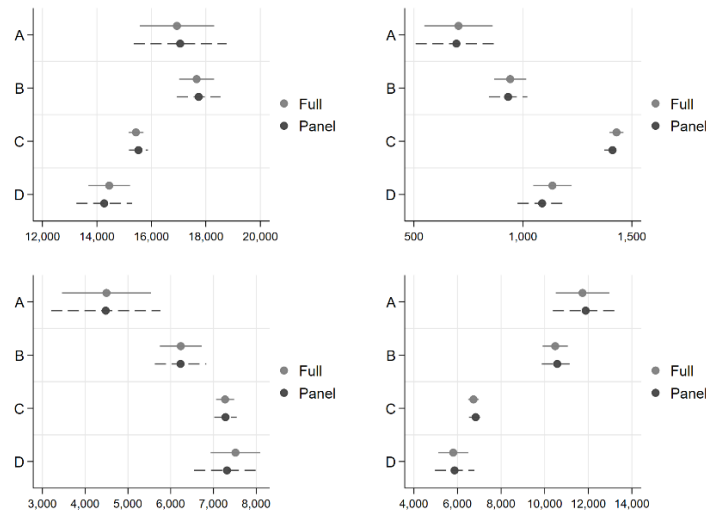
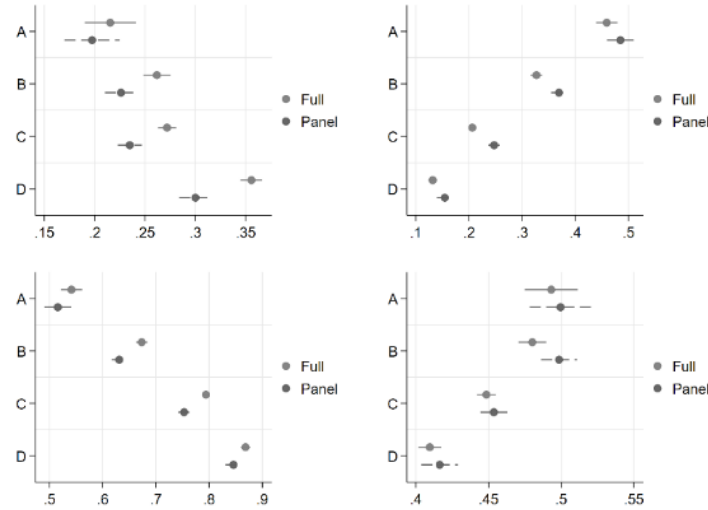


Figure B.5: HOLC A-D Coefficient Plots, Full vs. Panel, 2017-18, Finance (USD 2018)



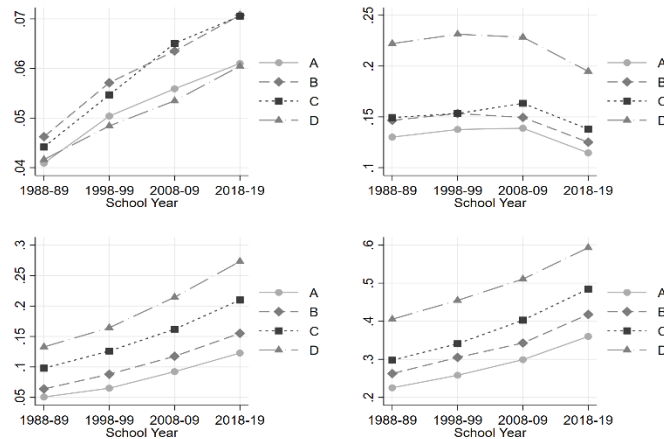
Notes: [top left] Per-Pupil Total Revenue; [top right] Per-Pupil Federal Revenue; [bottom left] Per-Pupil State Revenue; [bottom right] Per-Pupil Local Revenue. All values represented in USD 2018. Coefficients (solid dots) are HOLC A-D weighted averages from regressions of respective finance outcomes on HOLC A-D indicators without a constant term. The 95% confidence intervals (lines) are calculated using Huber-White heteroskedasticity-consistent standard errors. Regressions are run using both the 2017-18 full sample and the 2017-18 data from the panel sample. The full sample is the original 2017-18 cross-sectional sample and includes all $d = 1,760$ districts. The panel sample is the time series sample spanning 1989-90 through 2017-18 school years and includes $d_p = 1,109$ districts.

Figure B.6: HOLC A-D Coefficient Plots, Full vs. Panel, Diversity Outcomes



Notes: [top left] Percent Black; [top right] Percent White; [bottom left] %Non-White; [bottom right] Simpson's Diversity (1-D). Comments on full and panel samples here. Coefficients (solid dots) are HOLC A-D weighted averages from regressions of respective diversity outcomes on HOLC A-D indicators without a constant term. The 95% confidence intervals (lines) are calculated using Huber-White heteroskedasticity-consistent standard errors. Regressions are run using both the 2017-18 full sample and the 2017-18 data from the panel sample. The full sample is the original 2017-18 cross-sectional sample and includes all $n = 9,709$ schools. The panel sample is the time series sample spanning 1988-89 through 2018-19 school years and includes $np = 4,677$ schools.

Figure B.7: HOLC A-D Averages Over Time, Exposure Index (1988-2018)



Notes: [top left] White-Asian Exposure Index; [top right] White-Black Exposure Index; [bottom left] White-Hispanic Exposure Index; [bottom right] White-Non-White Exposure Index. Exposure index calculations for each race pair factor in all other student race groups.

Appendix C: Additional Tables and Figures from Chapter 3

Table C.1: Summary of Cellphone Policies and Metal Detector Status in Responding Schools

School year	Cellphone bans	Ban: High	Ban: Low	Ban: None	Missing	Schools*	% Ban**
<i>Pre-Policy</i>							
Fulltime	16	10	6	2	2	20	89%
Non-Fulltime	28	5	23	10	5	43	74%
Totals	44	15	29	12	7	63	100%
<i>Post-Policy</i>							
Fulltime	14	6	8	4	2	20	78%
Non-Fulltime	24	1	23	16	3	43	60%
Totals	38	7	31	20	5	63	100%

Notes: This table provides the count of active cellphone bans based on responding surveys. NYC public middle and high school principals and parent coordinators were asked what their cellphone policies were over 2 intervals: (1) pre-NYCDOE ban repeal (SY 2005-06 to SY 2014-15) and (2) post-NYCDOE ban repeal (SY 2015-16 to 2019-20). I classify a cellphone policy as a “high compliance” ban if students were not allowed to bring cellphones onto school premises or if students were allowed phones on school premises but were forced to either turn phones in before classes or keep them secure in their locker. A ban is classified as “low compliance” if students were allowed to bring phones into schools and allowed to use them at lunch or in study hall but not during regular classroom instruction. Survey respondents were asked questions related to discipline as well. I use the presence of metal detectors as a proxy for compliance, assuming schools with active full-time metal detectors during ban periods had perfect enforcement. Those schools without metal detectors present were assumed to have a low-compliance ban. *Sources:* Author-conducted cellphone and metal detector survey and supplemental metal detector data from NYPD SSA and InsideSchools.com. * Includes all responding schools, including those with incomplete survey responses. A total of 63 schools responded to my survey; of those 7 had missing cellphone ban responses for the pre-policy period (i.e., 2 Full-time; 5 Non-Full-Time) and 5 had missing cellphone ban responses for the post-policy period (i.e., 2 Full-time; 3 Non-Full-Time). ** Calculation based off non-missing school totals.

Table C.2: Impact of NYCDOE Cellphone Ban on Suspensions and Select School Culture Outcomes, SY2006-2016

	Per-Pupil Suspensions			School Behavior (Students)			School Safety (Students)			School Safety (Teachers)		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Cellphone Ban	-0.617 [0.868]	-1.381 [0.861]	-1.121 [0.867]	-0.066 [0.065]	-0.079 [0.068]	-0.09 [0.069]	-0.182*** [0.064]	-0.191*** [0.066]	-0.202*** [0.066]	-0.146* [0.083]	-0.156* [0.084]	-0.181** [0.085]
% FRPL Eligible		X	X		X	X		X	X		X	X
% ELL Students		X	X		X	X		X	X		X	X
% SWD Students		X	X		X	X		X	X		X	X
% Race Covariates		X	X		X	X		X	X		X	X
% Female		X	X		X	X		X	X		X	X
Enrollment		X	X		X	X		X	X		X	X
Leadership Changes			X			X			X			X
Sample Size	6,149	6,149	6,094	6,149	6,149	6,094	6,149	6,149	6,094	6,149	6,149	6,094
R2	0.063	0.191	0.189	0.06	0.407	0.404	0.07	0.366	0.364	0.01	0.188	0.19

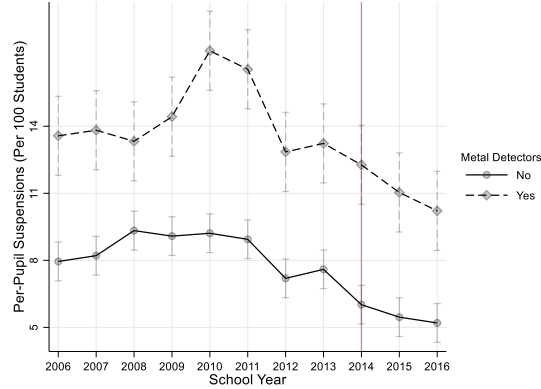
Notes: Model (1) is the baseline DiD model specification, Model (2) includes school-level characteristics, and Model (3) adds an indicator variable to account for principal changes at the time of treatment. The school culture measures are derived from NYCDOE school-level School Survey data. I use robust standard errors clustered at the school-level. Race covariates include: (1) Asian, (2) Black, (3) Hispanic, (4) White. A SY2013-14 pre-treatment cutoff is used for this analysis. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table C.3: Impact of NYCDOE Cellphone Ban on School Safety (Students) by Individual Outcome, SY2006-2016

	Safe Classroom (1)	Safe Outside (2)	Safe Hallways (3)
Cellphone Ban	-0.232*** [0.071]	- [0.061]	- [0.069]
% FRPL Eligible	X	X	X
% ELL Students	X	X	X
% SWD Students	X	X	X
% Race Covariates	X	X	X
% Female	X	X	X
Enrollment	X	X	X
Leadership	X	X	X
Changes			
Sample Size	6,094	6,094	6,094
R2	0.363	0.332	0.337

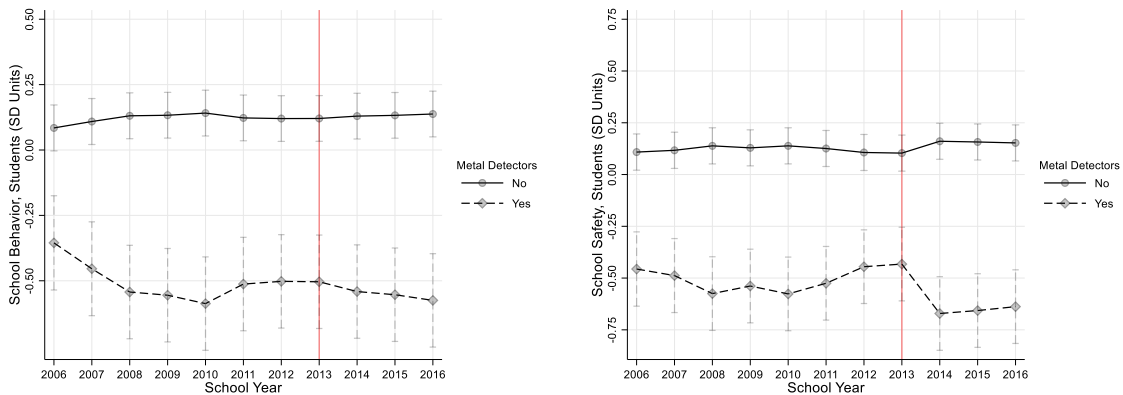
Notes: This table provides 2x2 DID regression estimates for the individual standardized outcomes contained in the student-based school safety index. Each model (1) – (3) uses the 2x2 DiD regression model specification (3). The school culture measures are derived from NYCDOE school-level School Survey data. I use robust standard errors clustered at the school-level. Race covariates include: (1) Asian, (2) Black, (3) Hispanic, (4) White. A SY2013-14 pre-treatment cutoff is used for this analysis. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure C.1: Per-Pupil Suspensions by Scanning Status, SY2006-2016



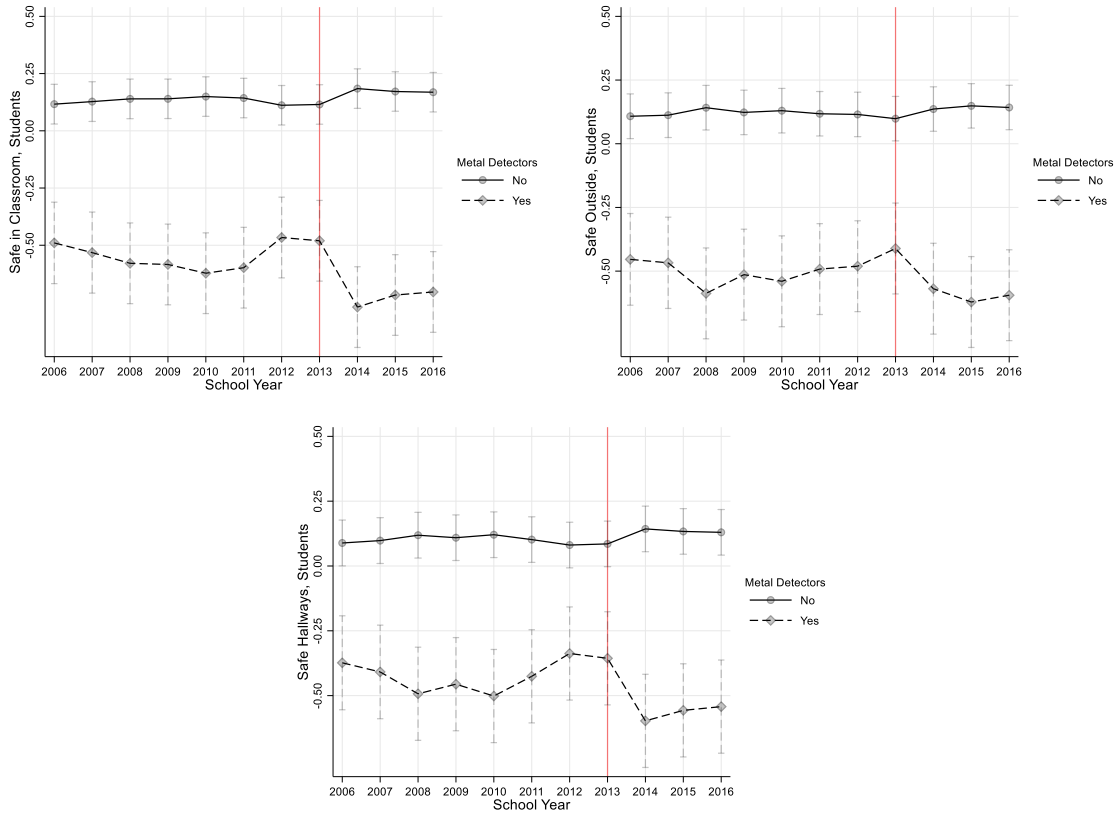
Notes: Per-pupil suspensions (per 100) is based on the sum of principal and superintendent suspensions for all NYC public middle and high schools in the analysis sample. Suspensions were calculated using NYCDOE student-level data aggregated at the school level and linked to an author-constructed metal-detector panel from NYPD SSA incident data. Schools coded “Yes” had full-time metal detector scanning, those coded “No” did not have full-time metal detector scanning (i.e., none, random, hand wands). The vertical red line marks the SY2014-15, the last year before NYCDOE lifted their ban.

Figure C.2: School Culture (Students) by Scanning Status, SY2006-2016



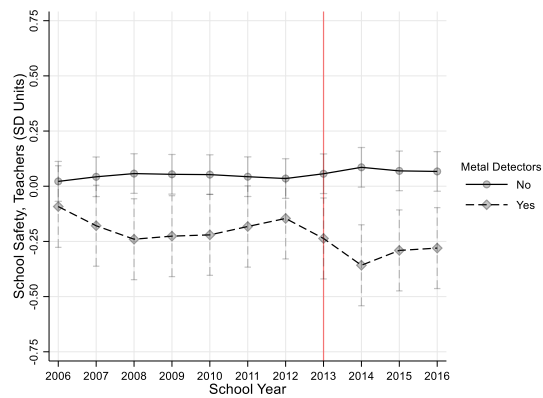
Notes: [left] Student Behavior (Students) [right] School Safety (Students). All measures calculated from student responses to annual questions on the NYCDOE School Survey. School Survey data cleaned and linked to an author-constructed metal-detector panel from NYPD SSA incident data. Schools coded “Yes” had full-time metal detector scanning, those coded “No” did not have full-time metal detector scanning (i.e., none, random, hand wands). The vertical red line marks the SY2013-14, the last full school year before NYCDOE lifted their districtwide cellphone ban in March of 2015. Students’ perceptions of school respect are omitted from this supplemental analysis due to data limitations.

Figure C.3: School Culture (Students) by Scanning Status, Individual Outcomes, SY2006-2016



Notes: [top left] Safe in Classroom (Students) [top right] Safe Outside (Students) [bottom] Safe in Hallways (Students). All measures calculated from student and teacher responses to annual questions on the NYC School Survey. NYC School Survey data cleaned and linked to an author-constructed metal-detector panel from NYPD SSA incident data. Schools coded “Yes” had full-time metal detector scanning, those coded “No” did not have full-time metal detector scanning (i.e., none, random, hand wands). The vertical red line marks the SY2013-14, the last full school year before NYCDOE lifted their districtwide cellphone ban in March of 2015. Students’ perceptions of school respect are omitted from this supplemental analysis due to data limitations.

Figure C.4: School Culture (Teacher) by Scanning Status, SY2006-2016



Notes: [middle] School Safety (Teachers). All measures calculated from teacher responses to annual questions on the NYCDOE School Survey. School Survey data cleaned and linked to an author-constructed metal-detector panel from NYPD SSA incident data. Schools coded “Yes” had full-time metal detector scanning, those coded “No” did not have full-time metal detector scanning (i.e., none, random, hand wands). The vertical red line marks the SY2013-14, the last full school year before NYCDOE lifted their districtwide cellphone ban in March of 2015. Students’ perceptions of school respect are omitted from this supplemental analysis due to data limitations.

Appendix D: Co-Authorship Statement

I. Preamble

This statement is prepared in accordance with the updated guidelines (October 2021) on the admissibility of co-authored works in milestone papers. This document is the co-authorship statement for Dylan J. Lukes and the following two papers in which he serves as a first author, co-equal or otherwise.

II. Paper 1

Citation

Pages, R., Lukes, D., Bailey, D., & Duncan, G. (2020). Elusive Longer-Run Impacts of Head Start: Replications Within and Across Cohorts. *Educational Evaluation and Policy Analysis*, 42(4), 471-492.

Statement

Remy Pages and Dylan Lukes share co-equal first authorship of this publication. Dylan's contribution as a co-equal first author includes sharing all paper-related responsibilities equally with Remy Pages, including Conceptualization, Data Curation, Methodology, Formal Analysis, Project Administration, Coding, Data Visualization, Writing – original draft, Writing – review & editing, Writing – response to reviewers.

III. Paper 2

Citation

Lukes, D., & Cleveland, C. (2021). A Lingering Legacy: The Relationship Between 1930s HOLC “Redlining” Maps and School Funding, Diversity, and Performance. *EdWorkingPapers*, 21-363, pp. 1-85.

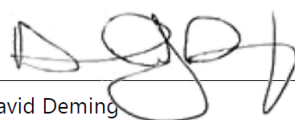
Statement

Dylan J. Lukes: First author; Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration – funding acquisition & grant application, Software, Supervision, Visualization, Writing – original draft, Writing – review & editing. Christopher H. Cleveland: Second author; Conceptualization, Writing – original draft, Writing – review & editing, Project Administration – funding identification.

IV. Signatures

x 

Dylan Lukes
Ph.D. Candidate

x 

David Deming
Dissertation Committee Chair

References

- Aaron, K., & Ye, J. (2015). More Than 90,000 New York City Students Are Searched Before School. *WYNC News*. Retrieved from <https://www.wnyc.org/story/school-metal-detectors/>.
- Aaronson, D. (1998). Using Sibling Data to Estimate the Impact of Neighborhoods on Children's Educational Outcomes. *The Journal of Human Resources*, 33(4), 915–946.
- Aaronson, D., Hartley, D., Mazumder, B. (2020). The Effects of the 1930s HOLC “Redlining” Maps. *Federal Reserve Bank of Chicago*. Working Paper.
- Aaronson, D., Faber, J., Hartley, D., Mazumder, B., & Sharkey, P. (2021). The long-run effects of the 1930s HOLC “redlining” maps on place-based measures of economic opportunity and socioeconomic success. *Regional Science and Urban Economics*, 86, 103622, 1-15.
- Abadie. (2005). Semiparametric difference-in-differences Estimators. *The Review of Economic Studies*, 72(1), 1–19.
- Alcaino, M., & Jennings, J. (2020). How Increased School Choice Affects Public School Enrollment and School Segregation. *EdWorkingPaper* No. 20-258. Providence, RI: Annenberg Institute at Brown University. Retrieved December 8, 2020, from <https://www.edworkingpapers.com/ai20-258>.
- Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *The Journal of Political Economy*, 113(1), 151–184.
- Anders, J. (2018). The Long Run Effects of De Jure Discrimination in the Credit Market: How Redlining Increased Crime. Working Paper. Retrieved March 1, 2021, from https://johnanders625665825.files.wordpress.com/2018/12/anders_redlining_12_16_2018.pdf
- Anderson, K., Foster, J., & Frisvold, D. (2010). Investing in Health: The Long-Term Impact of Head Start on Smoking. *Economic Inquiry* 48(3), 587-602.
- 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.
- Augustine, J. M., Prickett, K. C., Kendig, S. M., & Crosnoe, R. (2015). Maternal Education and the Link Between Birth Timing and Children's School Readiness. *Social Science Quarterly*, 96(4), 970-984.
- Angrist, J., & Lavy, V. (2002). New Evidence on Classroom Computers and Pupil Learning. *The Economic Journal*, 112(482), 735-765.

- Appel, I., Nickerson, J. (2016). Pockets of Poverty: The Long-Term Effects of Redlining. *SSRN Electronic Journal*, SSRN Electronic Journal.
- Bailey, M. J., Sun, S., & Timpe, B. (2018). Prep School for Poor Kids: The Long-Run Impacts of Head Start on Human Capital and Economic Self-Sufficiency. Unpublished manuscript, Department of Economics, University of Michigan, Ann Arbor, MI.
- Baker, B.D., Farrie, D., Sciarra, D., Luhm, T. (2018) Is School Funding Fair? 2018 Edition. Education Law Center of New Jersey & Rutgers GSE.
- Banerjee, A., Cole, S., Duflo, E., & Linden, L. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. *Quarterly Journal of Economics*, 122(3), 1235-1264.
- Barr, A., & Gibbs, C. R. (2019). *Breaking the Cycle? Intergenerational Effects of an Anti-Poverty Program in Early Childhood*. Manuscript submitted for publication.
- Barrera-Orsorio, F., & Linden, L.L. (2009). The Use and Misuse of Computers in Education: Evidence from a Randomized Experiment in Colombia. *World Bank Policy Research Working Paper* No. 4836. Retrieved December 14, 2021, from <https://ssrn.com/abstract=1344721>.
- Barrow, L., Markman, L. & Rouse, C.E. (2009). Technology's Edge: The Educational Benefits of Computer-Aided Instruction. *American Economic Journal: Economic Policy*, 1(1), 52-74.
- Bauer, L., & Schanzenbach, D.W. (2016). The Long-term Impact of the Head Start Program. Policy report. Retrieved from The Hamilton Project website: http://www.hamiltonproject.org/papers/the_long_term_impacts_of_head_start?_ga=2.58041645.688673869.1554397866-1708819741.1554397866
- Bayer, P., Blair, P., & Whaley, K. (2020). A National Study of School Spending and House Prices. NBER Working Paper No. w28255. Cambridge, MA: National Bureau of Economic Research. Retrieved March 1, 2021, from <https://www.nber.org/papers/w28255>.
- Beard, J. H., Morrison, C. N., Jacoby, S. F., Dong, B., Smith, R., Sims, C. A., & Wiebe, D. J. (2017). Quantifying disparities in urban firearm violence by race and place in Philadelphia, Pennsylvania: A cartographic study. *American Journal of Public Health*, 107(3), 371-373.
- Beland, & Murphy. (2016). Ill Communication: Technology, distraction & student performance. *Labour Economics*, 41, 61-76.
- Bergman, P. (2020). Parent-Child Information Frictions and Human Capital Investment: Evidence from a Field Experiment. *The Journal of Political Economy*, 129(1), 286-322.

- Bitler, M.P., Hoynes, H. W., & Domina, T. (2014). Experimental Evidence on Distributional Effects of Head Start. *NBER Working Paper Series*, 20434.
- Blinder, A. (1973). Wage Discrimination: Reduced Form and Structural Estimates. *The Journal of Human Resources*, 8(4), 436-455.
- Brittain, J., Willis, L., & Cookson, P. (2019). Sharing the Wealth: How Regional Finance and Desegregation Plans Can Enhance Educational Equity. *EdWorkingPaper* No. 19-187. Providence, RI: Annenberg Institute at Brown University. Retrieved March 1, 2021, from <https://www.edworkingpapers.com/index.php/ai19-187>.
- Caird, J., Willness, C., Steel, P., & Scialfa, C. (2008). A meta-analysis of the effects of cellphones on driver performance. *Accident Analysis and Prevention*, 40(4), 1282-1293.
- Callaway, B., & Sant'Anna, P.H. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, 225(2), 200–230.
- Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). Effectiveness of Reading and Mathematics Software Products: Findings from Two Student Cohorts. NCEE 2009-4041. In *National Center for Education Evaluation and Regional Assistance*. National Center for Education Evaluation and Regional Assistance.
- Card, D. (1999). The causal effect of education on earnings. In *Handbook of labor economics* (Vol. 3, pp. 1801-1863). Elsevier.
- Card, D., Mas, A., Moretti, E., Saez, E., 2012. Inequality at work: the effect of peer salaries on job satisfaction. *The American Economic Review*, 102 (6), 2981–3003.
- Carneiro, P., & Ginja, R. (2014). Long-term impacts of compensatory preschool on health and behavior: Evidence from Head Start. *American Economic Journal: Economic Policy*, 6(4), 135-73.
- Carter, Greenberg, & Walker. (2017). The impact of computer usage on academic performance: Evidence from a randomized trial at the United States Military Academy. *Economics of Education Review*, 56, 118-132.
- Chetty, R, Friedman, J., Hendren, N., Jones, M., & Porter, S. (2018). The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility. *NBER Working Paper* No. w25147. Cambridge, MA: National Bureau of Economic Research. Retrieved January 5, 2021, from <https://www.nber.org/papers/w25147>.
- Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, 126(4), 1593-1660.
- Chetty, R., & Hendren, N. (2018a). The Impacts of Neighborhoods on Intergenerational Mobility

- I: Childhood Exposure Effects. *The Quarterly Journal of Economics*, 133(3), 1107-1162.
- Chetty, R., & Hendren, N. (2018b). The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates. *The Quarterly Journal of Economics*, 133(3), 1163-1228.
- Chetty, R., Hendren, N., & Katz, L. (2016). The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment. *American Economic Review*, 106(4), 855-902.
- Clotfelter, C.T., Hemelt, S.W., Ladd, H.F., & Turaeva, M. (2019). School segregation in the era of immigration, school choice and color-blind jurisprudence - the case of North Carolina. *EdWorkingPaper* No. 19-101. Providence, RI: Annenberg Institute at Brown University. Retrieved March 1, 2021, from <https://www.edworkingpapers.com/ai19-101>.
- Clotfelter, C.T., Ladd, H.F., Clifton, C.R., & Turaeva, M.R. (2021). School Segregation at the Classroom Level in a Southern 'New Destination' State. *Race and Social Problems*, 13(2), 131-160.
- Connolly, N. (2014). *A World More Concrete: Real Estate and the Remaking of Jim Crow South Florida*. University of Chicago Press.
- Craig, A., & Martin, D. (2021). Discipline Reform, School Culture, and Student Achievement. Unpublished manuscript. Retrieved from https://sites.lsa.umich.edu/ashcraig/wp-content/uploads/sites/716/2021/06/suspensions_paper.pdf.
- Crespo, C., Smit, E., Troiano, R., Bartlett, S., Macera, C., & Andersen, R. (2001). Television Watching, Energy Intake, and Obesity in US Children: Results From the Third National Health and Nutrition Examination Survey, 1988-1994. *Archives of Pediatrics & Adolescent Medicine*, 155(3), 360-365.
- Cristia, J., Ibararán, P., Cueto, S., Santiago, A., & Severín, E. (2017). Technology and Child Development: Evidence from the One Laptop per Child Program. *American Economic Journal: Applied Economics*, 9(3), 295-320.
- Crossney, K. B., & Bartelt, D. W. (2005). The legacy of the Home Owners' Loan Corporation. *Housing Policy Debate*, 16(3-4), 547-574.
- Cunha, F., Heckman, J. J., Lochner, L., & Masterov, D. V. (2006). Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education*, 1, 697-812.
- Currie, J., & Almond, D. (2011). Human capital development before age five. In *Handbook of labor economics* (Vol. 4, pp. 1315-1486). Elsevier.
- Currie, J., & Neidell, M. (2007). Getting Inside the "Black Box" of Head Start Quality: What Matters and What Doesn't. *Economics of Education Review*, 26(1), 83-99.

- Currie, J., & Thomas, D. (1995). Does Head Start Make a Difference? *The American Economic Review*, 85(3), 341-364.
- Currie, J., & Thomas, D. (1999). Does head start help Hispanic children? *Journal of public Economics*, 74(2), 235-262.
- Dalane, K., & Marcotte, D.E. (2020). The Segregation of Students by Income in Public Schools. *EdWorkingPaper* No. 20-338. Providence, RI: Annenberg Institute at Brown University. Retrieved March 1, 2021, from <https://www.edworkingpapers.com/ai20-338>.
- De Haan, M., & Leuven, E. (2020). Head Start and the Distribution of Long-Term Education and Labor Market Outcomes. *Journal of Labor Economics*, 38(3), 000-000.
- Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics*, 1(3), 111-34.
- Dietz, & Henrich. (2014). Texting as a distraction to learning in college students. *Computers in Human Behavior*, 36, 163-167.
- Duncan, G. J., Lee, K. T. H., Rosales-Rueda, M., & Kalil, A. (2018). Maternal Age and Child Development. *Demography*, 55(6), 2229-2255.
- Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives*, 27(2), 109-32.
- Edunomics Lab (2021). NERD\$: National Education Resource Database on Schools (Version 1.0). *Georgetown University*. Retrieved 06/2021 from <https://edunomicslab.org/nerds/>
- Fairlie, R., & Robinson, J. (2013). Experimental Evidence on the Effects of Home Computers on Academic Achievement among Schoolchildren. *American Economic Journal: Applied Economics*, 5(3), 211-240.
- Fiel, J.E. (2013). Decomposing School Resegregation: Social Closure, Racial Imbalance, and Racial Isolation. *American Sociological Review*, 78(5), 828-848.
- Fishback, P., LaVoice, J., Shertzer, A., Walsh, R. (2020). Race, Risk and the Emergence of Federal Redlining. *NBER Working Paper* No. w28146. Cambridge, MA: National Bureau of Economic Research. Retrieved March 1, 2021, from <http://www.nber.org/papers/w28146.pdf>.
- Frisvold, D. E. (2006). *Head Start participation and childhood obesity*. [Vanderbilt university Economics Working Paper No. 06-WG01]. <https://ssrn.com/abstract=887433>.
- Frisvold, D. E., & Lumeng, J. C. (2011). Expanding exposure: Can increasing the daily duration of head start reduce childhood obesity? *Journal of Human resources*, 46(2), 373-402.

- Fryer. (2016). Information, non-financial incentives, and student achievement. *Journal of Public Economics*, 144(144), 109–121.
- Garces, E., Thomas, D., & Currie, J. (2002). Longer-term effects of Head Start. *American Economic Review*, 92(4), 999-1012.
- Gewertz, C. (2006). NYC Schools Take Hard Line on Cellphones. *EducationWeek*. Retrieved from <https://www.edweek.org/education/n-y-c-schools-take-hard-line-on-cellphones/2006/07>.
- Glass, A., & Kang, M. (2018). Dividing attention in the classroom reduces exam performance. *Educational Psychology*, 1-14.
- Goodman-Bacon. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2), 254–277.
- Green, P., Baker, B., & Oluwole, J. (2021). School Finance, Race, and Reparations. *Washington and Lee Journal of Civil Rights and Social Justice* (forthcoming). Retrieved January, 15, 2021, from <https://ssrn.com/abstract=3766279>.
- Hall, R.E., & Krueger, A. B. (2012). Evidence on the Incidence of Wage Posting, Wage Bargaining, and On-the-Job Search. *American Economic Journal: Macroeconomics*, 4(4), 56–67.
- Hancox, R., Milne, B., & Poulton, R. (2004). Association between child and adolescent television viewing and adult health: A longitudinal birth cohort study. *The Lancet*, 364(9430), 257-262.
- Hardy, L., Dobbins, T., Denney-Wilson, E., Okely, A., & Booth, M. (2009). Sedentariness, Small-Screen Recreation, and Fitness in Youth. *American Journal of Preventive Medicine*, 36(2), 120-125.
- Harris, E., & Schweber N. (2015). An Industry Will Vanish, Along With a Cellphone Ban in New York City Schools. *The New York Times*. Retrieved from <https://www.nytimes.com/2015/01/08/nyregion/industry-will-vanish-along-with-cellphone-ban.html>.
- Heckman, J. J., & Karapakula, G. (2019). Intergenerational and Intragenerational Externalities of the Perry Preschool Project. *NBER Working Paper Series*, 25889.
- Heckman, J. J., & Mosso, S. (2014). The economics of human development and social mobility. *Annual Review of Economics*, 6(1), 689-733.
- Heckman, J. J., & Pinto, R. (2015). Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs. *Econometric reviews*, 34(1-2), 6-31.

- Heffetz, O. (2011). A test of conspicuous consumption: visibility and income elasticities. *The Review of Economics and Statistics*, 93(4), 1101–1117.
- Hillier, A. E. (2003). Redlining and the Home Owners' Loan Corporation. *Journal of Urban History*, 29(4), 394-420.
- Hirschman, A. O. (1964) The paternity of an index. *The American Economic Review*, 54(5), 761-762.
- Hoffman, J. S., Shandas, V., & Pendleton, N. (2020). The Effects of Historical Housing Policies on Resident Exposure to Intra-Urban Heat: A Study of 108 US Urban Areas. *Climate*, 8(1), 12.
- Hotz, V. J., McElroy, S. W., & Sanders, S. (2005). Teenage childbearing and its life cycle consequences: Exploiting a natural experiment. *Journal of Human Resources*, 40(3), 683-715.
- Hoynes, H., Schanzenbach, D. W., & Almond, D. (2016). Long-run impacts of childhood access to the safety net. *American Economic Review*, 106(4), 903-34.
- Isikman, E., MacInnis, D., Ülkümen, G., Cavanaugh, L., & Brewer, Neil. (2016). The Effects of Curiosity-Evoking Events on Activity Enjoyment. *Journal of Experimental Psychology: Applied*, 22(3), 319-330.
- Jackson, K. T. (1980). Race, Ethnicity, and Real Estate Appraisal. *Journal of Urban History*, 6(4), 419–452.
- Jacobsen, W., & Forste, R. (2011). The Wired Generation: Academic and Social Outcomes of Electronic Media Use Among University Students. *Cyberpsychology, Behavior, and Social Networking*, 14(5), 275-280.
- Jacoby, S. F., Dong, B., Beard, J. H., Wiebe, D. J., & Morrison, C. N. (2018). The enduring impact of historical and structural racism on urban violence in Philadelphia. *Social Science and Medicine*, 199, 87-95.
- Jann, B. (2008). The Blinder–Oaxaca Decomposition for Linear Regression Models. *The Stata Journal: Promoting Communications on Statistics and Stata*, 8(4), 453-479.
- Johnson, R., & Jackson, C. (2019). Reducing Inequality through Dynamic Complementarity: Evidence from Head Start and Public School Spending. *American Economic Journal: Economic Policy*, 11(4), 310-349.
- Junco, R. (2012). In-class multitasking and academic performance. *Computers in Human Behavior*, 28(6), 2236-2243.
- Kraushaar, J.M., & Novak, D.C. (2010). Examining the Affects of Student Multitasking with

- Laptops during the Lecture. *Journal of Information Systems Education*, 21(2), 241-251.
- Krimmel, J. (2018). Persistence of Prejudice: Estimating the Long Term Effects of Redlining. *SocArXiv*, Center for Open Science.
- Kuznekoff, J., & Titsworth, S. (2013). The Impact of Mobile Phone Usage on Student Learning. *Communication Education*, 62(3), 233-252.
- Lai, F., Luo, R., Zhang, L., Huang, X. & Rozelle, S. (2015). Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing. *Economics of Education Review*, 47, 34-48.
- Levine, L., Waite, B., & Bowman, L. (2013). Use of Instant Messaging Predicts Self-Report but Not Performance Measures of Inattention, Impulsiveness, and Distractibility. *Cyberpsychology, Behavior, and Social Networking*, 16(12), 898-903.
- Light, J. (2010). Nationality and Neighborhood Risk at the Origins of FHA Underwriting. *Journal of Urban History*, 36(5), 634-671.
- Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly journal of economics*, 122(1), 159-208.
- Ludwig, J., & Phillips, D. A. (2008). Long-term effects of Head Start on low-income children. *Annals of the New York Academy of Sciences*, 1136(1), 257-268.
- Lutz, B. (2011). The end of court-ordered desegregation. *American Economic Journal. Economic Policy*, 3(2), 130-168.
- Malamud, O., & Pop-Eleches, C. (2011). Home computer use and the development of human capital. *Quarterly Journal of Economics*, 126(2), 987-1027.
- Marshall, S.J., Biddle, H., Gorely, T., Cameron, N., & Murdey, I. (2004). Relationships between media use, body fatness and physical activity in children and youth: A meta-analysis. *International Journal of Obesity*, 28(10), 1238-1246.
- Massey, D., & Denton, N. (1988). The Dimensions of Residential Segregation. *Social Forces*, 67(2), 281-315.
- Massey, D., & Denton, N. (1993). *American apartheid: Segregation and the making of the underclass (Democracy and urban landscapes)*. Cambridge, MA: Harvard University Press.
- Matthews, K. (2012). A phone home: NY teens pay valets to store devices. *USA Today*. Retrieved from <https://www.usatoday.com/story/tech/2012/10/04/phone-valet/1612725/>.

- Matthews, K. (2014). NYC poised to end school cellphone ban. *AP News*. Retrieved from <https://apnews.com/article/577d2718eedd478e9da91b9021eac902>.
- Miller, A. R. (2011). The effects of motherhood timing on career path. *Journal of Population Economics*, 24(3), 1071-1100.
- Miller, D. L., Shenhav, N. A., & Grosz, M. Z. (2019). Selection into Identification in Fixed Effects Models, with Application to Head Start. *NBER Working Paper Series*, 26174.
- Mitchell, B., & Franco, J. (2018). HOLC Redlining Maps: The Persistent Structure of Segregation and Economic Inequality. Washington, DC: National Community Reinvestment Coalition. Retrieved March 1, 2021, from https://ncrc.org/wp-content/uploads/dlm_uploads/2018/02/NCRC-Research-HOLC10.pdf.
- Monarrez, Tomas, Brian Kisida, and Matthew M. Chingos. (2020). The Effect of Charter Schools on School Segregation. *EdWorkingPaper* No. 20-308. Providence RI: Annenberg Institute at Brown University. Retrieved March 1, 2021, from <https://www.edworkingpapers.com/ai20-308>.
- Murnane, R.J. (2021). Can policy interventions reduce inequality? Looking beyond test Scores for evidence. William T. Grant Foundation.
- Nardone, A. L., Casey, J. A., Rudolph, K. E., Karasek, D., Mujahid, M., & Morello-Frosch, R. (2020). Associations between historical redlining and birth outcomes from 2006 through 2015 in California. *PloS One*, 15(8), E0237241.
- Nelson, R.K., Winling, L., Marciano, R., Connolly, N., et al. (2021). Mapping Inequality. American Panorama. Retrieved October 2, 2020, from <https://dsl.richmond.edu/panorama/redlining>.
- NYC Department of City Planning. (2020). School Districts (Clipped to Shoreline). Retrieved December 4, 2020, from <https://www1.nyc.gov/site/planning/data-maps/open-data/districts-download-metadata.page>.
- New York City Department of Education. (2022). NYC School Survey. Retrieved from <https://infohub.nyced.org/reports/school-quality/nyc-school-survey>.
- New York City Press Office. (2015). Mayor de Blasio and Chancellor Fariña to Lift School Cellphone Ban. Retrieved from <http://www1.nyc.gov/office-of-the-mayor/news/013-15/mayor-de-blasio-chancellor-fari-a-lift-school-cell-phone-ban>.
- New York State Education Department. (2021). School Safety and the Educational Climate. Retrieved May 18, 2021, from https://www.p12.nysed.gov/irs/school_safety/school_safety_data_reporting.html
- Ophir, E., Nass, C., & Wagner, A. D. (2009). Cognitive Control in Media Multitaskers.

- Proceedings of the National Academy of Sciences*, 106(37), 15583–15587.
- Oaxaca, Ronald. (1973). Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, 14(3), 693-709.
- Patterson, Richard W., & Patterson, Robert M. (2017). Computers and productivity: Evidence from laptop use in the college classroom. *Economics of Education Review*, 57, 66-79.
- Pew Research Center. (2013). Teens and Technology 2013. Retrieved from <https://www.pewresearch.org/internet/2013/03/13/teens-and-technology-2013/>.
- Pew Research Center. (2018a). Mobile Fact Sheet. Retrieved from <http://www.pewinternet.org/fact-sheet/mobile>.
- Pew Research Center. (2018b). Teens, Social Media & Technology 2018. Retrieved from <http://www.pewinternet.org/2018/05/31/teens-social-media-technology-2018>.
- Ravizza, S.M., Hambrick, D.Z., & Fenn, K.M. (2014). Non-academic internet use in the classroom is negatively related to classroom learning regardless of intellectual ability. *Computers & Education*, 78, 109-114.
- Ravizza, S., Uitzlugt, M., & Fenn, K. (2017). Logged In and Zoned Out: How Laptop Internet Use Relates to Classroom Learning. *Psychological Science*, 28(2), 171-180.
- Reardon, S.F., Grewal, E.T., Kalogrides, D. & Greenberg, E. (2012). Brown Fades: The End of Court-Ordered School Desegregation and the Resegregation of American Public Schools. *Journal of Policy Analysis and Management*, 31(4), 876-904.
- Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., Jang, H., & Chavez, B. (2021). Stanford Education Data Archive (Version 4.0). Retrieved February 8, 2021, from <http://purl.stanford.edu/db586ns4974>.
- Reardon, S.F., Weathers, E.S., Fahle, E.M., Jang, H., & Kalogrides, D. (2019). Is Separate Still Unequal? New Evidence on School Segregation and Racial Academic Achievement Gaps CEPA Working Paper No.19-06. Stanford, CA: Stanford Center for Education Policy Analysis. Retrieved March 1, 2021, from <http://cepa.stanford.edu/wp19-06>.
- Rich, M., & Taylor, K. (2014). With School Ban Nearing End, New York City Works on How and When to Allow Cellphones. *The New York Times*. Retrieved from <https://www.nytimes.com/2014/11/01/nyregion/with-school-ban-nearing-end-new-york-city-works-on-how-and-when-to-allow-cellphones.html>.
- Robinson, T. (1999). Reducing Children's Television Viewing to Prevent Obesity: A Randomized Controlled Trial. *JAMA*, 282(16), 1561-1567.
- Rouse, C.E., & Krueger, A.B. (2004). Putting Computerized Instruction to the Test: A

- Randomized Evaluation of a “Scientifically Based” Reading Program. *Economics of Education Review*, 23(4), 323-338.
- Roza, M., Hagan, K. & Anderson, L. (2020). Variation is the Norm: A Landscape Analysis of Weighted Student Funding Implementation. *Public Budgeting and Finance*. Retrieved March 1, 2021, from <https://doi.org/10.1111/pbaf.12276>.
- Sana, F., Weston, T., & Cepeda, N. (2013). Laptop Multitasking Hinders Classroom Learning for Both Users and Nearby Peers. *Computers & Education*, 62, 24-31.
- Sant’Anna, P.H., & Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1), 101–122.
- Shapley, K., Sheehan, D., Maloney, C., & Caranikas-Walker, F. (2011). Effects of Technology Immersion on Middle School Students' Learning Opportunities and Achievement. *Journal of Educational Research*, 104(5), 299-315.
- Simpson, E. (1949). Measurement of diversity. *Nature*, 163, 688.
- Smith, A., Rainie, L., & Zickuhr, K. (2011). College students and technology. *Pew Research Center*. Retrieved from <http://www.pewinternet.org/2011/07/19/college-students-and-technology>.
- Stothart, C., Mitchum, A., Yehnert, C., & Enns, James T. (2015). The Attentional Cost of Receiving a Cellphone Notification. *Journal of Experimental Psychology: Human Perception and Performance*, 41(4), 893-897.
- Suhr, K.A., Hernandez, D.A., Grimes, D., & Warschauer, M. (2010). Laptops and Fourth-Grade Literacy: Assisting the Jump over the Fourth-Grade Slump. *Journal of Technology, Learning, and Assessment*, 9(5), 46.
- Taylor, K. (2015). Ban on Cellphones in New York City Schools to Be Lifted. *The New York Times*. Retrieved from <https://www.nytimes.com/2015/01/07/nyregion/ban-on-cellphones-in-new-york-city-schools-to-be-lifted.html>.
- Urban Institute (2017). How do school funding formulas work? Washington, DC: Urban Institute. Retrieved March 1, 2021, from <https://apps.urban.org/features/funding-formulas/>.
- U.S. Census Bureau. (2020). 2019 TIGER/Line Shapefiles (machine readable data files). Retrieved November 8, 2020, from <https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html>.
- U.S. Census Bureau. (2021). Public School Spending Per Pupil Increases by Largest Amount in 11 Years. Retrieved June 8, 2021, from <https://www.census.gov/newsroom/press-releases/2021/public-school-spending-per-pupil.html>.

- U.S. Department of Education. (2009). The American Recovery and Reinvestment Act of 2009: Saving and Creating Jobs and Reforming Education. Retrieved March 1, 2021, from <https://www2.ed.gov/policy/gen/leg/recovery/implementation.html>
- U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2019). Status and Trends in the Education of Racial and Ethnic Groups 2018. Publication no. NCES 2019-038. Washington, DC: National Center for Education Statistics. Retrieved February 25, 2020, from <https://nces.ed.gov/pubs2019/2019038.pdf>.
- U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2020). Retrieved October 2, 2020, from <https://nces.ed.gov/programs/edge/Home>.
- U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2021). Retrieved December 11, 2021, from <https://nces.ed.gov/ccd/elsi>.
- U.S. Department of Health and Human Services, Administration for Children and Families. (2018). *Head Start Program Information Reports*. Retrieved from: <https://hses.ohs.acf.hhs.gov/pir/>
- U.S. Department of Health and Human Services, Administration for Children and Families. (2019). *Head Start program facts: Fiscal year 2019*. Retrieved from: <https://eclkc.ohs.acf.hhs.gov/about-us/article/head-start-program-facts-fiscal-year-2019>
- Wells, A.S., Fox, L., Cordova-Cobo, D., & Kahlenberg, R.D. (2016). How racially diverse schools and classrooms can benefit all students. *The Education Digest*, 82(1), 17.
- Ward, A., Duke, K., Gneezy, A., & Bos, M. (2017). Brain Drain: The Mere Presence of One's Own Smartphone Reduces Available Cognitive Capacity. *Journal of the Association for Consumer Research*, 2(2), 140-154.
- Woods, L. L. (2012). The Federal Home Loan Bank Board, Redlining, and the National Proliferation of Racial Lending Discrimination, 1921–1950. *Journal of Urban History*, 38(6), 1036–1059.
- Zigler, E., & Valentine, J. (1979). *Project Head Start: A legacy of the war on poverty*. The Free Press.