ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-808

When Does School Autonomy Improve Student Outcomes?

C. Kirabo Jackson
Northwestern University

This paper presents new evidence on the benefits of decentralization in public education, focusing on a Chicago policy that granted school principals more control over budgeting and operations. Meta-analysis of similar policies shows a small average effect with significant variation across settings. To explain this heterogeneity, I adopt theories from public finance, contract theory and psychology that suggest that the impact of autonomy depends on motivation effects, principal objectives, and the alignment between district and school choices. In event-study models, on average, increased school-level control improved math and English passing rates by about four percentage points (0.1σ), comparable to interventions costing over $1,000 per pupil but achieved at nearly zero cost. Affected schools also see reduced principal turnover and improved school climate, indicating increased stability and effort. Deconvolution-based analysis of the distribution of true effects reveals a range from zero at the 20th percentile to a ten percentage-point increase at the 80th percentile (approximately 0.2σ). I provide design-based evidence supporting the theoretical literature: (a) High-quality principals with a track record of strong test score growth experience more positive autonomy effects – underscoring the role of local capacity and well-aligned incentives. (b) Schools with atypical student populations benefit more from autonomy and allocate resources to services tailored to their student's specific needs – indicating that heterogeneity plays a key role.

VERSION: July 2023

Suggested citation: Jackson, C. Kirabo. (2023). When Does School Autonomy Improve Student Outcomes?. (EdWorkingPaper: 23-808). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/cdj7-rg41

When Does School Autonomy Improve Student Outcomes?

*

C. Kirabo Jackson
Northwestern University

July 12, 2023

Abstract

This paper presents new evidence on the benefits of decentralization in public education, focusing on a Chicago policy that granted school principals more control over budgeting and operations. Meta-analysis of similar policies shows a small average effect with significant variation across settings. To explain this heterogeneity, I adopt theories from public finance contract theory and psychology that suggest that the impact of autonomy depends on motivation effects, principal objectives, and the alignment between district and school choices. In event-study models, *on average*, increased school-level control improved math and English passing rates by about four percentage points (0.1σ), comparable to interventions costing over \$1,000 per pupil but achieved at nearly zero cost. Affected schools also see reduced principal turnover and improved school climate, indicating increased stability and effort. Deconvolution-based analysis of the distribution of true effects reveals a range from zero at the 20th percentile to a ten percentage-point increase at the 80th percentile (approximately 0.2σ). I provide design-based evidence supporting the theoretical literature: (a) High-quality principals with a track record of strong test score growth experience more positive autonomy effects – underscoring the role of local capacity and well-aligned incentives. (b) Schools with atypical student populations benefit more from autonomy and allocate resources to services tailored to their student's specific needs – indicating that heterogeneity plays a key role.

*Jackson: kirabo-jackson@northwestern.edu. The author thanks Claire Mackevicius for their invaluable research assistance and feedback on this project. The author also thanks Jerry Travlos, Allison Tingwall, and Sarah Dickson for providing helpful contextual information about the ISP program. The content is solely the responsibility of the author.

I Introduction

In the 1960s, public school systems in many nations were highly centralized, with key decisions made by district, state, or national authorities. However, recent education reforms aim to decentralize decision-making to the school level. During the 1980s, several nations transitioned toward "school-based management," and more recently, policies like the US Race to the Top, charter schools, UK Free Schools, Australian Independent Public Schools, and Dutch Freedom of Education Law have granted greater decision-making authority to schools. This policy shift is driven by the theory positing that assigning policy responsibility to the smallest feasible level of government promotes efficiency, as lower levels of government are better equipped to understand and respond to specific needs (Oates 1999, Wallis and Oates 1988). Theory indicates that while the benefits of decentralization are greatest in settings with heterogeneous needs (Acemoglu et al., 2007; Neri et al., 2022), granting more authority to principals can also lead to a loss of control, potentially resulting in ineffective decisions that are misaligned with the district's objectives (Aghion and Tirole, 1997; Grossman and Hart, 1986; Bishop et al., 2004). Therefore, the benefits of greater principal autonomy may depend on the dominant forces in a given context. While this is well-understood theoretically, empirical evidence based on design-based approaches is lacking. To address this gap, this study examines a 2016 Chicago policy that increased control over budgeting and operations for some school principals. This study investigates the average effect of this policy, offers fresh evidence for the US context, provides design-based evidence guided by theory on when decentralization is most effective, and explores underlying mechanisms.

The existing evidence on the effect of school autonomy on student achievement is inconclusive regarding the *average* effect and suggests context matters. In the US, attending charter and pilot schools, which enjoy greater autonomy than traditional public schools, does not improve outcomes *on average* (Cohodes and Parham, 2021; Abdulkadiroğlu et al., 2011). Similarly, a limited number of design-based US studies on policies that increase principal autonomy indicate limited effects. While Stiefel et al. (2003) and Merkle (2022) find generally positive effects of two reforms in New York City that increased autonomy, and both studies report small and statistically insignificant effects in conservative models. Moreover, Steinberg (2014), examining a similar 2005 policy in Chicago using a regression-discontinuity design, finds imprecise estimates consistent with a wide range of true effects. In the UK, Clark (2009) and Eyles and Machin (2019) demonstrate substantial gains when high schools convert to more autonomous models within broader educational reforms to increase school accountability. However, Eyles et al. (2017) and Regan-Stansfield (2018) find little evidence of average effects at the primary school level, with recent evidence suggesting negative effects of individual school-level autonomy (Neri and Pasini, 2023). Formal meta-analysis of these studies (Section II) reveals considerable heterogeneity – underscoring the importance of context.

To guide empirical work, I present a framework for identifying schools benefiting most and least from increased autonomy. While acknowledging that test scores don't capture everything valued by students and parents (Jackson, 2018; Jackson et al., 2020; Beuermann and Jackson, 2022; Beuermann et al., 2023), I define improvement on standardized tests as "better" outcomes. Drawing from psychology (Deci and Ryan, 1985) and management science (Muecke and Iseke, 2019), increased autonomy boosts motivation, reduces turnover, and enhances performance for all treated principals. However, decentralization benefits are particularly pronounced for schools with unique unmet needs (Acemoglu et al., 2007; Neri et al., 2022). Highly effective principals aligned with improving student achievement gain the most from autonomy (Merkle, 2022). Accountability aligns incentives between principals, parents, and districts, explaining why autonomy is associated with better outcomes in states with strong accountability (Loeb and Strunk, 2007) and nations with better institutions and more accountability (Fuchs and Wößmann, 2007). Testing this framework using design-based empirical strategies sheds light on heterogeneity within and across contexts.

This paper investigates the impact of the Chicago Independent School Program (ISP) implemented in 2016, which granted certain school principals increased autonomy in day-to-day operations, budgeting, and reduced district oversight. The analysis focuses on outcomes from 2009 to 2019 and examines 44 public elementary schools where principals transitioned to ISPs between 2016 and 2018. The study utilizes publicly available school-level data on Chicago Public Schools (CPS) obtained from city and state websites, linked to the public-use Common Core of Data (CCD), and a list of ISP principals and their designation dates, allowing for easy replication.

To isolate the effects of being granted greater autonomy under the program, I compare the change in outcomes before and after being granted increased autonomy to the change for a carefully selected set of comparison schools over the same time period. This approach involves first matching each treated school to a set of caparison schools based on pre-reform characteristics, stacking the data for each treated school (and its controls), then implementing a differences-in-differences-type model allowing each treated school to have its own counterfactual time path (based on its own set of matched comparison schools), similar to Deshpande and Li (2019) and Cengiz et al. (2019). This within-school difference-in-difference approach relies on the assumption that the *timing* of ISP designation is exogenous to other changes within treated schools. It identifies the treatment effects on the treated if (a) the treated and comparison schools share similar time shocks, and (b) the timing of ISP designation is unrelated to other changes at ISP schools. While there is no way to show this for certain, I present evidence suggesting that both conditions are likely satisfied. Moreover, I present several patterns supported by theory that are inconsistent with a pure bias story.

While treated schools are on very similar trajectories to comparison schools *before* treatment, after being granted autonomy, *on average* treated schools see relative improvements of about four percentile points in math and English (p-value<0.01). The average ISP effects increase monotonically over time, reaching about seven percentage points in both subjects after three years (p-value<0.01). These correspond to sizable effects between 0.10σ and 0.18σ. This intervention achieved gains similar to well-known interventions that cost over \$1,000 per pupil at a cost of under \$50 per pupil. With a benefit-cost ratio above 40:1, this policy was very cost-effective.

To document heterogeneity, I exploit the stacked data and estimate individual treatment effects for each school. I decompose the observed distribution of estimated effects into components attributed to sampling variability and true treatment effects. I find significant heterogeneity in treatment effects, indicating that any two randomly selected schools from our sample may exhibit true effects on average differing by six percentage points (approximately 0.15σ). Applying deconvolution-based estimates of the distribution of true effects (Wang and Wang, 2011), approximately one-quarter of the true treatment effects are negative. These findings challenge the notion that greater autonomy always improves student outcomes and supports the idea that effects can be either positive or negative, as theorized. The observed heterogeneity within Chicago aligns with meta-analytic results from existing design-based studies, emphasizing the importance of better understanding this variability.

To *explain* this effect heterogeneity, I employ meta-regression analysis with school-specific effect estimates as data points, similar to Card and Krueger (1992). To examine the significance of principal alignment/ quality, I link each school's estimated ISP effect to proxies representing principal alignment, including past performance in maintaining high test scores and teacher-rated measures of quality. Both measures reveal larger positive ISP effects for more capable principals, while the effects are *negative* for principals with a history of not improving scores and being deemed ineffective by teachers. Improved autonomy results in approximately three percentage points (0.075σ) higher passing rates for principals one standard deviation above the mean, highlighting the importance of principal quality.

To test the notion that the benefit will be larger when there is heterogeneity (Acemoglu et al., 2007) because schools may have specialized needs, schools were considered specialized if they had a proportion of bilingual students, special education students, or white, and Hispanic students above the 95^{th} percentile. Schools that were outliers in any of these categories had larger treatment effects than those that were not. Furthermore, schools that were outliers in multiple categories had even larger effects. These findings support the idea that autonomy is more beneficial when needs are heterogeneous. Further reinforcing this idea, I find that ISP schools that have students with identifiable specific needs (bilingual or special education) tend to allocate personnel spending differently than other ISP schools, and provide suggestive evidence that they allocate money toward the specific needs of their particular student population.

To examine the principal motivation effect, I analyze principal turnover and find that it decreases following ISP designation. Moreover, ISP principals tend to remain in their schools even after the required two-year commitment, indicating that autonomy is an "amenity" for principals. Consistent with providing stability, there are observed improvements in school climate at ISP schools. Notably, the increases in school climate and test scores align with the association found between school climate and test score value-added in cross-sectional studies (Porter et al., 2023; Jackson and Mackevicius, 2024). This suggests that stability and improved school climate may serve as mediating mechanisms in successful schools.

Using the estimated relationship between the ISP effect, principal quality, and outlier status, I project the predicted ISP effect to all schools in Chicago. Although the ISP effect is larger in existing ISP schools, it would still be substantial if applied to all schools. This is because many non-ISP schools have specialized student populations, albeit with relatively weaker principals. Given the sizable treatment effects observed at nearly zero cost, these findings suggest that granting greater autonomy to individual school principals can be highly effective in contexts with significant school heterogeneity, such as large urban areas with sizable immigrant populations. Additionally, if principals prioritize maximizing achievement due to strong accountability, social norms, or other factors, this policy can yield non-negligible allocative efficiencies. Consequently, there is an opportunity to improve outcomes with minimal financial cost.

The results contribute to the existing literature on school autonomy, providing clear evidence of its effects in the United States while emphasizing the significance of context and heterogeneity. Moreover, the results contribute to the principal quality literature, as evidenced by the predictive power of value-added measures in relation to treatment effects, despite their shortcomings (Chiang et al. (2016); Grissom et al. (2012)). The findings also validate established associations between survey-based assessments of leadership quality and student outcomes (Bloom et al. (2015); Liu et al. (2014)). In a broader sense, this study provides fresh evidence highlighting the importance of effective leaders (Bertrand and Schoar (2003); Jones and Olken (2005); Frick et al. (2007)). Lastly, while there is existing theoretical and empirical work on decentralized decision-making (Acemoglu et al. (2007); Neri et al. (2022); Colombo and Delmastro (2004)), this paper directly presents design-based evidence on when such decentralization may improve outcomes.

The remaining sections of the paper are organized as follows: Section II provides a meta-analysis of existing studies, Section III presents a theoretical framework, Section IV describes the data utilized, Section V outlines the estimation strategy employed, and Section VI presents the results. Finally, Section VII concludes the paper.

II Meta-Analysis of the Related Design-Based Studies

One of the first contributions of this paper is to document the extent of heterogeneity in design-based studies across contexts. While observational studies have documented heterogeneity across nations (Fuchs and Wößmann 2007) and states (Stiefel et al. 2003), this heterogeneity has not been well documented among design-based (i.e., credibly causal) studies.

As pointed out in Jackson and Mackevicius (2024), simply comparing point estimates across studies without regard to estimation errors can lead one to overstate heterogeneity. Accordingly, to assess whether the observed estimates lie above zero *on average*, and to assess whether the estimates truly differ from one another above and beyond that explained by sampling variability (an indication of true heterogeneity), I take 28 estimates (and the associated standard errors) from the eight design-based papers and perform a random-effects meta-analysis – estimating both the pooled average and the spread of true effect across these contexts. Note that I include multiple estimates from the same paper when results are presented for different subjects or tests or using different (reasonable) specifications.[1] Papers that report effects on passing rates are converted to standardized effects using the inverse normal transformation as in Ho (2009).[2]

I briefly outline the random effects meta-analysis model. In the simplest formulation, there is some grand mean (Θ) representing the pooled effect of all studies. Due to heterogeneity, the true effect in any context or individual study (θ_j) deviates from this pooled average with variance τ^2 – which represents true treatment heterogeneity. We do not observe the true effect (θ_j), but only a noisy estimate of if ($\hat{\theta}_j$) which has variance σ_j^2. If the estimation errors are unrelated to the underlying effect, the estimates will be centered on the grand mean Θ with total variance $\sigma_j^2 + \tau^2$. The optimal pooled average is an inverse-variance precision-weighted average where each estimate receives weight $(\sigma_j^2 + \tau^2)^{-1}$. To implement this estimate, one takes the squared reported standard error of each estimate se_j^2 as the true value of σ_j^2. To estimate τ^2, I follow DerSimonian and Laird (1986), which takes a precision-weighted variance of the raw estimates, and subtracts the variance expected due to sampling variability (based on the *observed* estimation errors) to provide an estimate of the variance of true heterogeneity (τ^2). This approach does not rely on any distributional assumptions regarding heterogeneity. To account for the dependence of estimates from the same study, I cluster estimates from the same study using Robust Variance Estimation (Hedges et al., 2010).[3]

I present the forest plot (i.e., the individual point estimates and their respective 95 percent

[1] For Merkle (2022), while not the model emphasized by the authors, I include models that include school-fixed effects given that the source of variation is within schools. Similarly, for Stiefel et al. (2003), I also include models that control for the change in outcomes before schools were actually granted autonomy (i.e., that account for anticipation or planning effects). In Neri and Pasini (2023), I include estimates for *all* schools that converted to the autonomous models and not only those that join school chains – which was the emphasis of the authors. For Steinberg (2014), I include estimates on both achievement scores and passing rates for both math and English – resulting ins 4 estimates for this study.

[2] Intuitively, one can compute the shift in a standardized latent normal outcome that would generate a given change in the passing rate.

[3] I implement these estimators using the "robumeta" package in Stata (Hedberg et al., 2017)

confidence intervals) in Figure 1. The 95 percent confidence interval for the pooled average is depicted in the lavender-shaded region. The plot shows a wide range of estimates across the studies (ranging from -0.15 in Steinberg and Cox (2016) to 0.17 in Clark (2009)). However, these largest estimates are relatively imprecise, and there is considerable overlap in the confidence intervals across most estimates. As such, a simple comparison of the estimates may drastically overstate the extent of true heterogeneity. Indeed, the pattern of results implies that these larger point estimates are likely imprecise estimates of real effects closer to the pooled average.

The precision-weighted pooled average across the studies is 0.0245σ (p-value=0.19) with heterogeneity τ of 0.036σ. That is, while the model rejects that all the effects are the same, it fails to reject that the pooled average differs from zero. In laypersons terms, *assuming the true effects are approximately normally distributed*, while school-level autonomy tends to be associated with better test scores on average, the true effect in any given context may be negative about one-quarter of the time but could generate meaningful positive effects (above 0.07σ) about ten percent of the time. It is important to note that this measure of heterogeneity is across study contexts. Heterogeneity across schools or districts within study contexts could be appreciably larger – underscoring the likely important of heterogeneity. In sum, existing work suggests that greater school autonomy *can* improve student outcomes but underscores the need to better understand (a) the role of heterogeneity and (b) whether theoretical predictions regarding which settings where such policies will be most effective hold in the data. This study seeks to help fill these knowledge gaps.

III The Policy and Theoretical Framework

III.1 The History of Decentralization Policy in Chicago Public Schools

Operational decisions in Chicago Public Schools (CPS), such as hiring and curriculum, are made at the district or state level, similar to most public schools in the United States. CPS organizes its district-run schools into 18 school networks based on location and school type. These networks provide administrative support, strategic direction, and leadership development to the schools within each network. Network chiefs are intermediaries between the district's administration and individual school principals. The Independent School Principal (ISP) Program was implemented in 2016. The ISP program rewards high-performing principals with increased autonomy. The program involved the following four key components: (1) Exemption from network membership and network chief oversight. (2) Exemption from budget and Work Plan approval.[4] (3) Increased flexibility with their budget and purchases. (4) Requirement to remain in the principal role at the current school for at least two years. Principals within CPS must apply to receive the

[4]The CIWP is the strategic planning process of schools that also meets the federal and state requirements of a school improvement plan.

ISP designation and undergo a competitive application, interview, and review process.

To be eligible for the program, the principal's school had to show strength in at least three categories used to compute the School Quality Rating Policy (SQRP) – the district's system for measuring annual school performance. The rating weights school growth percentiles in Math and ELA at about 35%, the percentage of students exceeding national growth norms at 10%, average school attainment at 15%, percent of English Language Learners making annual progress at 5%, average school attendance at 20%, school climate at 10%, and data quality at 5% (see Appendix Figure A2). All but one ISP principal was at a school with an "*Exemplary*" or "*Commendable*" rating. This is in line with the district aims of *"Reward high-performing principals"* (Appendix Figure A1). Also, applicants must demonstrate in their application and through an interview that they have the necessary internal capacity and a plan to compensate for the loss of centralized support from the network.[5] In sum, ISPs must (a) have demonstrated strong performance and (b) have a plan to address the loss of support from the district office. Successful principals receive ISP designation for the next school year.

In an important study of ISP principals (Travlos, 2020), they express being "*free from a network structure of oversight and accountability that is fragmented, stressful, and consumes valuable leadership time*" and report feeling "*valued and rewarded*." This highlights that while autonomy is at the heart of the reform, it could generate positive effects through increased principal effort, time, and motivation. Shedding light in what ISPs do, Travlos (2020) reports that ISPs "*have more authority and time to be collaborative, creative, and resourceful in meeting the needs of their students, teachers, and communities*" and "*use their autonomy to select curricula, assessments, and professional development that work best for their schools.*" This echoes descriptive work of a similar 2008 policy in Chicago (Steinberg and Cox, 2016; Steinberg, 2014). At the same time, the increased independence led to reports of feeling "*isolated*" and expressed a greater need for "*budget, network, and management supports.*" This suggests that the program allowed high-performing principals to implement changes that would improve their students' outcomes, but that the increased independence may have led a small number of ISPs to struggle. I formalize these ideas below.

III.2 Theoretical Framework and Testable Predictions

I identify two channels through which school autonomy can impact student outcomes. The first channel is by promoting increased principal effort and motivation. By granting autonomy as a form of reward, the ISP program can serve as an amenity that enhances motivation, effort, and reduces burnout (Deci and Ryan, 1985). Evidence from a recent meta-analysis of 318 studies supports the association between decision-making autonomy, work motivation, and reduced strain (Muecke and Iseke, 2019). Additionally, the commitment to remain in the position for at least two years is

[5] In fact, the school's network chief must approve the plan.

likely to enhance principal effort and reduce turnover. These factors contribute to school stability and create a better climate, ultimately improving student outcomes. I refer to this as the **stability channel**, which has the potential to benefit all schools similarly.

The second channel is the **allocative efficiency channel**. Autonomy in budget and operations can impact student outcomes through improved or reduced allocative efficiencies. For instance, the choice of textbooks and teacher training methods have been shown to affect student performance (Jackson et al. 2014; Koedel and Polikoff 2017; Kraft et al. 2018; van den Ham and Heinze 2018). Therefore, the freedom to select curriculum and professional development, which are often determined by the district, can directly influence student outcomes. Additionally, budget autonomy allows principals to allocate funds across categories and spend more effectively. Prior to becoming ISPs, principals have expressed frustration about the need for approval to reallocate funds from one designated use to a more effective one (Travlos, 2020).[6] Building upon this, I propose a framework that formalizes the role of allocation efficiencies and suggests that increased autonomy can have both positive and negative effects. Importantly, this framework generates testable predictions regarding the schools that are most likely to benefit from enhanced autonomy.

FORMALIZING THE ALLOCATIVE EFFICIENCY CHANNEL

OPTIMAL CHOICES: I consider a simplified model with two inputs, x_1 and x_2, with respective prices, p_1 and p_2. Each school has production function, $F_i(x_1, x_2)$, but *all* schools have the same fixed budget, B. All schools exhaust their budgets so that each school's output can be expressed as a function of a single input, $f_i(x_1)$. Due to heterogeneity in the production function, schools have different levels of input 1 that maximize their output (i.e., x_{1i}^*). So long as the production technology is convex, student achievement is maximized at x_{1i}^* and monotonically decreases as one uses x_{1i} farther from x_{1i}^* in any one direction. This basic setup is depicted in Figure 2.

CENTRALIZATION: Under centralization, the district chooses input level x_{1d} that *all* schools must employ. This captures the idea that "*central office leaders can limit school leaders' decision-making power*" (Wong et al., 2020). Because $x_{1d} \neq x_{1i}^*$ for many schools, some schools can increase achievement by using more or less of x_1. This captures the idea that principals often report "*misalignment between school and district priorities*" (Steinberg and Cox 2016; Travlos 2020).

PRINCIPAL QUALITY: Each principal in the study has a unique utility function, with "aligned" principals seeking to maximize student achievement and "misaligned" principals pursuing alternative objectives. Misalignment can arise due to self-interest, where principals prioritize non-academic factors that benefit themselves, such as having a winning football team or providing laptops for teachers. Alternatively, misalignment may stem from different priorities, such as principals prioritizing socioemotional learning over test-preparation services based on their belief that

[6] Liebman et al. (2017) shows that intertemporal non-fungibility due to expiring budgets leads to wasteful spending.

test scores do not fully capture student success. Lastly, misalignment can result from differences in principal skill or capacity to identify student needs, leading to sub-optimal decision-making. The focus of the study is student achievement, measured through standardized tests, and from this perspective, greater alignment is desirable, as ”aligned” principals choose the optimal input level x^*_{1i} while ”misaligned” principals choose their privately optimal level $x^*_{1p} \neq x^*_{1i}$ when granted autonomy. *As a shorthand, I use alignment as a measure of principal quality*.

Result 1: The potential benefits of autonomy are increasing in the distance between x^*_{1i} and x_{1d}. Intuitively, the gains to autonomy are larger when individual schools have very specific needs that may differ from the district's choice. This captures the Acemoglu et al. (2007) result that the gains to decentralization are increasing in heterogeneity –because greater heterogeneity increases variability in this distance.

Result 2: The achievement gains of autonomy will be largest for "aligned" principals (*all else equal*). This holds since $x^*_{1p} \neq x^*_{1i}$. This captures the idea that autonomy may have better effects in setting with higher capacity (Galiani et al., 2008) and where accountability may align principal's incentives toward improved student outcomes (Stiefel et al., 2003; Fuchs and Wößmann, 2007).

Result 3: The benefits of autonomy will be heterogeneous and can be positive or negative. Outcomes will improve under autonomy so long as the principal is better aligned with the optimum than the district (i.e., the distance between x^*_{1i} and x_{1d} is large while that between x^*_{1i} and x^*_{1p} is small). This condition highlights that even misaligned principals can improve outcomes if the district and school choices are badly misaligned. However, outcomes can decline under autonomy if the principal is misaligned and the district was previously well-aligned with the school (i.e., the distance between x^*_{1i} and x_{1d} is small while that between x^*_{1i} and x^*_{1p} is large).

The framework suggests that while stability has a positive effect on all schools, allocative efficiency can have either a positive or negative effect – making the overall effect ambiguous in sign. Considering this theoretical ambiguity and the heterogeneous effects documented in Section II, I examine the *average* effect of the ISP on student outcomes in Section VI.1, and also explore the causal channels outlined in the model in Section VI.4, highlighting how the results can inform broader policy debates about the advantages of autonomy.

IV Data

I collect data from a variety of public sources freely available on the internet. I obtain a list of each school principal designated to be an ISP and the year of designation from the (CPS website). I matched each school to a database of all CPS schools in 2014-2015 academic year (the year immediately before the first ISP schools were designated). Achievement data come from two sources (ISAT reports) from 2009 through 2014 and the Partnership for Assessment of Readiness for College and Careers (PARCC) exams from 2015 through 2020. Both these exams were the state assessment and accountability measure for Illinois students enrolled in a public school district for the respective years.[7] These data report the percentage of grade 3 through 8 students who attain the proficiency standard at a school from 2001 through 2022. Demographic data for each school in each year are obtained from the Common Core of Data (CCD). We obtain from these data information on school enrolment, percent of free or reduced-price lunch, race, bilingual status, and special education stats from 2001 through 2019. These performance data and demographic data are matched to the list of ISP principals by school and year.

Additional data on school climate and organization come from the 5Essentials for the years 2014 through 2020. The five essentials measures school climate in the following five domains: (1) Effective Leaders: The principal works with teachers to implement a clear and strategic vision for school success. (2) Collaborative Teachers: Teachers collaborate to promote professional growth. (3) Involved Families: The entire school staff builds strong relationships with families and communities to support learning. (4) Supportive Environment: The school is safe and orderly. Teachers have high expectations for students and support students to realize their goals. Classmates also support one another. (5) Ambitious Instruction: Classes are academically demanding and engage students by emphasizing the application of knowledge. Publicly available data report each school from 1 through 5 across all these domains.[8] Individual survey questions included in the underlying 5Es surveys have been used to measure student socioemotional learning (Jackson et al., 2020, Jackson et al., 2023) and school climate (Porter et al., 2023).

Because principal quality is an important variable for our analysis, I obtained specific scores for the effective leadership domain (manually entered from CPS reports) in 2016 – the first year of the ISP program. The metric is designed to identify schools in which the leadership orients people, programs, and resources toward a focused vision for sustained improvement.[9] The leadership score is a summary measure of four dimensions of leadership quality: Teacher influence, Program coherence, Instructional leadership, and Teacher-Principal Truest. These data are also linked to personnel files from 2010 through 2020.[10] These data include each position in CPS and the salary attached to each position. This file allows one to code up principal turnover and spending on particular kinds of personnel (instruction, support staff, special education). I also obtain official measures of teacher turnover (one year) and principal turnover (six-year) from the Illinois Report Card.

The final dataset comprises all elementary schools in CPS between 2010 and 2019. I exclude

[7]The PARCC incorporates the Common Core for English and Math.

[8]Obtained from the SQRP reports.

[9]From this perspective, good leaders practice shared leadership, set high goals for quality instruction, maintain mutually trusting and respectful relationships, support professional advancement for faculty and staff, and manage resources for sustained program improvement.

[10]These were downloaded as pdf files and then digitized to create the personnel dataset.

data for 2020 through 2022 to avoid conflating any results with any ill-effects of school shutdowns associated with Covid-19. As discussed above, the set of schools that became ISPs was not typical of schools in the district. Table 1 presents summary statistics for ISP schools (columns 1 through 3) and all non-ISP schools (columns 4 through 6). In general, ISP schools had higher proficiency rates than non-ISP schools (50.47 percent for ELA and 49.95 percent for Math for ISP schools compared to 36.17 percent for ELA and 34.86 percent for Math for non-ISP schools). Also, ISP schools scored 4.3 on the 5Essentials compared to 3.8 for non-ISP schools. This 0.5 difference in climate score is a meaningful 0.33σ (or 12 percentile points) difference. The ISP schools also had more white students (14.6 versus 7.19 percent), fewer black students (19.7 versus 55 percent), and fewer students on free or reduced-price lunches (76 versus 85 percent) than non-ISP schools. In sum, while many non-ISP schools may have similar attributes to the ISP schools, the non-ISP schools are quite different from all other schools in general. This motivates my use of within-school variation for identification and matching to identify a set of comparison schools that may have been on a similar trajectory as ISP schools. I detail these approaches below.

V Empirical Strategy

To address potential bias arising from differences between ISP schools and non-ISP schools, I employ a design that compares the change in outcomes before and after ISP designation among ISP schools. This helps eliminate the influence of time-invariant disparities, such as variations in student composition or pre-ISP performance. To account for underlying outcome changes over time, I use a comparison group of schools that did not adopt the ISP during the same period. I first present average results using a standard (or "*naive*") difference-in-differences (DiD) model represented by equation (1). Here, Y_{st} denotes the outcome of interest for school s in year t, ISP_s equals one if school s was designated as an ISP school within our sample period. For treated schools, τ is the year relative to the year of ISP designation (i.e., event time), while for non-treated units, τ is -10. 1_τ equals 1 for all observations of ISP schools during year τ, γ_s represents a school fixed effect to account for time-invariant differences, and γ_t captures year fixed effects to account for common time shocks across schools. Standard errors are clustered at the school level.

$$Y_{st} = \sum_{\tau=-8,\tau\neq-1}^{3} \beta_\tau (ISP_s \times 1_\tau) + \gamma_s + \gamma_t + \varepsilon_{st} \tag{1}$$

The β_τ map out the evolution of outcomes in the ISP schools relative to the comparison schools before and after being designated an ISP.

While the "naive" model above is valid under certain conditions, it *can* lead to problems when treated units are used as control units (due to the negative weights problem), and if the time shocks are not the same between treated and comparison units. To address both these concerns, I use a carefully selected group of comparison schools that (a) were not ISP designated in the sample (eliminating the negative weight problem), and (b) with similar observable characteristics *before* the ISP program was implemented (making the common time shock condition much more plausible). My *preferred* model assesses whether, among schools that were observationally similar before the ISP program was introduced, ISP schools had more improved outcomes than comparison non-ISP schools after ISP designation. The key identifying assumptions are that (1) the carefully selected set of non-ISP schools would have had a similar trajectory in outcomes as the ISP-designated schools and that (2) the *timing* of ISP designation is unrelated to any *other* contemporaneous changes at ISP schools. I present empirical tests indicating the validity of both these assumptions.

To form a comparison group of schools that were never designated an ISP (avoiding problems of using already treated units as controls), I matched each ISP school to a set of non-ISP schools with similar math scores, principal seniority, demographics, and school climate scores in 2014 (two years before any school was designated an ISP). The average match quality is high with fewer matches, so the bias due to different time shocks is likely small. However, with few matches, the sample size is small so estimates may be imprecise. As such, there is a bit of a tradeoff between bias and efficiency when choosing the number of matches. To balance bias and efficiency, each ISP-treated school is put in a dataset with fifteen non-ISP matches, constituting group g. Importantly, **the main results are robust to using a smaller or larger number of matches** (see Table A1 and Figure A5). All of these group g datasets are appended to create a stacked dataset that includes a mini-dataset for each treatment-school group. I then estimate models as below for outcome Y for each school s, in treated group g, in each year t, Y_{sgt}.

$$Y_{sgt} = \sum_{\tau=-8,\tau\neq-1}^{3} \beta_\tau (ISP_s \times 1_\tau) + \gamma_s + \gamma_{t,g} + \varepsilon_{sgt} \tag{2}$$

In (2), all variables are as in (1), but now the subscript g connotes school-group g. To (a) account for differences in time effects for different kinds of schools, and (b) ensure that never-treated schools are the basis for comparison, I include school-group by year fixed effects γ_{gt}. Finally, ε_{sgt} is a school-level error term that I allow to vary by group (for control schools that may be used for more than one treated school). Standard errors are clustered at the school level. This is essentially the stacked difference-in-difference used in Cengiz et al. (2019) and Deshpande and Li (2019).

VI Results

VI.1 Average Effects

To demonstrate how outcomes change over time for ISP schools before and after ISP designation, I present the standard event-study model in Figure 3. The year before ISP designation is grey, while the post-designation years are blue. The figure also presents the 95 percent confidence interval for each relative year effect. For ELA passing rates (right), there is some suggestive evidence of differential pre-trends, but this is not statistically significant. For math passing rates (left), there is no evidence of differential pre-trends – consistent with the validity of the identifying assumptions. Consistent with some relative improvement associated with ISP destination, for both subjects, pass rates improve in year zero and after by about five percentage points.

To summarize these naive DiD average results, I estimate simple before vs. after effects using the full set of public elementary schools in Chicago – columns 1 and 2 of Table 2. ELA and math passing rates increase by 4.423 percentage points (p-value$<$0.001) and 3.153 percentage points (p-value$<$0.001), respectively. I also estimate the effects of having the ISP destination for two or more years (the second panel). In such models, after two years of ISP destination, ELA and math passing rates increase by 6.756 percentage points (p-value$<$0.001) and 5.576 percentage points (p-value$<$0.001), respectively. In sum, in naive difference-in-difference models, the effect on passing rates in simple before versus after models is about 3.8 percentage points – which is equivalent to an increase of about 0.095σ using the inverse-normal transformation (Ho 2009), while the effect after three years of the program is about 6.16 percentage points – equivalent to an increase of 0.155σ.

While the standard DiD event study is highly suggestive of a real effect, potential bias in such models is now well-established (Roth et al., 2023), and the event study is suggestive of some differential pre-trending between ISP schools and all other schools for ELA pass rates. As such, it is important to show that these effects persist in a more credible version of this model. I now turn to the results using the stacked DiD approach on the set of matched schools. These are presented in Figure 4. The first notable pattern is that the event-study plots for the matched DiD are similar to those from the standard DiD. For both ELA (left) and Math (right) passing rates, there is now no evidence of differential pre-trends. This suggests that eliminating comparison schools that were dissimilar from the treated schools led to a comparison set of schools that had a very similar trajectory in outcomes over time and, therefore, very likely shared common time shocks to the ISP schools. Also, note that this model does not require any additional covariates – avoiding potential problems with models that rely on time-varying covariates for credible identification (Caetano and Callaway 2023). As in the standard DiD models, pass rates improve in year zero for both subjects and increase by around five percentage points three years later. To assuage concerns that the results are driven by choice of method, I present the Callaway and Sant'Anna (2021) estimator – an alternative approach to addressing the negative weight problem (in Appendix Figure A5). This approach yields results very similar to other models, indicating no differential pre-trending and improved passing rates (by about five percentage points) after ISP designation.

To summarize these preferred average results, I estimate a simple before versus after effect using the treated and matched schools only – columns 5 and 6 of Table 2. ELA and math passing rates increase by 4.207 percentage points (p-value$<$0.001) and 3.113 percentage points (p-value$<$0.001), respectively. These estimates are very similar to the standard DiD estimates. After two years of ISP destination, ELA and math passing rates increase by 6.486 percentage points (p-value$<$0.001) and 4.934 percentage points (p-value$<$0.001), respectively. In sum, in the preferred models, the effect on passing rates in simple before versus after models is about 3.66 percentage points – equivalent to an increase of about 0.092σ, while the effect after three years of the program is about 5.7 percentage points – equivalent to an increase of about 0.144σ.

Putting The Average Effects Estimates in Context

To put these average effects in perspective, it is helpful to consider the pooled results from the meta-analysis. The simple before versus after effect of 0.085σ is larger than that of the pooled average across studies (0.023σ), but similar to, and within the range of, many reported estimates from the literature. Because my estimates are noisy estimates of the true average effect in this context, one can use the existing literature to provide an improved estimate of the true average effect in the Chicago context. That is, taking a Bayesian perspective, I can borrow strength from other studies (Efron and Morris 1973; Morris 1983) to inform the true effect in this context. This is analogous to creating an empirical Bayes estimate for teacher effects (as in Kane and Staiger (2008); Jackson and Bruegmann (2009) and others) by using a weighted average of the noisy estimate (0.085σ) and the pooled average (0.023σ), where more precise estimates are more heavily weighted toward the estimate while noisier estimates are weighed closer to the pooled average. Formally, if one assumes that the estimates are normally distributed around the grand mean with variance $\sigma_j^2 + \tau^2$, then the expected value of the true effect for study j is (3) where $B = (\sigma_j^2)/(\sigma_j^2 + \tau^2)$.

$$E(\theta_j|\hat{\theta}_j, \sigma_j, \tau) = B \times \Theta + (1 - B) \times \hat{\theta}_j \tag{3}$$

Replacing σ_j, τ, and Θ with their estimates, (3) yields the Best Linear Unbiased Prediction (BLUP) of the true effect for study j given the data (i.e., the best estimate of the *true effect* in a similar setting). The BLUPs ($\tilde{\theta}_j$) – also Empirical Bayes – are weighted averages of the individual estimates and the pooled average, where more precise estimates receive greater weight. Constructing an Empirical Bayes estimate yields a predicted true effect (given both the raw estimate and the information from the extant literature) of 0.06σ.

To put the magnitude of this effect in perspective, increasing teacher quality by one-half of a standard deviation increases test scores by this same amount (0.06σ) and raises lifetime earnings by $3,500 (Chetty et al. 2014). The ISP program costs roughly $30,000/800 pupils = $37 per pupil – implying a cost-benefit ratio of almost 1 to 100. Put differently, from Jackson and Mackevicius (2024) this effect is similar to the average effect of increasing school spending by about $1,700 per pupil. Comparing it to highly effective uses of funds (not just the average), the average ISP effect is similar to the 95^{th} percentile of the distribution of effects of increased school spending by $800 per pupil. In sum, it compares very favorably to effective uses of school resources. The relative cost-effectiveness of this intervention is not surprising because (as the theory suggests) the benefits arise not from a movement along the production possibilities frontier (as with hiring more teachers or hiring better teachers) but rather a shift toward the frontier through increased allocative efficiency or effort. It is worth noting that while the average effects indicate very high cost-effectiveness, the overall effects are not large enough to bring *all* children up to the proficiency standard (the average proficiency rate is around 47 percent). However, after just a few years, they are economically meaningful (cutting the non-proficiency rate by around a tenth) and can be achieved at low-cost.

VI.2 Threats to Validity

The estimates in Section VI.1 rely on two identifying assumptions: (1) the treated school's trajectory would have been similar to that of the comparison schools in the absence of any intervention, and (2) no other confounding policies or changes systematically affected the schools simultaneously with the ISP designation. Providing evidence for assumption (1), I show (a) the similarity in outcome trends between the treated and comparison schools before ISP designation (indicating that the comparison schools serve as a suitable counterfactual for the treated schools), and (b) comparable pre-treatment trends of observed predictors of outcomes between the ISP and comparison schools. To provide evidence for assumption (2), I show that, despite significant changes in outcomes following the ISP designation in ISP schools, observed predictors of outcomes *do not* exhibit changes after receiving the ISP designation (indicative of the absence of confounding shocks). These observed patterns support a causal interpretation of the documented effects in Section VI.1.

To show that the treated schools were on a similar trajectory as comparison schools before ISP destination, I first look to the event study models. While there is no visible evidence of differential pre-trending, I use a more formal test. For ELA and Math passing rates, the p-value associated with the joint hypothesis that all the pre-treatment year effects (year t-2 through t-9) are equal to that in year t-10 is 0.57 and 0.96, respectively. That is, the formal tests are consistent with what is visually apparent – no differential pre-trending in outcomes.

To examine differential pre-trending not just on the outcome but also strong predictors of the outcome, I create predicted student performance based on a linear regression of the passing rate in ELA and math on school enrollment, the percent of students on free and reduced-price lunch, and the proportion of reduced-price lunch before 2016. The R-squared is above 0.6, indicating a strong predictor of demographic and other changes within schools. Importantly, the predicted passing rates strongly predict variation within schools over time, and across schools at any point in time. Using this prediction, I estimate event-study models on *predicted* outcomes and present this in Figure 4. For both ELA and Math, there is no visual evidence of differential pre-trending in strong predictors of passing rates and little evidence of a change in predicted outcome after ISP designation. In *predicted* ELA and Math models, the p-value associated with the joint hypothesis that all the pre-treatment year effects (year t-2 through t-9) are equal to that in year t-10 yields a p-value of 0.77 and 0.55, respectively. The formal tests are consistent with what is visually apparent – no differential pre-trending in *predicted* outcomes.

Next, I turn to the evidence of no confounding. Similar to a test for smoothness of latent outcomes (as modeled by covariates) in regression discontinuity models, I estimate event-study on predicted outcomes and test for a sudden change in predicted outcomes after ISP designation. It is clear that for both math and ELA passing rates, despite a clear improvement in actual outcomes after ISP destination, there is no apparent change in predicted outcomes after ISP designation. As summarized in Table 2, the ISP designation effect on predicted ELA and Math rates were 0.366 (p-value>0.1) and 0.266 (p-value>0.1), respectively. To assuage concerns that this result is driven by how the individual characteristics are weighted to create predicted outcomes, Table 3 reports regression estimates on ISP designation on individual predictors of passing rates. Of the 32 estimates, none are significant at the 5 percent level, and none of the before versus after estimates is significant at the 10 percent level – indicating little change in observable predictors of outcome.

As a final check, I focus on results that use the single best non-ISP matched school as a comparison – where bias due to imperfect matching is minimized. Consistent with this, the ISP effects on observed covariates are *very* small when only using the best single match (See Appendix Table A2). The estimated effect on predicted math and ELA passing rates are merely 0.0072 and -0.0391, respectively. This suggests that these schools form a very good counterfactual for what would have happened at ISP schools where there is no evidence of confounding. Accordingly, the estimated ISP effect using this sample may be a cleaner estimate of the true average ISP effect. Reassuringly, the effects are very similar (actually larger) – indicative of the estimated gains being real.

Because I only use aggregate school-level data for the analysis, the lack of effect of predicted outcomes is important. That is, one may worry that the improved outcomes could be driven by motivated parents deciding to send their children to schools after the principal is designated an ISP – leading to improved outcomes due to compositional changes rather than the effects of autonomy *per se*. To assuage these concerns I examine the ISP effects on individual measures of student composition in Table 3. I find no effects on student demographic composition (measures of income, ethnicity, language spoken at home, special education status, and measures of mobility (a direct measure of students transferring in and out of schools over time). While this does not rule out changes in *unobserved* dimensions of student composition, the results suggest that this is unlikely to drive the results. It is also worth noting that elementary school enrollment is driven mainly by residential location, so changes in residential location would have to be implausibly large (and only in unobserved dimensions) to generate the composition effects that could drive the main effects.

In sum, the evidence shows that both actual outcomes and predicted outcomes were on very similar trajectories for ISP and comparison schools before ISP designation. Importantly, predicted outcomes continued to be very similar between ISP and comparison schools after ISP designation, while actual ELA and math scores improved sharply for ISP schools only after ISP designation. Taken together, they indicate that the common time shocks assumption is likely valid and that the no confounding effects assumption likely holds. Moreover, Section VI.3 will present several patterns in the data predicted by the theoretical model that are inconsistent with a selection story.

VI.3 Heterogeneity

The results indicate that the ISP effect is positive, *on average*. Indeed, the standard errors indicate that the average effects are clearly greater than zero. It is important to note that the precision of the average effect is not an indication of a lack of effect heterogeneity. Indeed, I will demonstrate that while the pooled average is clearly above zero (with a tight confidence interval for the pooled average), one may observe negative true effects a sizable amount of the time. Indeed, the framework in Section III.2 suggests that the ISP effects will be heterogeneous (i.e., above and beyond that explained by sampling variability) and that some effects may be negative. By exploiting the stacked data structure, I can test this directly. Because each school has its own matched set of control schools, one can estimate the treatment effects of each ISP individually. That is, for each treated school (and the comparison schools matched to that treatment school), I estimate a simple difference-in-difference specification as below (using only data from group g).

$$Y_{sgt} = \beta_g(ISP_s \times 1_\tau) + \gamma_s + \gamma_{t,g} + \varepsilon_{sgt} \tag{4}$$

This model yields individual ISP effects ($\hat{\beta}_g$) for each school that received the ISP designation between 2016 and 2018. Because the individual ISP effects are estimated with noise (the variance of which is approximated by the squared standard error, se_g^2), the spread of the estimates will overstate the spread of true causal effects. I use an approach motivated by a hierarchical model to uncover the distribution of true effects.

Each school has a real ISP effect θ_g, and due to sampling error, estimates are distributed as in (5). This normality assumption follows from the central limit theorem.

$$\hat{\beta}_g \sim \mathcal{N}(\theta_g, \sigma_g^2) \tag{5}$$

The true effects for individual schools deviate from the grand mean (Θ) due to heterogeneity with variance τ^2 according to some unknown distribution $g(\tau)$. The empirical challenge is to obtain estimates of the variability of true effects (τ^2), and, using this, estimate the distribution $g(\tau)$.

Assuming that the sampling variance for each estimate (σ_g^2) is well approximated by its standard error squared (se_g^2), one can uncover estimates of true variability – i.e., the variability in estimates not attributable to sampling variability. To estimate τ, I follow DerSimonian and Laird (1986), which takes a precision-weighted variance of the raw estimates, and subtracts the variance expected due to sampling variability (based on the *observed* estimation errors) to provide an estimate of the variance of true heterogeneity (τ^2). This approach does not rely on any distributional assumptions regarding heterogeneity. The estimates are reported in Table 4. Note that this table also reports a precision-weighted average ISP effect across all schools.[11] This is a more efficient estimate of the average effects than the equal-weighted estimates from Table 2.

Pooling both subjects (and accounting for clustering at the school level), the estimated heterogeneity (τ) is 5.97, while it is 5.63 for ELA and 6.21 for math. One cannot reject that the distribution of effects is the same for the two subjects. This indicates that for any two randomly chosen schools, the true ISP effect on passing rates will typically differ by about six percentage points. Given that the average pooled effect reported in column 1 is 3.666 percentage points (p-value$<$0.01), this suggests a large degree of true heterogeneity such that some true effects may be negative while others are large and positive. I present more evidence on this below.

The Distribution of True Effects

To estimate the distribution of true effect $g(\tau)$ based on the observed estimates, I use a deconvolution kernel density estimator (Carroll and Hall (1988)) introduced by Kato and Sasaki (2018) that identifies the distribution of X_i even if X_i is not observed, but where two measurements X_{1i} and X_{2i} of X_i are observed.[12] To obtain two measures of the ISP effect for each school, I estimate the treatment effect on math and ELA passing rates (for each school)– treating these as two measures of the school's ISP effect. To assess the plausibility of this assumption, I test whether the ISP effects on math and reading have the same mean using a two-sample t-test, and have the same distribution using the Kolmogorov – Smirnov test. Using both approaches, I fail to reject equality of effects across subjects such that the ISP effects on Math and ELA can be reasonably considered multiple measures of the same underlying treatment effect.

This approach does not require that the measures be independent, and allows for estimating a uniform confidence band for the density. Setting the tuning parameter (in this case, the kernel bandwidth) so that the resulting density matches moments in the data helps balance local and global properties of the density (Efron and Tibshirani, 1996). As such, following Walters (2022) and Jackson and Mackevicius (2024), I select the bandwidth so that the standard deviation of the true effects (based on the deconvolved distribution) equals the unbiased estimate of the standard

[11] That is, each estimate has weight $1/(se_g^2 + \hat{\tau}^2)$.

[12] This approach approximates the true distribution by fitting it to a Fourier transform.

deviation of true effects (from above). Figure 5 plots the distribution of ISP effects, the deconvolved density, and the implied distribution assuming normality. It also includes the distribution of raw estimates for math and ELA. The figure shows that the distribution of estimated effects covers a wide range of values, from negative 15 percentage points to positive 24 percentage points. However, this range may be due to sampling variability rather than true treatment heterogeneity. Despite this, the deconvolved estimated density of true effects is similar to that of the raw effects. The deconvolved density covers effects both below and above zero, indicating that real ISP effects can be both negative and large or positive and large.

Assessing Normality

To assess whether the distribution of effects differs from a normal distribution, I include a normal distribution plot with the same spread but centered on the deconvolved average.[13] The deconvolved distribution has a bell shape similar to a normal distribution and its confidence intervals overlap with the normal density. This suggests that the distribution of true effect is approximately normal. As a formal test of this hypothesis, I implement a test in Jackson and Mackevicius (2024) following Wang and Lee (2020). Specifically, I implement the Shapiro-Wilk normality tests on appropriately standardized effect sizes– an approach found to detect deviations from normality. This test fails to reject that the distribution of effects is normal with a p-value of 0.71. A quantile-quantile plot of the estimates is in Appendix Figure A5 – showing minimal deviation from normality.

This figure illustrates two key points. Firstly, the distribution is centered above zero, indicating that the majority of ISPs experience positive effects. However, there are also a significant number of negative estimates. In fact, the deconvolved density suggests that approximately one-quarter of the true effects are below zero. Additionally, the deconvolved distribution shows that around two-thirds of the effects fall between -3 and 9 percentage points, while improvements greater than 20 percentage points are highly unlikely. This suggests that there is considerable treatment heterogeneity with generally positive effects, but with some true effects being notably negative. I will now explore whether this heterogeneity can be accounted for by the mechanisms outlined in the theoretical framework.

VI.4 Explaining the Heterogeneity and Testing the Theory

Testing the Allocative Efficiency Chanel

The theoretical framework highlights that one of the mechanisms through which greater autonomy (or decentralization) can affect outcomes is through its effect on allocative efficiency. The

[13]The deconvolved average is computed by integrating values of treatment effects over the estimated density distribution. The deconvolved estimated mean is 2.968. This is slightly below the meta-analytic mean because the model estimates a slightly smaller density at the lower tails and slightly less at the upper tails than implied by the raw estimates.

model highlights two important dimensions: principal-school alignment (a proxy for principal quality narrowly construed) and heterogeneity (school-district alignment). This section presents evidence of both mechanisms that undergird the allocation efficiency channel.

Principal Alignment/Quality

A key prediction from the theoretical framework is that principals more aligned toward improving test scores will have larger ISP effects on math and ELA passing rates. I test this notion using two separate measures of principal alignment toward improved outcomes: (a) The first is the residual ELA and Math passing rate in 2015 unexplained by the school poverty rate. This is a measure of the extent to which a school (in 2015) had better outcomes than would have been predicted based on the socioeconomic status and demographics of the student population. The underlying logic is that school principals that tended to have better achievement than would be expected in the past (by revealed preference) are likely to be those whose orientation was already toward improved student achievement. (b) The second measure is a proxy for principal quality (measured in 2016) based on teacher survey responses.[14] While this measures leadership oriented toward "sustained improvement" in a general sense, academic achievement is explicitly part the districts definition of success (based on the school-rating criteria), and Laing et al. (2016) show that this measure is correlated with principal test-score value-added. While I will show relationships for each individually, to summarize principal quality in a single metric, I created a principal quality index, which is the standardized average of the z-scores for the two measures (as in Kling et al. (2007)).

Using these measures, I regress the ISP effect (the estimated β_gs from Equation (4)) on the measured principal alignment proxy. If the *true effects* are normally distributed (as shown above), and the sampling variability is unrelated to the true effect,[15] the optimal weight for each observation is the inverse of its precision $1/(se_g^2+\tau^2)$. This regression model is implemented by weighted-least-squares. This is analogous to the approach used in Card and Krueger (1992), but differs in that it also accounts for noise due to true heterogeneity (not just sampling variability). Note that this is also a standard random effect meta-regression (Berkey et al. (1995)).

Both measures reveal the same basic pattern – schools with "better" principals (i.e., more aligned toward improved achievement) have more positive ISP effects. Specifically, a one standard deviation increase in principal alignment is associated with an IPS effect on ELA passing rates 3.25 percentage points higher (p-value$<$0.01) and math passing rates 1.77 percentage points higher (p-value$<$0.1). Looking at the different dimensions, both the leadership and residual pass

[14]Note that the relationship between principal alignment and the ISP effects is very similar (*albeit less precise*) when restricting to schools that are designated an ISP in 2017 and 2018 – assuaging concerns that the principal quality measure in endogenous.

[15]A regression of the estimated effects against their precision yields p-value above 0.1 – suggesting that this condition is satisfied.

rate measures predict larger ISP effects for both subjects. However, these relationships are only statistically significant for ELA passing rates (columns 2 and 3). Pooling both subjects (column 12), the coefficient on this standardized measure is 3.009 (p-value<0.01) and one cannot reject that there are differences in the relationship between math and ELA passing rates at the 10 percent level. That is, the ISP effect is 3 percentage points larger for a principal that is at the 85^{th} percentile of the alignment distribution than one at the average. The pooled average effect is 3.66 percentage points, so the average ISP effects are positive for principals with z-scores greater than -1.222. That is, the ISP improves outcomes for most principals, other than those with measured alignment (a proxy for quality) in the bottom 12 percent of the distribution of the ISP principals. The fact that the negative effects are concentrated among principals that tend to have poor outcomes historically or teachers rate poorly on leadership skills strongly suggests that the negative ISP effects are due to some principals not making decisions that are well aligned with improving student achievement.

Heterogeneity (District-School Alignment)

Another key prediction from the framework is that the degree of alignment may matter. While this is hard to measure directly, I proxy for having specific needs that may not be well aligned with the generic district position by having a student population that is demographically quite different from the typical school in the district. Specifically, I code schools as outliers if they have more than the 95 percentile of percent white or percent Hispanic. I do not do this for percent black because many schools have more than 95 percent black, while this is not true for the other ethnic categories. I also code schools as outliers if they have more than the 95th percentile of percent free and reduced-price lunch, and more than the 95th percentile of percent bilingual and percent special education. Schools are coded as an outlier if they fall into any one of these categories, and I also code up the number of categories. I then regression the ISP effect on an indicator for outlier status.

For both subjects, there is strong evidence that outlier schools have larger ISP effects. The coefficient on the outlier indicator is 4.65 (p-value<0.05) for ELA passing rates and 4.445 (p-value<0.05) for math passing rates. This indicates that, *all else equal*, the ISP effect is considerably larger for schools that are outliers in at least one of these demographic categories. In pooled models (both subjects), one fails to reject that the effect is the same for both subjects, and the coefficient on outlier is 4.004 (p-value<0.05). That is, all else equal, schools that are outliers in some demographic category (a proxy for having specific needs that may not be well aligned with the district choices) have ISP effects that are about four percentage points larger than those that are not. Note that four percentage points corresponds to an increase of more than 0.1σ – an economically significant effect by most metrics. This indicates that the benefits to increased autonomy (or decentralization) may be very large in settings with considerable heterogeneity (as indicated in both Oates (1999) and Acemoglu et al. (2007)).

How Much Heterogeneity Does this Explain

To present visual evidence of these mechanisms, Figure 6 plots the standardized principal alignment measure against the raw estimated ISP effects. As indicated by the regression results, there is a clear positive relationship between the two. Indeed, the meta-regression indicates that this measure explains roughly 15 percent of the unobserved heterogeneity in effects. The right panel presents a box plot of the raw estimated ISP effect against the number of categories in which a school is an outlier. As one can see, not only is there an increase in the ISP effect from 0 to 1, but there is a large jump from 1 to 2 – that is, being an outlier in two categories is associated with an even larger ISP effect than being an outlier in only one or none. This lends further credibility to the heterogeneity result. To show that this pattern is robust to the number of matches used, Appendix Figure A6 shows figures analogous to Figure 6 with very similar patterns using different numbers of matches. In meta-regression model, both the principal alignment measure and indicators for the number of outlier categories explain roughly one-quarter (R^2 is 0.2697) of the true heterogeneity in treatment effects. That is, using relatively crude proxies for the extent to which granting a school greater autonomy is likely to increase allocative efficiency explains more than a quarter of the heterogeneity in effects across schools.

Testing the Stability Channel

The other mechanism highlighted in Section III.2 is a stability channel. A primary motivation for the ISP program was retaining effective principals by making the job more rewarding (via increased autonomy) and also requiring a commitment to stay at the current school for two years. For both reasons, one would expect reduced turnover in the first two years. However, if there were no motivation effects (i.e., the principals only stayed because of the commitment), then turnover would have reverted back to pre-treatment levels after two years. In contrast, if principals value the job more (because they value the increased autonomy), then the reduction in turnover would be sustained past the first two years of the ISP.

To investigate this, I estimated the ISP's effect on the likelihood that a school has a new principal in the first two years after ISP designation and also in the years after this. The regression estimates are in Table 6. As one can see in column 1, ISP schools are roughly 2.57 percentage points less likely to have a new principal in the first year of being an ISP. To some extent, this is mechanical because if the principal had left the school, it would not have been an ISP. However, as one can see, this effect persists in year 2 (where the point estimate is largely the same at 2.27 percentage points) – indicative of some persistent reduction in principal turnover among ISPs. Again, given the two-year commitment, this reduction is not surprising. However, the reduction in year 3 is similar to that in years 1 and 2, indicative of some persistent reduction in turnover that is not driven only by the two-year commitment to stay at the current school. I also examine effects on a publicly available measure of principal turnover. The district reports the fraction of principals in the school in a given year that were not at the school six years before. As with the new principal measure, there is a notable reduction in principal turnover upon ISP adoption (see column 2), indicating that some principals who might have otherwise left their schools chose to stay due to the ISP program. Importantly, the reduction in principal turnover lasted for more than two years – indicative that school principals may have been happier under the ISP program than before. Note interpretation is corroborated by principals' accounts, indicating that before they became ISP, they "*started looking for other jobs*" and that they "*described independence as a motivational boost*" and "*felt rewarded for their demonstrated success*" Travlos (2020).

VI.5 Evidence on Mechanisms

By design, schools are expected to implement diverse strategies when granted increased autonomy, leading to inherent heterogeneity in the changes made by individual schools. This heterogeneity poses a challenge in uncovering the mechanisms underlying the effects of the ISP. However, certain variables can offer valuable insights in this regard.

School Climate: Theoretically, reduced principal turnover and increased effort should contribute to greater stability and an enhanced school climate. Furthermore, evidence suggests that changes in principal decision-making, such as implementing positive reforms or adopting practices valued by teachers and students, would translate into improved school climate. To assess this, the study estimates the average effect of the ISP on school climate, measured on a scale of 1 to 5. Consistent with this notion above, ISP schools witnessed an increase in school climate by approximately 0.3 points (p-value$<$0.01)—equivalent to an effect size of roughly 0.23σ. This finding indicates that teachers and students perceive ISP schools as having become better learning environments following the ISP designation.

To evaluate whether these observed increases in school climate align with the effects on test scores, this analysis draws upon two relevant studies. Porter et al. (2023) highlights that a one-standard-deviation (1SD) increase in school climate corresponds to approximately a 0.8 increase in effectiveness, while Jackson et al. (2023) indicates that a (1SD) increase in effectiveness results in a 0.08σ improvement in test scores. As such, the observed increase in school climate may be associated with a $0.08\sigma \times 0.8 = 0.066\sigma$ rise in test scores. Notably, this estimate closely aligns with the best linear unbiased predictor of the ISP effect. While it is unclear whether the improved school climate led to the gains in test scores or *vice versa*, or if both changes were caused by general improvement resulting from the ISP, it is clear that the program facilitated an enhancement in school climate.

Given the improvement in school climate, one might expect to observe an improvement in student attendance and a reduction in teacher absenteeism. I assess this using the regression model

(Table 6 columns 3, 4, and 5). There is no evidence that ISP schools saw a reduction in the chronic truancy rate or an increase in the student attendance rate. Also, there is no evidence of a reduction in teacher absenteeism. This suggests that the gains observed at the school are not driven by students and teachers showing up to school more often (essentially a kind of dosage mechanism), but rather by improvements in the schooling environment conditional on attendance. It is worth noting that the lack of any effect on student attendance measures further reinforces the fact that the achievement gains are not likely to be driven by changes in student composition.

Inputs: As suggested by theory, each school has unique needs, so specific school policies employed should differ from one school to another. Accordingly, one would not expect similar policy or input changes at ISP schools on average. I assess this notion empirically. One crude measure for changes in policy is changes in personnel spending categories. Because schools spend on more than personnel, this is an imperfect reflection of resources, and shifting away from or toward non-personnel spending will not be observed. However, most public schools spend about 70 percent of their budgets on personnel (Jackson et al. 2016), so an analysis of personnel spending may be instructive. Moreover, the categories of spending are much more detailed than in many other quantitative studies, so there may be something to learn from such data.

Using personnel data linked to individual schools, I calculate school spending on various types of personnel (principals, teachers, and other). The other category is further broken down into spending on teaching assistants, nurses, social workers, counselors, bilingual teachers, custodial staff, school bus staff, lunch staff, and misc. I estimate the ISP effect on the natural log of school spending in each category and report these effects in the top panel of Table 7. Looking at all ISP schools, there is no appreciable change in overall personnel spending. The coefficient on the log of personnel spending is 0.0029 (p-value$>$0.1). Somewhat mechanically, there is a 9 percent increase in spending on principal salary (this is part of the ISP package). However, there is no change, on average, in teacher salary spending or spending on non-teacher personnel. This is consistent with no observed change in average class size shown in Table 6. Breaking up the "other" spending categories does reveal some shifts in school spending. Specifically, on average, ISP schools reduced special education personnel spending by about 8.19 percent (p-value$<$0.05) and increased spending on school counselors by roughly 8.49 percent (p-value$<$0.05) — indicative of a shifting away from special education teachers toward more counselors. These two roles are likely highly substitutable for some tasks but not others. The key differences are that special education teachers may be involved in in-class pedagogy while counselors are not. While these patterns are interesting, given the heterogeneous needs of schools, it is difficult to interpret these average effects.

Ideally, one would observe a suboptimally low allocation of spending in a particular high-needs area and then see if those schools increased spending in that area after being granted more autonomy. I approximate this thought experiment by interacting the ISP indicator with indicators for whether a school was an outlier in the proportion of bilingual students or the percentage of special education students. Because I cannot observe whether spending was suboptimally low in special education or bilingual education for these schools before ISP designation, it is more likely to be true for these schools than for other schools. As such, while this test cannot disprove this kind of behavior, it can potentially confirm it.

The lower panel of Table 7 presents these results. I focus on spending in the separate sub-categories of other personnel. The pattern of results is consistent with greater alignment of funding toward school-specific priorities for the special education outlier schools but less so for bilingual education schools. Overall, both outlier schools increase spending for "other" personnel by between 14 and 20 percent (both significant at the 5 percent level). Also consistent with greater alignment of budgets toward school-specific needs, schools with high shares of bilingual students increase spending on bilingual education by a modest five percent (0.0638 - 0.0102 = 0.0536), and schools with high shares of special education students increase spending on special education by a sizable 44 percent (0.556 - 0.111 = 0.445). However, only the effect for special education is statistically significant. The results in column 8 indicate that schools with large shares of special education students reduced spending on teaching assistants, indicating some substitution away from general education in-class teacher support towards teacher support focused on students with specialized needs. This is precisely the shifting of resources toward the special needs of a school that may lead to some allocative efficiencies. Finally, in column 9, one can see that, although not statistically significantly different, the outlier schools both spend more on counselors than ISP schools with more typical student populations. Again, this is consistent with schools with more specific student needs using the added budget flexibility to focus funds toward inputs more tailored to their needs than the typical school.

VI.6 Extrapolating to All Schools

Because the principals (and schools) that opted to gain more autonomy likely faced the lowest costs and/or the highest gains, the benefits to voluntarily gaining greater autonomy, such as with the ISP program and also those in the UK (e.g., Clark (2009)) may overstate the benefits for the average school. To speak to this point, I provide evidence on the treatment's effect on the treated (ATT) and the average treatment effect (ATE). The difference between these distributions will determine how applicable the results are to different settings. We can measure principal alignment and have proxies for alignment for all schools. Assuming that the relationship between these observable measures and the ISP effect is the same among ISP principals as all principals, we can use the distribution of principal alignment and school alignment to estimate the effect of expanding ISP-type autonomy to all schools. I use a meta-regression with both variables to construct predicted ISP effects for all schools.

Figure 7 shows a density plot of the predicted ISP effect for both ISP and non-ISP schools. Surprisingly, the distribution of predicted effects among actual ISP schools is similar to that for non-ISP schools – indicating that schools granted autonomy would benefit similarly to those that were not. Indeed, a Kolmogorov–Smirnov test for differences between the distributions of predicted effects for ISP and non-ISP schools yields a p-value of 0.53. The similarity in predicted effects reflects two offsetting and opposing forces: ISP schools have principals with about 0.5 standard deviations better alignment but are about 25 percent less likely to be outlier schools.

The similarity in the distributions of predicted effects between ISP schools and non-ISP schools indicates the potential for significant gains by extending greater autonomy to a larger number of schools. The data reveals weakly positive ISP effects in more than 80 percent of schools. A rough estimation suggests that expanding greater autonomy to the top 85 percent of schools with the most promising predicted effects (those with exemplary principals and unique needs) could raise the average passing rates at the district level by approximately five percentage points (equivalent to around 0.12σ). Alternatively, adopting a conservative approach and assuming that the impact of implementing the ISP across all schools aligns with the estimated BLUP for the treated ISP schools (noting that the predicted effects are actually larger for non-ISP schools than ISP schools), one could anticipate an effect size of 0.06σ, corresponding to an increase in passing rates of roughly 2.4 percentage points. While effects of this magnitude are not transformative, the cost-effectiveness of this policy makes it a relatively easy way to improve outcomes. It is important to acknowledge that this extrapolation (a) assumes that the observed predictors of the ISP effect operate similarly for both ISP and non-ISP schools, and (b) does not account for any potential general equilibrium effects that may arise if all schools were granted greater autonomy. Considering these factors, the policy extrapolation should be viewed as merely suggestive.

VII Conclusions

Contemporary policy reforms often involve granting more decision-making and budgetary autonomy to school principals, but the evidence of improved student outcomes is limited to specific contexts or lacks conclusive results due to validity and statistical power concerns. Theoretical work suggests that the effects of such policies vary across settings, which is confirmed by a meta-analysis of design-based studies in Section II. Although there is evidence of an association between increased autonomy and better outcomes in certain settings (e.g., high accountability and capacity), this paper goes beyond associations by presenting a theoretical framework and empirically testing it using design-based models.

This paper helps make sense of the documented associations in the literature and provides evidence on the importance of context by (a) presenting a theoretical framework that highlights the settings in which autonomy can positively and negatively affect student achievement and (b) testing the model using design-based empirical approaches. This paper provides evidence of a significant impact of principal autonomy in a large urban environment in the United States. The findings align with those found in highly competitive settings. However, I found considerable variability in the effect across schools that cannot be attributed to sampling variability alone. This variability is consistent with the theoretical framework, which suggests that schools with high-quality principals, and student populations requiring atypical policy decisions benefited more from autonomy. Furthermore, schools in which the ISP resulted in greater stability, as evidenced by reduced principal turnover and improved school climate. This finding suggests that increased stability (afforded by the reduced principal turnover) also played a significant role in the success of the schools. An examination of school personnel spending reveals patterns consistent with schools using increased autonomy to cater to the specific needs of their students – in line with predictions from canonical work in public finance (Oates (1999) and others).

The patterns observed highlight the benefits of increased school autonomy, particularly in heterogeneous settings (Acemoglu et al., 2007). However, they also indicate that autonomy should be granted to effective and motivated school leaders and may lead to worse outcomes in settings with agency problems or low principal capacity, as suggested by comparative descriptive work across nations (Fuchs and Wößmann, 2007) or states within nations (Loeb and Strunk, 2007). The results underscore the significance of leader (principal) and management quality (Bloom et al., 2015; Branch et al., 2012) and emphasize the need to consider context when examining policies in different settings (Jackson and Mackevicius, 2024). Finally, the findings suggest that granting more autonomy to high-quality school leaders may improve student outcomes at minimal cost.

References

Atila Abdulkadiroğlu, Joshua Angrist, Susan Dynarski, Thomas J. Kane, and Parag Pathak. Accountability and flexibility in public schools: Evidence from boston's charters and pilots. *The Quarterly Journal of Economics*, 126:699–748, 5 2011. ISSN 00335533. doi: 10.1093/QJE/QJR017.

Daron Acemoglu, Philippe Aghion, Claire Lelarge, John Van Reenen, and Fabrizio Zilibotti. Technology, information, and the decentralization of the firm. *The Quarterly Journal of Economics*, 122:1759–1799, 11 2007. ISSN 0033-5533. doi: 10.1162/QJEC.2007.122.4.1759.

Philippe Aghion and Jean Tirole. Formal and real authority in organizations. *Journal of Political Economy*, 105(1):1–29, 1997. ISSN 00223808, 1537534X.

C. S. Berkey, D. C. Hoaglin, F. Mosteller, and G. A. Colditz. A random-effects regression model for meta-analysis. *Statistics in Medicine*, 14:395–411, 2 1995. ISSN 1097-0258. doi: 10.1002/SIM.4780140406.

Marianne Bertrand and Antoinette Schoar. Managing with style: The effect of managers on firm policies. *The Quarterly Journal of Economics*, 118:1169–1208, 11 2003. ISSN 0033-5533. doi: 10.1162/003355303322552775.

Diether W. Beuermann and C. Kirabo Jackson. The short- and long-run effects of attending the schools that parents prefer. *Journal of Human Resources*, 57:725–746, 5 2022. ISSN 0022-166X. doi: 10.3368/JHR.57.3.1019-10535R1.

Diether W Beuermann, C Kirabo Jackson, Laia Navarro-Sola, and Francisco Pardo. What is a good school, and can parents tell? evidence on the multidimensionality of school output. *The Review of Economic Studies*, 90:65–101, 1 2023. ISSN 0034-6527. doi: 10.1093/RESTUD/RDAC025.

John Bishop, Ludger Wossmann, John Bishop, and Ludger Wossmann. Institutional effects in a simple model of educational production. *Education Economics*, 12:17–38, 4 2004. ISSN 0964-5292. doi: 10.1080/0964529042000193934.

Nicholas Bloom, Renata Lemos, Raffaella Sadun, and John Van Reenen. Does management matter in schools? *The Economic Journal*, 125:647–674, 5 2015. ISSN 1468-0297. doi: 10.1111/ECOJ.12267.

Gregory F Branch, Eric A Hanushek, Steven G Rivkin, and Texas Schools. Estimating the effect of leaders on public sector productivity: The case of school principals. 2 2012. doi: 10.3386/W17803.

Carolina Caetano and Brantly Callaway. Difference-in-differences with time-varying covariates in the parallel trends assumption. *Papers*, 2023.

Brantly Callaway and Pedro H.C. Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225:200–230, 12 2021. ISSN 0304-4076. doi: 10.1016/J.JECONOM.2020.12.001.

David Card and Alan B Krueger. Does school quality matter? returns to education and the characteristics of public schools in the united states. *The Journal of Political Economy*, 100:1–40, 1992.

Raymond J. Carroll and Peter Hall. Optimal rates of convergence for deconvolving a density.

Journal of the American Statistical Association, 83:1184–1186, 1988. ISSN 1537274X. doi: 10.1080/01621459.1988.10478718.

Doruk Cengiz, Arindrajit Dube, Attila Lindner, and Ben Zipperer. The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics*, 134:1405–1454, 8 2019. ISSN 0033-5533. doi: 10.1093/QJE/QJZ014.

Raj Chetty, John N. Friedman, and Jonah E. Rockoff. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, 104(9):2593–2632, September 2014. ISSN 0002-8282. doi: 10.1257/aer.104.9.2593.

Hanley Chiang, Stephen Lipscomb, and Brian Gill. Is school value added indicative of principal quality? *Education Finance and Policy*, 11:283–309, 7 2016. ISSN 15573079. doi: 10.1162/EDFP_A_00184.

Damon Clark. The performance and competitive effects of school autonomy. *Journal of Political Economy*, 117:745–782, 8 2009. ISSN 00223808. doi: 10.1086/605604.

Sarah R. Cohodes and Katharine S. Parham. Charter schools' effectiveness, mechanisms, and competitive influence. 2 2021. doi: 10.3386/W28477.

Massimo G. Colombo and Marco Delmastro. Delegation of authority in business organizations: An empirical test. *Organizational Behavior*, 52:53–80, 3 2004. ISSN 00221821. doi: 10.1111/J.0022-1821.2004.00216.X.

Edward L. Deci and Richard M. Ryan. Intrinsic motivation and self-determination in human behavior. *Intrinsic Motivation and Self-Determination in Human Behavior*, 1985. doi: 10.1007/978-1-4899-2271-7.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, September 1986. ISSN 01972456. doi: 10.1016/0197-2456(86)90046-2.

Manasi Deshpande and Yue Li. Who is screened out? application costs and the targeting of disability programs. *American Economic Journal: Economic Policy*, 11:213–48, 11 2019. ISSN 1945-7731. doi: 10.1257/POL.20180076.

Bradley Efron and Carl Morris. Stein's Estimation Rule and its Competitors—An Empirical Bayes Approach. *Journal of the American Statistical Association*, 68(341):117–130, March 1973. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1973.10481350.

Bradley Efron and Robert Tibshirani. Using specially designed exponential families for density estimation. *https://doi.org/10.1214/aos/1032181161*, 24:2431–2461, 12 1996. ISSN 0090-5364. doi: 10.1214/AOS/1032181161.

Andrew Eyles and Stephen Machin. The introduction of academy schools to england's education. *Journal of the European Economic Association*, 17:1107–1146, 8 2019. ISSN 1542-4766. doi: 10.1093/JEEA/JVY021.

Andrew Eyles, Stephen Machin, and Sandra McNally. Unexpected school reform: Academisation of primary schools in england. *Journal of Public Economics*, 155:108–121, 11 2017. ISSN 0047-2727. doi: 10.1016/J.JPUBECO.2017.09.004.

Bernd Frick, Robert Simmons, Bernd Frick, and Robert Simmons. The impact of managerial quality on organizational performance: Evidence from german soccer. 2007.

Thomas Fuchs and Ludger Wößmann. What accounts for international differences in student performance? a re-examination using pisa data. *Empirical Economics*, 32:433–464, 5 2007. ISSN 03777332. doi: 10.1007/S00181-006-0087-0/METRICS.

Sebastian Galiani, Paul Gertler, and Ernesto Schargrodsky. School decentralization: Helping the good get better, but leaving the poor behind. *Journal of Public Economics*, 92:2106–2120, 10 2008. ISSN 0047-2727. doi: 10.1016/J.JPUBECO.2008.05.004.

Jason A. Grissom, Demetra Kalogrides, and Susanna Loeb. Using student test scores to measure principal performance. *Educational Evaluation and Policy Analysis*, 37:3–28, 2012. ISSN 19351062. doi: 10.3102/0162373714523831.

Sanford J. Grossman and Oliver D. Hart. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy*, 94:691–719, 8 1986. ISSN 0022-3808. doi: 10.1086/261404.

Eric Hedberg, J Pustejovsky, and E Tipton. robumeta: A macro for Stata. 2017.

Larry V. Hedges, Elizabeth Tipton, and Matthew C. Johnson. Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(1):39–65, January 2010. ISSN 17592879. doi: 10.1002/jrsm.5.

Andrew Dean Ho. A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics*, 34:201–228, 6 2009. ISSN 10769986. doi: 10.3102/1076998609332755/ASSET/IMAGES/LARGE/10.3102_1076998609332755-FIG7.JPEG.

C. Kirabo Jackson. What do test scores miss? the importance of teacher effects on non-test score outcomes. *Journal of Political Economy*, 5 2018. doi: 10.3386/w22226.

C. Kirabo Jackson and Elias Bruegmann. Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics*, 1:85–108, 10 2009. ISSN 1945-7782. doi: 10.1257/APP.1.4.85.

C. Kirabo Jackson and Claire Mackevicius. What impacts can we expect from school spending policy? evidence from evaluations in the u.s. *American Economic Journal: Applied Economics*, page p.43, 2024. doi: 10.4/JQUERY-UI.MIN.JS.

C. Kirabo Jackson, Jonah E. Rockoff, and Douglas O. Staiger. Teacher effects and teacher-related policies. *Annual Review of Economics*, 6(1):801–825, 2014. doi: 10.1146/annurev-economics-080213-040845.

C. Kirabo Jackson, Rucker C. Johnson, and Claudia Persico. The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *The Quarterly Journal of Economics*, 131:157–218, 2 2016. ISSN 0033-5533. doi: 10.1093/qje/qjv036.

C. Kirabo Jackson, Shanette C. Porter, John Q. Easton, Alyssa Blanchard, and Sebastián Kiguel. School effects on socioemotional development, school-based arrests, and educational attainment. *American Economic Review: Insights*, 2:491–508, 12 2020. ISSN 2640-205X. doi: 10.1257/AERI.20200029.

C. Kirabo Jackson, Sebastian Kiguel, Shanette C. Porter, and John Q. Easton. Who benefits from attending effective high schools? *Journal of Labor Economics*, 2 2023. ISSN 0734-306X. doi: 10.1086/724568.

Benjamin F. Jones and Benjamin A. Olken. Do leaders matter? national leadership and growth since world war ii. *The Quarterly Journal of Economics*, 120:835–864, 8 2005. ISSN 0033-5533. doi: 10.1093/QJE/120.3.835.

Thomas J Kane and Douglas O Staiger. Estimating teacher impacts on student achievement: An experimental evaluation. Working Paper 14607, National Bureau of Economic Research, December 2008.

Kengo Kato and Yuya Sasaki. Uniform confidence bands in deconvolution with unknown error distribution. *Journal of Econometrics*, 207:129–161, 11 2018. ISSN 18726895. doi: 10.1016/j.jeconom.2018.07.001.

Jeffrey R Kling, Jeffrey B Liebman, and Lawrence F Katz. Experimental analysis of neighborhood effects. *Econometrica*, 75:83–119, 1 2007.

Cory Koedel and Morgan Polikoff. Executive summary big bang for just a few bucks: The impact of math textbooks in california. *Evidence Speaks Reports*, 2, 2017.

Matthew A. Kraft, David Blazar, and Dylan Hogan. The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, 88:547–588, 8 2018. ISSN 19351046. doi: 10.3102/0034654318759268/ASSET/IMAGES/LARGE/10.3102_0034654318759268-FIG4.JPEG.

Derek Laing, Steven G. Rivkin, Jeffrey C. Schiman, and Jason Ward. Decentralized governance and the quality of school leadership. *NBER Working Papers*, 2016.

Jeffrey B Liebman, Neale Mahoney, Liebman :, and John F Kennedy. Do expiring budgets lead to wasteful year-end spending? evidence from federal procurement. *American Economic Review*, 107:3510–49, 11 2017. ISSN 0002-8282. doi: 10.1257/AER.20131296.

Keke Liu, David Stuit, Jeff Springer, Jim Lindsay, and Yinmei Wan. The utility of teacher and student surveys in principal evaluations: An empirical investigation — american institutes for research. *AIR Report*, 11 2014.

Susanna Loeb and Katharine Strunk. Accountability and local control: Response to incentives with and without authority over resource generation and allocation. *Source: Education Finance and Policy*, 2:10–39, 2007. doi: 10.2307/educfinapoli.2.1.10.

Jessica Merkle. School-level autonomy and its impact on student achievement. *Peabody Journal of Education*, 97:497–519, 2022. ISSN 0161956X. doi: 10.1080/0161956X.2022.2109918.

Carl N. Morris. Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association*, 78(381):47–55, March 1983. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1983.10477920.

Simeon Muecke and Anja Iseke. How does job autonomy influence job performance? a meta-analytic test of theoretical mechanisms. *https://doi.org/10.5465/AMBPP.2019.145*, 8 2019. ISSN 0065-0668. doi: 10.5465/AMBPP.2019.145.

Lorenzo Neri and Elisabetta Pasini. Heterogeneous effects of school autonomy in england. *Economics of Education Review*, 94:102366, 6 2023. ISSN 0272-7757. doi: 10.1016/J.ECONEDUREV.2023.102366.

Lorenzo Neri, Elisabetta Pasini, and Olmo Silva. The organizational economics of school chains. *IZA Discussion Papers*, 2022.

Wallace E. Oates. An essay on fiscal federalism. *Journal of Economic Literature*, 37:1120–1149, 1999. ISSN 00220515. doi: 10.1257/JEL.37.3.1120.

S C Porter, C K Jackson, S Q Kiguel, and J Q Easton. Investing in adolescents: High school climate and organizational context shape student development and educational attainment. *University of Chicago Consortium on School Research*, 2023.

Joseph Regan-Stansfield. Does greater primary school autonomy improve pupil attainment? evidence from primary school converter academies in england. *Economics of Education Review*, 63:167–179, 4 2018. ISSN 0272-7757. doi: 10.1016/J.ECONEDUREV.2018.02.004.

Jonathan Roth, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 4 2023. ISSN 0304-4076. doi: 10.1016/J.JECONOM.2023.03.008.

Matthew P. Steinberg. Does greater autonomy improve school performance? evidence from a regression discontinuity analysis in chicago. *Education Finance and Policy*, 9:1–35, 1 2014. ISSN 1557-3060. doi: 10.1162/EDFP_A_00118.

Matthew P. Steinberg and Amanda Barrett Cox. School autonomy and district support: How principals respond to a tiered autonomy initiative in philadelphia public schools. *Leadership and Policy in Schools*, 16:130–165, 1 2016. ISSN 17445043. doi: 10.1080/15700763.2016.1197278.

Leanna Stiefel, Amy Ellen Schwartz, Carole Portas, and Dae Yeop Kim. School budgeting and school performance: The impact of new york city's performance driven budgeting initiative. *Journal of Education Finance*, 28:403–24, 2003. ISSN -0098-949.

Jerry Travlos. A phenomenological study of chicago's independent school a phenomenological study of chicago's independent school principals principals. 2020.

Ann Katrin van den Ham and Aiso Heinze. Does the textbook matter? longitudinal effects of textbook choice on primary school students' achievement in mathematics. *Studies in Educational Evaluation*, 59:133–140, 12 2018. ISSN 0191-491X. doi: 10.1016/J.STUEDUC.2018.07.005.

John Joseph Wallis and Wallace E. Oates. Decentralization in the public sector: An empirical study of state and local government. *NBER Chapters*, pages 5–32, 1988.

Patrick K Kline Evan Rose Christopher R Walters. Systemic discrimination among large u.s. employers. *The Quarterly Journal of Economics*, 137:1963–2036, 9 2022. ISSN 0033-5533. doi: 10.1093/QJE/QJAC024.

Chia-Chun Wang and Wen-Chung Lee. Evaluation of the Normality Assumption in Meta-Analyses. *American Journal of Epidemiology*, 189(3):235–242, March 2020. ISSN 0002-9262, 1476-6256. doi: 10.1093/aje/kwz261.

Xiao-Feng Wang and Bin Wang. Deconvolution Estimation in Measurement Error Models: The R Package decon. *Journal of Statistical Software*, 39(10):i10, March 2011. ISSN 1548-7660.

Lok Sze Wong, Cynthia E. Coburn, and Ayah Kamel. How central office leaders influence school leaders' decision-making: Unpacking power dynamics in two school-based decision-making systems. *Peabody Journal of Education*, 95:392–407, 8 2020. ISSN 0161956X. doi: 10.1080/0161956X.2020.1800175.

Tables and Figures

Table 1: Summary Statistics for ISP, Non-ISP, and Matched Schools

	1	2	3	4	5	6	7	8	9
	N	mean	sd	N	mean	sd	N	mean	sd
	ISP Schools			All Non-ISP Schools			Matched Non-ISP Schools		
% White	495	0.146	0.2	5,724	0.0719	0.15	6,817	0.158	0.196
% Black	495	0.197	0.279	5,724	0.55	0.421	6,817	0.189	0.296
% Hispanic	495	0.59	0.363	5,724	0.336	0.363	6,817	0.571	0.332
% Special Ed	333	0.122	0.0725	3,817	0.148	0.0863	4,760	0.127	0.0481
% Bilingual	328	0.255	0.192	3,451	0.163	0.18	4,607	0.276	0.173
Enrollment	425	790.4	415.6	4,079	617.1	629.6	6,045	699.2	345.1
Attednance Rate	422	0.955	0.0163	4,069	0.94	0.0417	6,039	0.955	0.0113
Mobility Rate	421	0.0926	0.0653	4,050	0.178	0.135	6,011	0.111	0.0742
Chronic Truancy Rate	408	0.178	0.16	3,823	0.285	0.222	5,839	0.175	0.155
% Free and Reduced Price Lunch	416	0.76	0.262	3,919	0.845	0.209	5,990	0.774	0.25
% Passed ELA	475	50.47	23.27	4,944	36.17	23.67	6,664	47.25	23.12
% Passed Math	475	49.95	26.86	4,943	34.86	27.26	6,664	46.04	27.28
ELA Percentile Score (2014 -2019)	253	70.83	20.24	2,535	51.41	27.41	3,969	66.81	22.3
Math Percentile Score (2014 -2019)	253	73.13	20.13	2,535	48.04	28.14	3,969	65.82	23.38
School Climate Score	265	4.298	0.968	2,870	3.847	1.333	4,030	4.096	1.116

Notes: ISP Schools are those that were designated ISP between 2016 and 2018. This is based on the set of all elementary schools in Chicago Public Schools between 2010 and 2019.

Table 2: Average ISP Effect on Proficiency Rates: Standard and Stacked DiD Models

	Naïve Difference-in-Difference: Full Smaple				Stacked Difference-in-Difference: Matched Sample					
	1	2	3	4	5	6	7	8	9	10
	Passed ELA	Passed Math	Predicted Passed Math	Predicted Passed ELA	Passed ELA	Passed Math	ELA Percentile	Math Percentile	Predicted Passed Math	Predicted Passed ELA
Treated (Year 1)	3.127***	1.787**	-0.116	0.509	2.685***	2.051**	1.212	2.956**	-0.0887	-0.131
	[0.963]	[0.881]	[0.542]	[0.486]	[1.022]	[0.883]	[1.061]	[1.306]	[0.446]	[0.408]
Treated (Year 2)	3.957***	2.688**	0.711	1.508***	3.914***	2.724**	0.878	2.411	0.916*	0.817
	[0.977]	[1.060]	[0.624]	[0.576]	[1.051]	[1.062]	[1.203]	[1.728]	[0.504]	[0.515]
Treated (Year 3)	6.756***	5.576***	1.137	0.933*	6.486***	4.934***	2.430**	4.416**	0.244	0.0729
	[1.768]	[1.856]	[0.767]	[0.532]	[1.767]	[1.693]	[1.075]	[1.902]	[0.572]	[0.506]
Treated (All)	4.423***	3.153***	0.527	0.987**	4.207***	3.113***	1.407	3.136**	0.366	0.266
	[1.031]	[1.075]	[0.576]	[0.462]	[1.072]	[1.034]	[0.970]	[1.448]	[0.416]	[0.399]
Observations	4,849	4,848	5,718	5,718	6,402	6,402	4,220	4,220	6,599	6,599

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

The top and lower panels of this table report coefficients from separate regression models. The top panel reports the dynamic treatment effects for the first three years after ISP destination, while the lower panel reports the simple before versus after comparison.

Table 3: Average ISP Effect on Individual Covariates: Stacked DiD Model

	1	2	8	3	4	5	6	7
	% White	% Hispanic	% Black	% Spec. Ed.	% Bilingual	Mobility Rate	Free-Lunch %	Enrollment
Treated (Year 1)	0.00605	0.00214	-0.00168	0.00106	-0.00247	0.00650*	0.00296	0.474
	[0.00468]	[0.00620]	[0.00506]	[0.00305]	[0.00906]	[0.00376]	[0.00735]	[24.52]
Treated (Year 2)	0.00321	0.00623	-0.00454	-0.00392	-0.00727	0.00491	-0.0122	13.08
	[0.00536]	[0.00716]	[0.00591]	[0.00349]	[0.00848]	[0.00528]	[0.00948]	[22.24]
Treated (Year 3)	-0.00208	0.0137*	-0.0112*	-0.00218	0.00836	0.00634	0.00269	-5.041
	[0.00757]	[0.00798]	[0.00627]	[0.00436]	[0.00984]	[0.00619]	[0.00906]	[31.97]
Treated (All)	0.00273	0.00687	-0.00539	-0.00163	-0.00142	0.00588	-0.00253	3.408
	[0.00545]	[0.00657]	[0.00516]	[0.00319]	[0.00725]	[0.00443]	[0.00702]	[23.21]
Observations	6,555	6,555	6,555	5,081	4,923	6,406	6,380	6,444

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

The top and lower panels of this table report coefficients from separate regression models. All models are based on the stacked-matched sample. The top panel reports the dynamic treatment effects for the first three years after ISP destination, while the lower panel reports the simple before versus after comparison.

Table 4: Heterogeneity in Treatment Effects Across Schools: By subject and Combined

	1	2	3
	Both	ELA	Math
	ISP Effect	ISP Effect	ISP Effect
Precision-Weighted Average	3.666***	4.314***	3.010***
	[0.863]	[0.992]	[0.895]
Observations	90	45	45
$\hat{\tau}$	5.974	5.631	6.208

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

This reports the precision-weighted pooled average of the ISP designation effect (pre versus post) across all schools on Math passing rates, ELA passing rates, and both combined. The estimated heterogeneity across schools (estimated using DerSimonian and Laird (1986)) is also reported.

Table 5: Testing the Allocative Efficiency Chanel

	1	2	3	4	5	6	7	8	9	10	11	12
	ELA Passing Rate					Math Passing Rate					Both Subjects	
Principal Alignment	3.252***				3.038***	1.771*				1.517*	2.275***	3.009***
	[0.920]				[0.889]	[0.919]				[0.879]	[0.749]	[0.824]
Leadership Score		0.141**					0.0873					
		[0.0592]					[0.0551]					
Residual Passing Rate			2.371***					1.199				
			[0.894]					[0.813]				
Outlier				4.650**	3.965**				4.445**	4.081**	4.004**	4.001**
				[2.006]	[1.806]				[1.783]	[1.765]	[1.792]	[1.803]
Math												-1.331
												[1.013]
Math×Principal Alignment												-1.473
												[0.909]
Math×Outlier												0.00951
												[1.469]
Observations	45	45	43	45	45	45	45	43	45	45	90	90
Pr(Both Subjects Same)												0.244

Standard errors in brackets adjusted for clustering at the school level
*** p<0.01, ** p<0.05, * p<0.1

This reports the precision-weighted meta-regression of the estimated ISP effect for each school against measure of principals alignment (residual passing rate is measured in 2015 while the reported principal quality is measured in 2016) and the extent to which the school had a student population (in 2015) that was somewhat different from the district average. In models that pool across subjects (columns 11 and 12), the probability that the relationships differ y subject is reported.

Table 6: ISP Effect on Measures of Stability and Other

	1	2	3	4	5	6	7
	New Principal	Principal Turnover	School Climate (5 Essentials)	Student Attendance rate	Chronic Truant Rate	Teacher Attendance Rate	Average Class Size
Treated (Year 1)	-0.0257**	-0.313***	0.218	4.63E-05	0.00635	-0.00333	-1.08
	[0.0126]	[0.119]	[0.135]	[0.000998]	[0.0129]	[0.00781]	[0.948]
Treated (Year 2)	-0.0227*	-0.446***	0.354**	0.000511	-0.00696	-0.000328	-0.927
	[0.0121]	[0.132]	[0.146]	[0.00102]	[0.0124]	[0.00912]	[0.728]
Treated (Year 3)	-0.0267*	-0.559***	0.331**	-0.000776	-0.00268	-0.00111	-0.521
	[0.0146]	[0.211]	[0.148]	[0.00105]	[0.0142]	[0.0119]	[0.733]
Treated (All)	-0.0249**	-0.426***	0.297***	-2.41E-05	-0.000979	-0.00163	-0.872
	[0.0118]	[0.129]	[0.110]	[0.000862]	[0.0118]	[0.00810]	[0.561]
Observations	6,599	4,281	4,293	6,435	6,228	4,998	4,899

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

This reports difference-indifference effect of ISP designation on various outcomes using the stacked sample. The top panel reports the dynamic treatment effects for the first three years after ISP destination, while the lower panel reports the simple before versus after comparison.

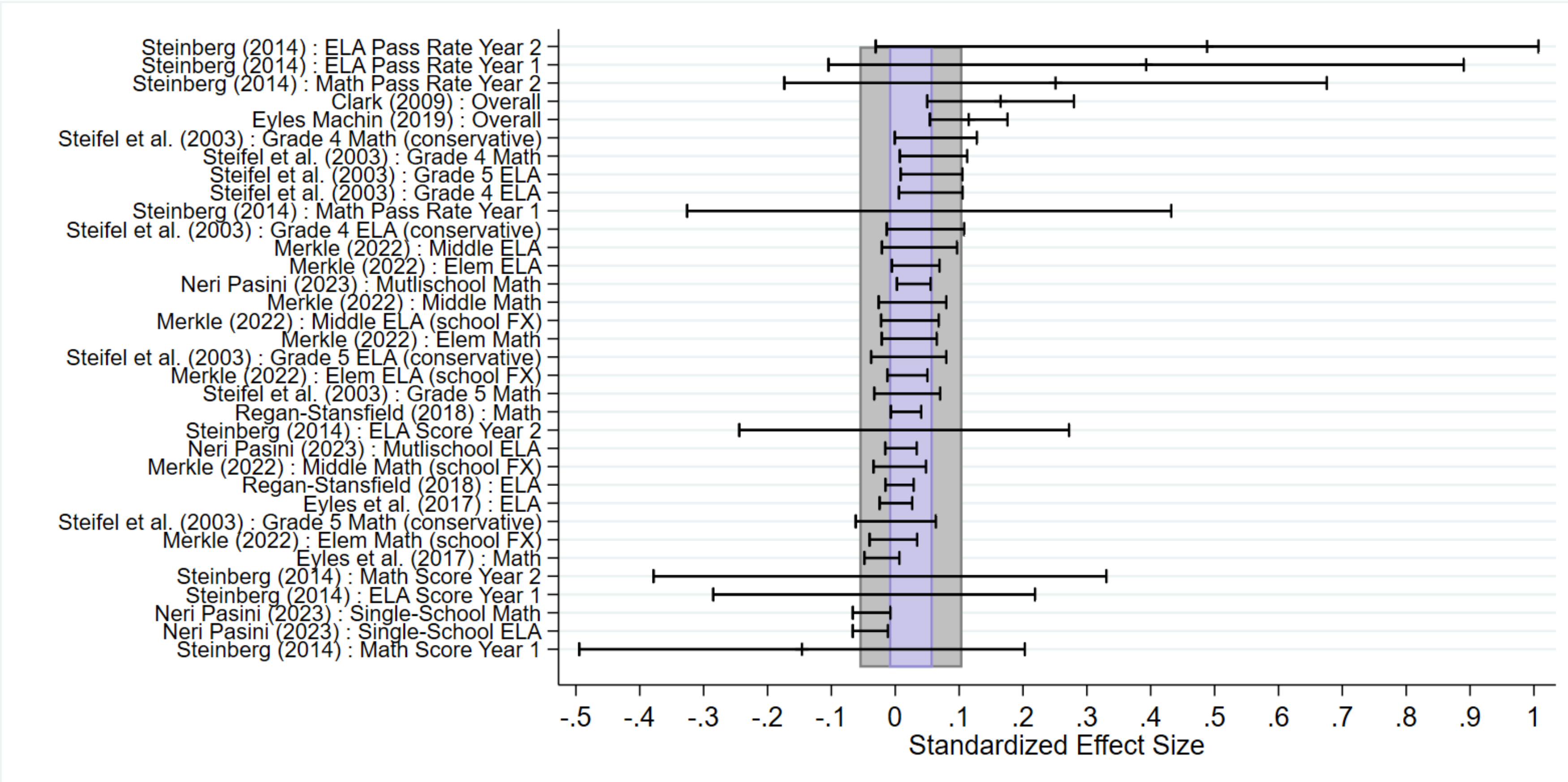
Table 7: Changes in Personnel Spending Categories

	1	2	3	4	5	6	7	8	9	10
	ln Total Personell	ln Teachers	ln Principal	ln Other	ln Special Ed	ln Bilingual	ln Secretary	ln Teaching Assistant	ln Counselor	ln Clerk
Treated	0.00299	-0.00583	0.0900***	0.0075	-0.0819**	0.0604	-0.0149	0.0108	0.0849**	0.00074
	[0.0161]	[0.0175]	[0.0217]	[0.0266]	[0.0377]	[0.0481]	[0.0272]	[0.0830]	[0.0331]	[0.0424]
Treated	-0.00693	-0.0121	0.0825***	-0.015	-0.111***	0.0638	-0.0195	0.0492	0.0782**	-0.011
	[0.0170]	[0.0190]	[0.0228]	[0.0277]	[0.0380]	[0.0506]	[0.0296]	[0.0869]	[0.0353]	[0.0458]
Treated×(outlier Spec. Ed)	0.0944***	0.0651**	-0.0590*	0.200***	0.556***	-0.0915	0.0204	-0.887***	0.0998	0.133
	[0.0227]	[0.0305]	[0.0329]	[0.0374]	[0.0543]	[0.0974]	[0.0495]	[0.0922]	[0.0846]	[0.0920]
Treated×(outlier Bilingual)	0.0612	0.038	0.0666	0.141**	0.132	-0.0102	0.0421	-0.0806	0.0392	0.0691
	[0.0455]	[0.0460]	[0.0538]	[0.0660]	[0.104]	[0.112]	[0.0587]	[0.158]	[0.0890]	[0.100]
Observations	4,298	4,291	4,298	4,298	4,286	3,561	3,537	2,784	4,290	4,298

*** p<0.01, ** p<0.05, * p<0.1
Standard errors in brackets

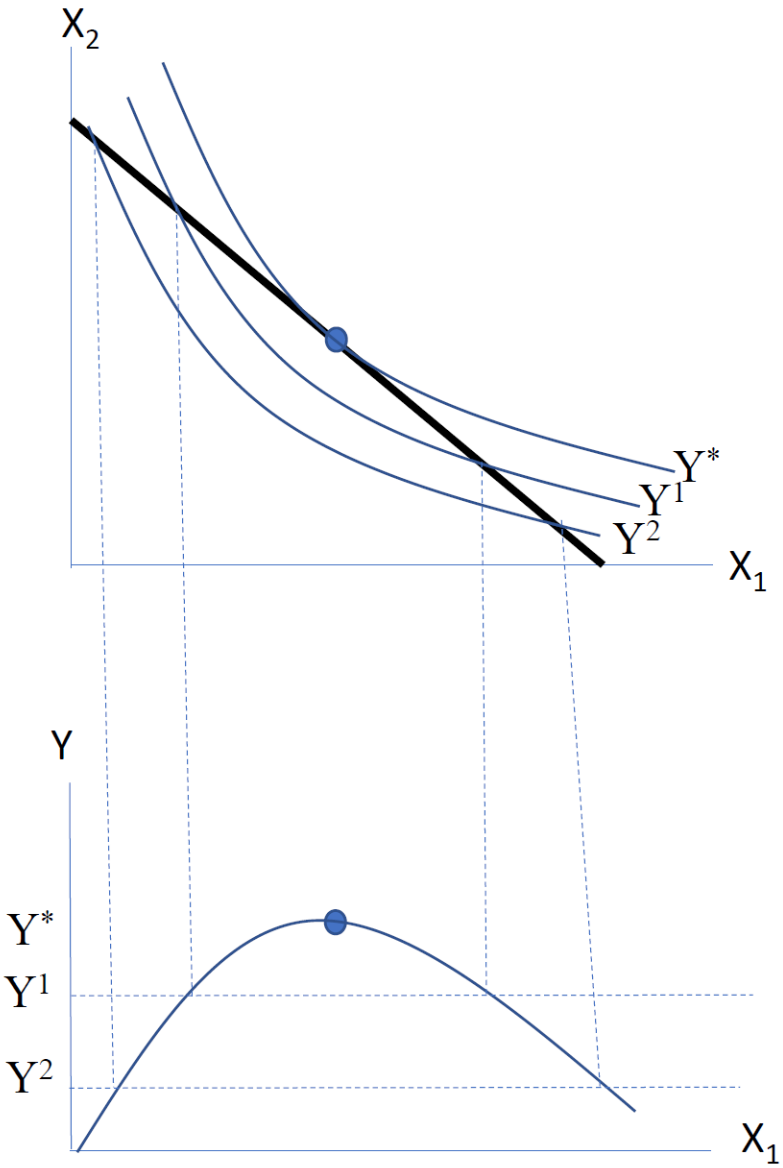
This reports difference-indifference effect of ISP designation on various outcomes using the stacked sample. The top and lower panels report estimates from different regression models. The top panel reports the simple before versus after effects. The lower panel reports the before versus after effects for schools that are outliers in terms of the percentage of special education students and the percentage of bilingual students.

Figure 1. *Summary of Design-Based Studies*



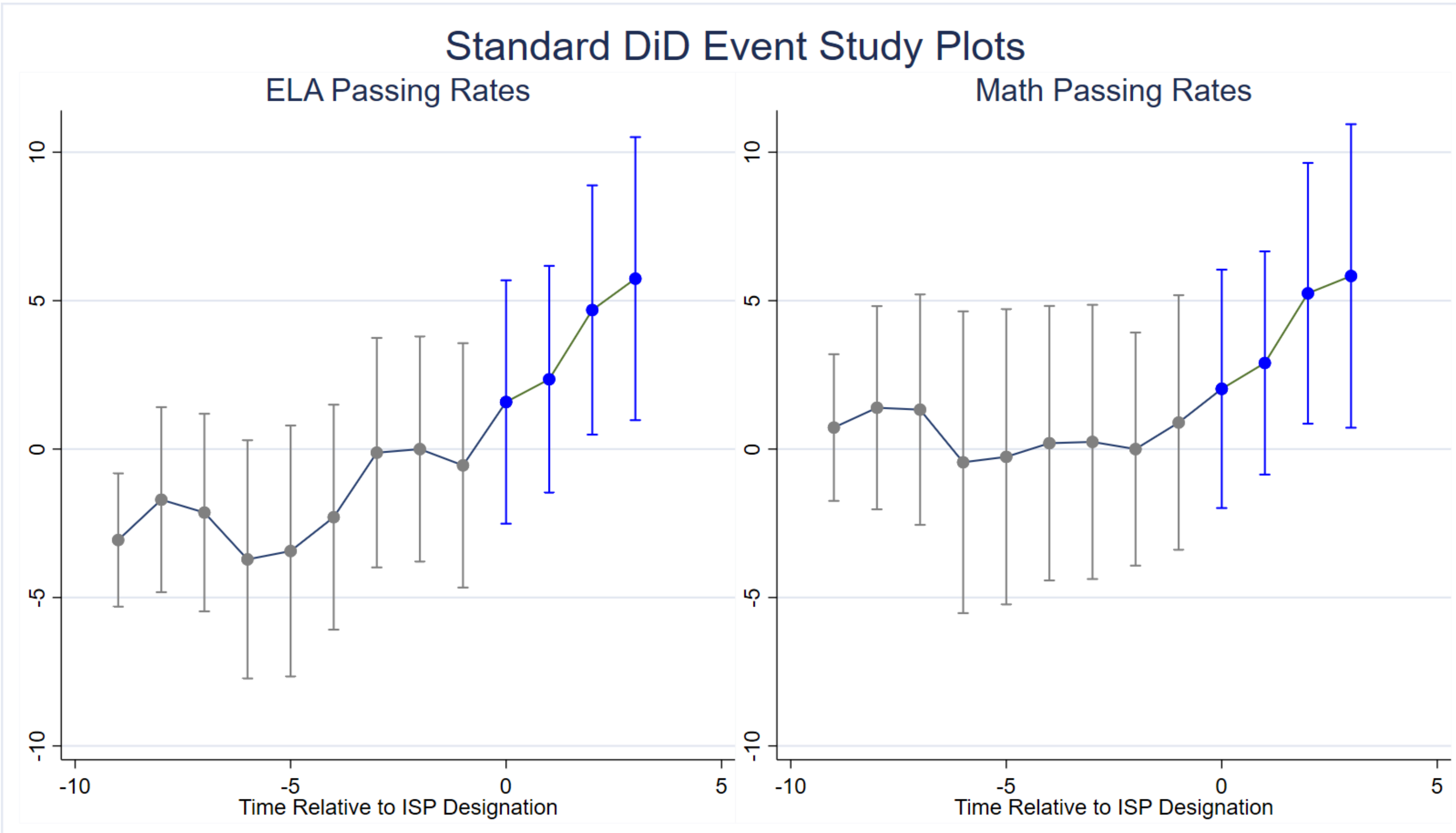
Note: This is a forest plot of all the estimates obtained from the design-based studies discussed in Section II. Each estimates is reported along with its 95 percent confidence interval. The purple area is the 95 percent confidence interval for the pooled average. The grey area is the 95 percent prediction interval indicating the range of likely true treatment effects.

Figure 2. *Sketch of Theory*



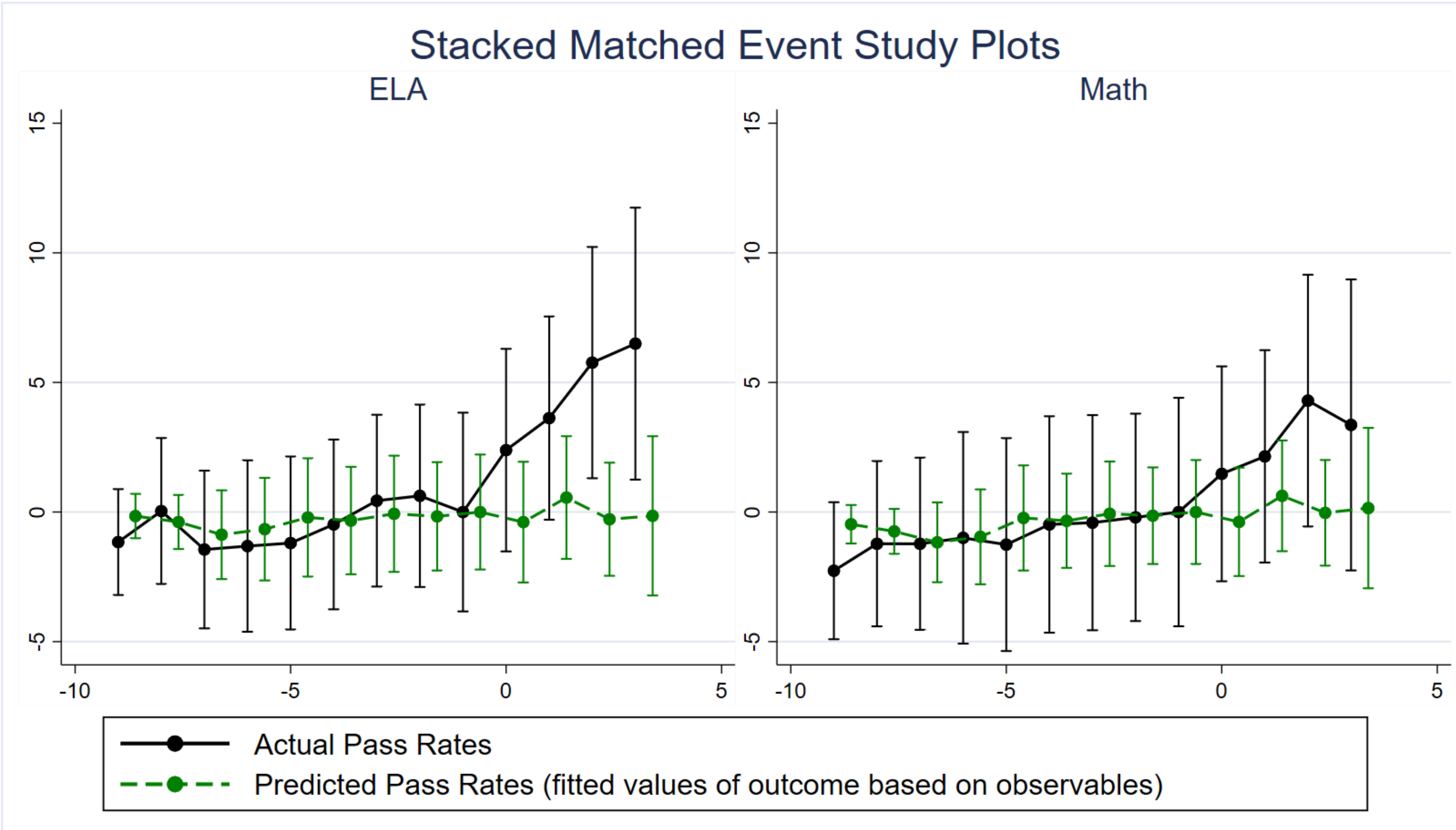
Note: With any smooth twice-differentiable concave production function, the production with respect to any single output (spending all the budget) will be inverse U-shaped, with a maximum at the output maximizing level of input 1.

Figure 3. *Student Achievement Before Versus After ISP Designation: Naive DiD*



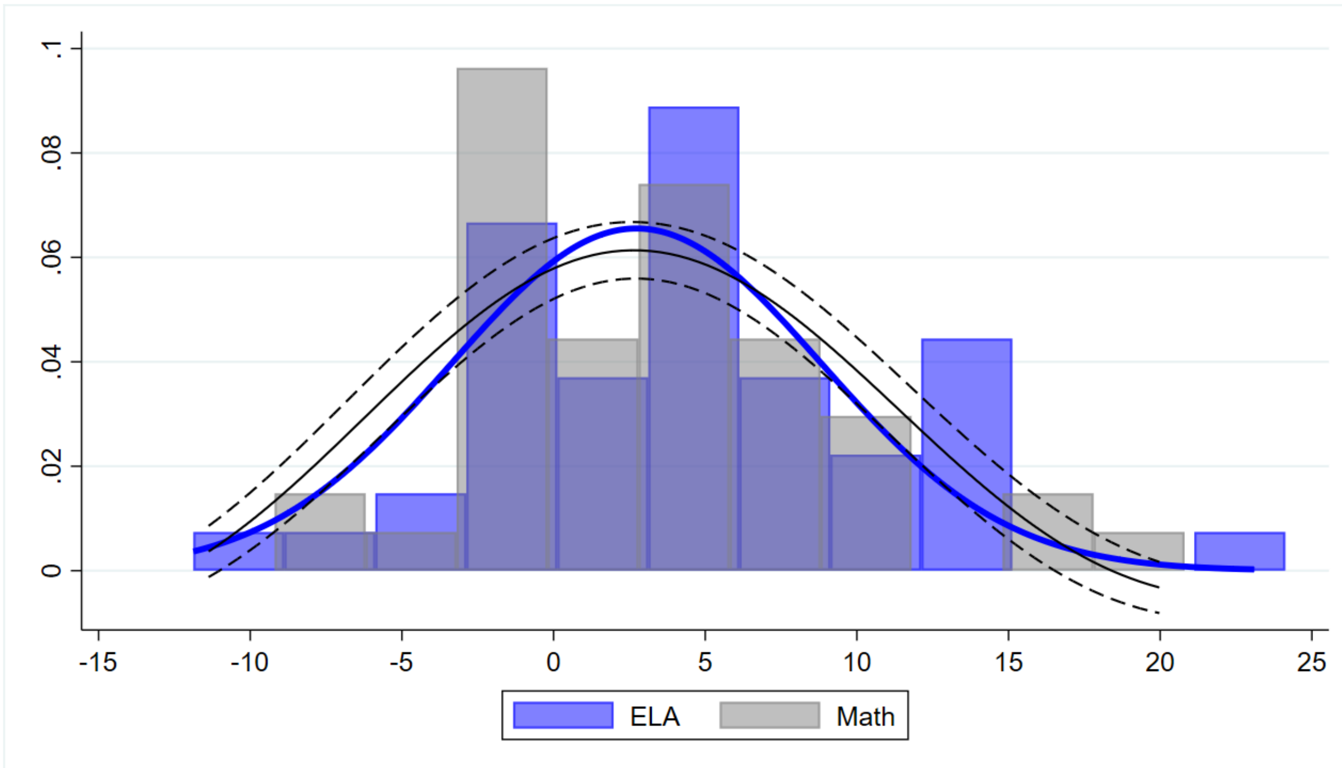
Note: This presents event-study estimates for the standard difference-in-difference models for the full dataset. Each point estimate is the relative year effect along with the associated 95 percent confidence interval. Year 0 is the first year after ISP designation (where there could be an effect).

Figure 4. *Student Achievement Before Versus After ISP Designation: Relative to Comparison Non-ISP Schools using different estimators*



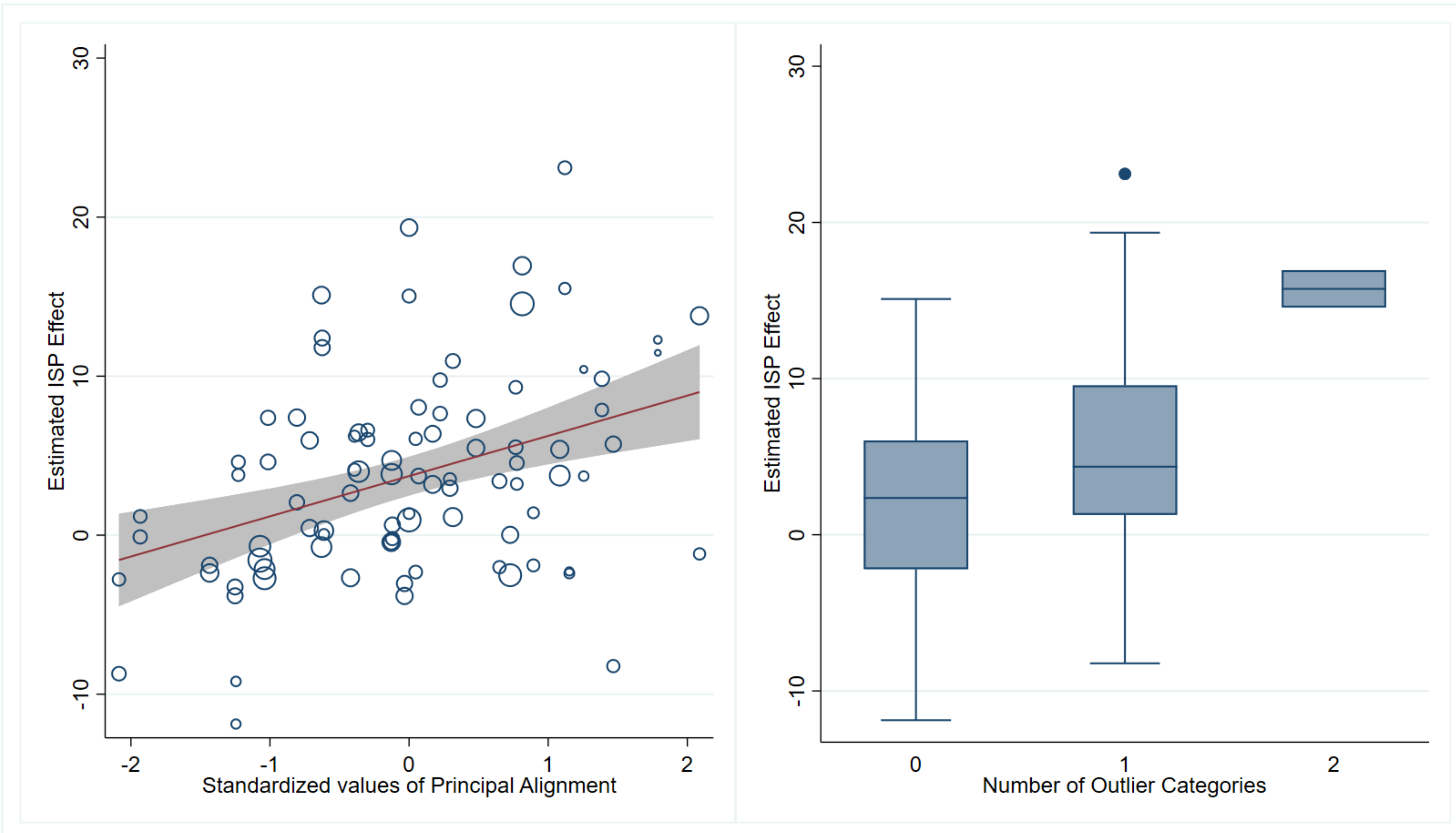
Note: These event-study estimates present the stacked difference-in-difference models using the matched sample following Cengiz et al. (2019) and Deshpande and Li (2019) with matching. The models include no covariates. The event study for actual passing rates is depicted in black, while the event study for predicted rates (based on an outcome-weighted average of all observable predictors of school passing rates) is shown in green. Year 0 is the first year after ISP designation (where there could be an effect).

Figure 5. *Distribution of Individual ISP Effects*



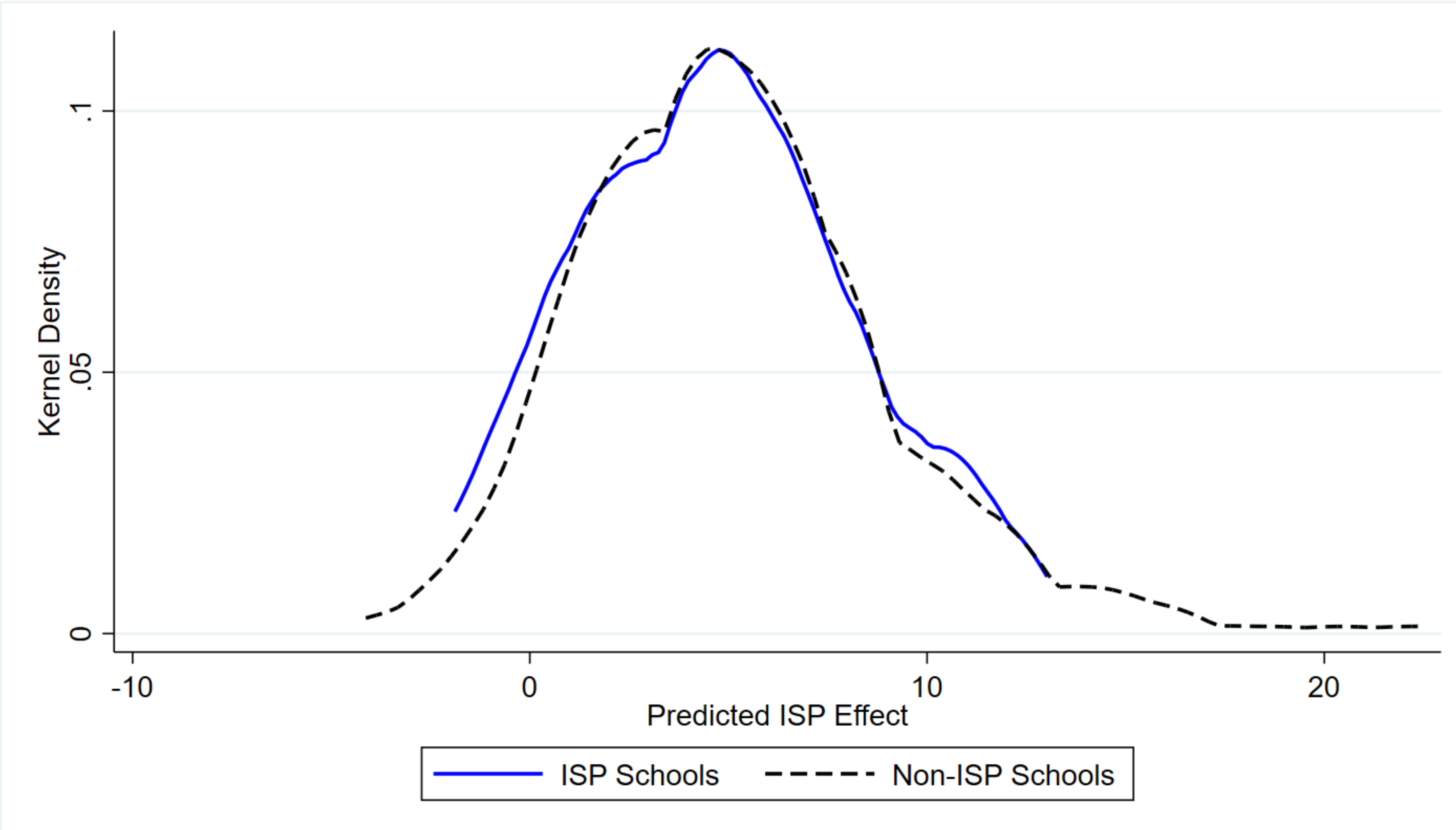
This figure displays a histogram of the estimated impact of the ISP on passing rates in ELA (blue) and Math (grey). The deconvolved density distribution is depicted, along with its 95 percent confidence interval (black). Additionally, a normal distribution (blue) is included for reference, with the same standard deviation as the estimated pooled average.

Figure 6. *Individual ISP Effects by Measures of Principal Quality and Outlier Status*



Note: **Left:** This is a bubble plot displaying the raw ISP effects plotted against the standardized principal alignment measure. More precise estimates (which receive greater weight) are presented as larger bubbles. The plot includes a precision-weighted line of best fit, along with the 95% confidence interval for the line. **Right:** This is a box plot showing the raw ISP estimates for schools categorized as outliers in zero, one, and two categories.

Figure 7. *Predicted ISP Effects for All Schools*



Note: **Left:** This presents kernel density plots for the predicted ISP effect for ISP-designated schools (blue) and also non-ISP-designated schools (black). The predicted ISP effects are estimated from a meta-regression among the treated schools. The predicted effects are fitted values from this regression applied to all schools.

VIII Appendix

Figure A1. *Description of the ISP Program from CPS Website*

6/2/23, 10:55 AM Independent School Principals | Chicago Public Schools

HOME / CAREERS / SCHOOL LEADERSHIP / PRINCIPAL QUALITY

MENU

Independent School Principals

The ISP program is designed for high–performing principals who can ensure continued strong performance with minimal oversight from the district, and who would benefit from additional independence to lead their schools.

The objectives of the program are to:

- Reward high-performing principals with increased autonomy.
- Expand ISP leadership impact through meaningful leadership capacities and innovative collaboration.
- Build streamlined systems and structures that support increased autonomy.

PROGRAM AUTONOMIES | PROGRAM EXPECTATIONS

ISPs are afforded the following autonomies:

- Exemption from network membership and Network Chief oversight.
- Exemption from budget and CIWP approval.
- Increased flexibility with budget and purchasing.
- Professional learning autonomy for ISPs and their staff except for CPS-mandated training.
- Modified principal evaluation within state requirements, including no requirement to submit evidence for SY2017 evaluation, and option for peer evaluation. Note: two formal observations and a final rating are state requirements.

Figure A2. *SQRP Indicators*

Reassignment Rules for Missing Elementary Indicators

Missing Elementary Indicator	**Standard Weight**	**Reassignment Rule***
National School Growth Percentile on the NWEA Reading Assessment	12.5%	School will not receive a rating.
National School Growth Percentile on the NWEA Math Assessment	12.5%	School will not receive a rating.
Priority Group National Growth Percentile on the NWEA Reading Assessment	5%	For each priority group with missing data, weight will be reassigned to National School Growth Percentile on the NWEA Reading Assessment.
Priority Group National Growth Percentile on the NWEA Math Assessment	5%	For each priority group with missing data, weight will be reassigned to National School Growth Percentile on the NWEA Math Assessment.
Percentage of Students Meeting or Exceeding National Average Growth Norms	10%	School will not receive a rating.
National School Attainment Percentile on the NWEA Reading Assessment for Grade 2	2.5%	National School Attainment Percentile on the NWEA Reading Assessment for Grades 3-8
National School Attainment Percentile on the NWEA Math Assessment for Grades 2	2.5%	National School Attainment Percentile on the NWEA Math Assessment for Grades 3-8
National School Attainment Percentile on the NWEA Reading Assessment for Grades 3-8	5%	School will not receive a rating.
National School Attainment Percentile on the NWEA Math Assessment for Grades 3-8	5%	School will not receive a rating.
Percentage of Students Making Sufficient Annual Progress on the ACCESS Assessment^	5%	In the case that any of these indicators are missing, the weight for that indicator will be split evenly between National School Growth Percentile on the NWEA Reading Assessment and National School Growth Percentile on the NWEA Math Assessment.
Average Daily Attendance Rate	20%	
My Voice, My School 5 Essentials Survey	10%	
Data Quality Index Score	5%	

*See Special Case box on page 13 for reassignment of weights for schools serving a highest grade level of Grade 3.

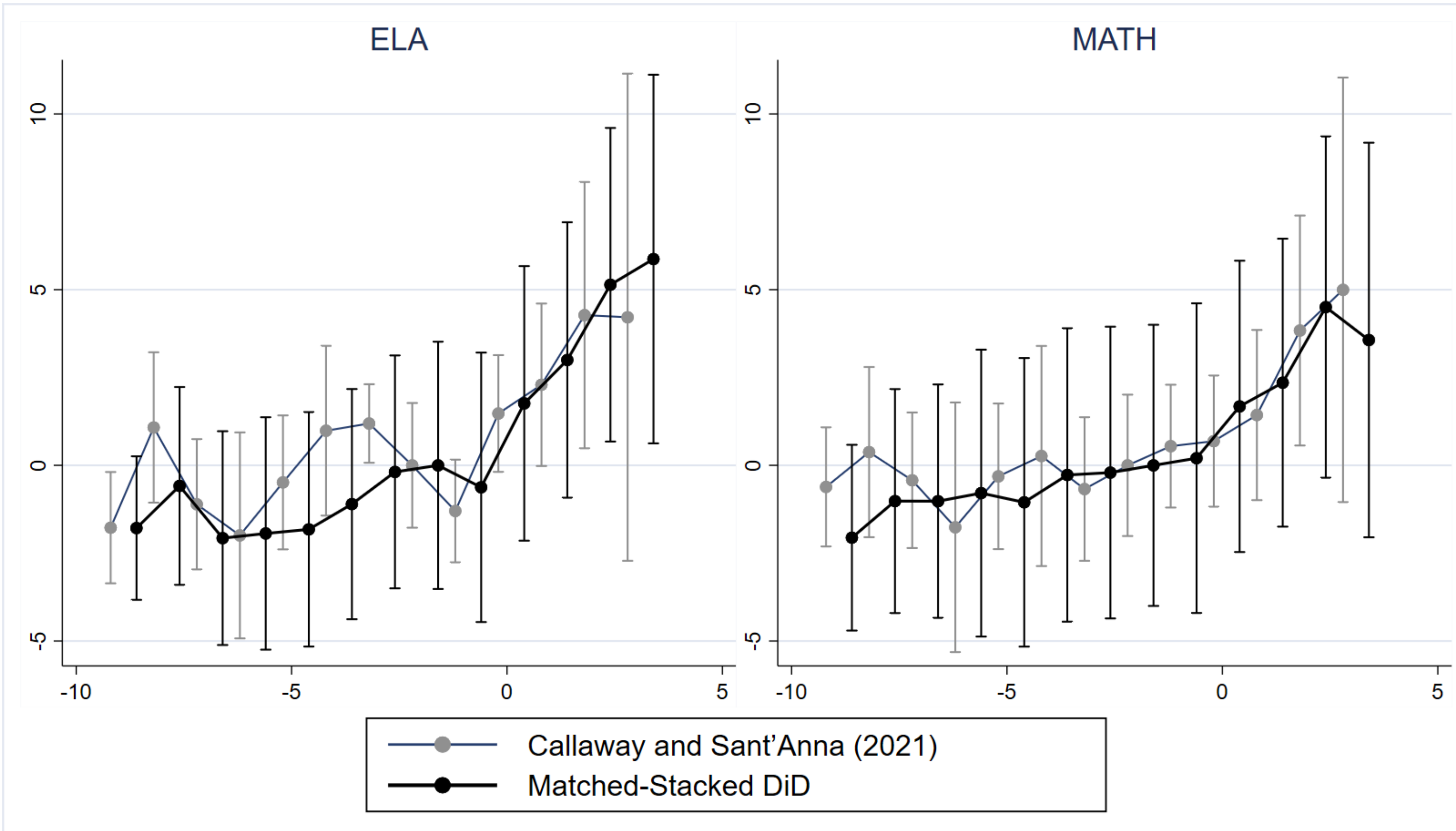
Figure A3. *List of ISPs (Pages 1 and 2)*

Principal	SCHOOL	ES OR HS	SY Joined
Ruth Walsh	ADDAMS	ES	2017
Mira Weber	AGASSIZ	ES	2017
Anna Pavichevich	AMUNDSEN HS	HS	2018
Otis Lee Dunson III	ARMSTRONG G	ES	2020
Takeshi White-James	AVALON PARK	ES	2019
Carmen Navarro	AZUELA	ES	2018
Patricia Brekke	BACK OF THE YARDS HS	HS	2018
Estuardo Mazin	BARRY	ES	2018
Stacy Stewart	BELMONT-CRAGIN	ES	2019
Naomi Nakayama	BUDLONG	ES	2019
Catherine Plocher	BURLEY	ES	2017
Richard Morris	BURROUGHS	ES	2018
Danielle Porch	CALDWELL	ES	2019
Stephen Harden	CAMERON	ES	2019
Clariza Dominicci	CAMRAS	ES	2020
Jeremy Feiwell	CARDENAS	ES	2018
Docilla Pollard	CARNEGIE	ES	2017
Javier Arriola-Lopez	CARSON	ES	2016
Eileen Scanlan	CASSELL	ES	2019
Joseph Peila	CHAPPELL	ES	2019
Barton Dassinger	CHAVEZ	ES	2016
William Hook	CHICAGO AGRICULTURE HS	HS	2017
Natasha Buckner	CLARK ES	ES	2019
Charles Anderson	CLARK HS	HS	2020
Eileen Marie Considine	COLUMBIA EXPLORERS	ES	2020
Wendy Oleksy	COLUMBUS	ES	2018
Gregory Alan Zurawski	COONLEY	ES	2020
Carol Devens-Falk	CORKERY	ES	2019
Carolyn Eggert	DEVRY HS	HS	2018
Kathleen Hagstrom	DISNEY	ES	2016
Beulah McLoyd	DYETT ARTS HS	HS	2018
Nneka Gunn	EBERHART	ES	2019
Serena Peterson	EBINGER	ES	2017
Judith Sauri	EDWARDS	ES	2017
Kurt Jones	FRANKLIN	ES	2018
Michelle Willis	GILLESPIE	ES	2018
Pamela Brandt	GOUDY	ES	2019
Kiltae Kim	GUNSAULUS	ES	2017
Jacqueline Hearns	HEFFERAN	ES	2019
Adam Stich	HITCH	ES	2020
Konstantinos Patsiopoulos	HOLDEN	ES	2019
Charles Smith	INFINITY HS	HS	2019
Paul Powers	JONES HS	HS	2016
Juan Ocon	JUAREZ HS	HS	2016
Suzanne Mazenis-Luzzi	JUNGMAN	ES	2019
Dawn Caetta	KINZIE	ES	2016
Lawanda Bishop	KIPLING	ES	2019
Paul Schissler	LARA	ES	2020
Lauren Albani	LASALLE II	ES	2017
Lisa Epstein	LEE	ES	2017
Angela Sims	LENART	ES	2016
Mark Armendariz	LINCOLN	ES	2019
Michael Boraz	LINCOLN PARK HS	HS	2016

Figure A4. *List of ISPs (Pages 3 and 4)*

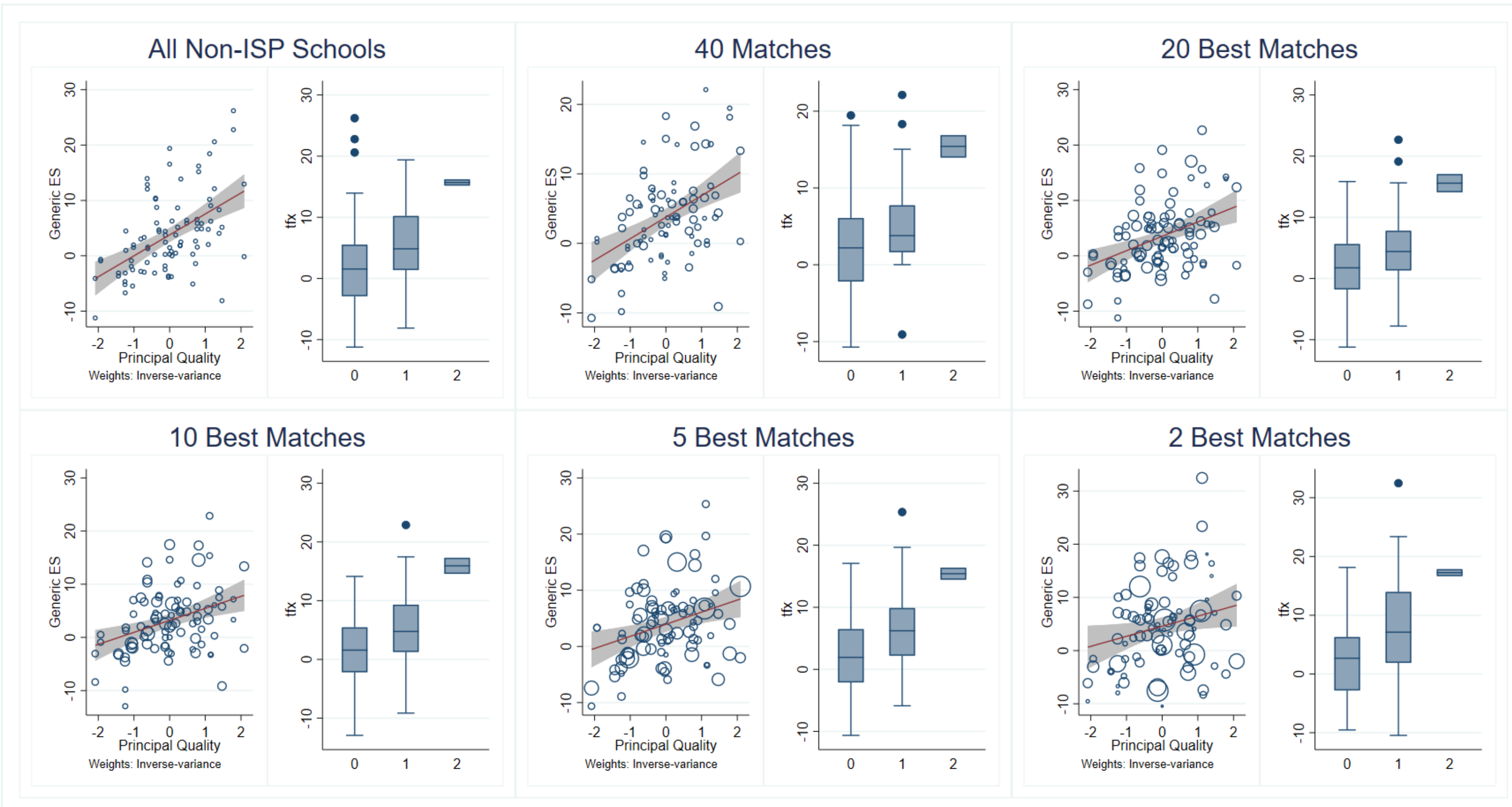
Lillian Lazu	LITTLE VILLAGE	ES	2018
Jay Thompson	LLOYD	ES	2016
July Cyrwus	LORCA	ES	2018
Erin Galfer	MARINE LEADERSHIP AT AMES HS	HS	2018
Jose Juan Torres	MARSH	ES	2020
Joseph Shoffner	MCCLELLAN	ES	2018
Jo Easterling-Hood	MCDOWELL	ES	2017
Karime Asaf	MOOS	ES	2016
Catherine Reidy	MOUNT GREENWOOD	ES	2017
Manuel Adrianzen	NOBEL	ES	2017
Kelly Mest	NORTHSIDE PREP HS	HS	2019
Angelica Herrera-Vest	ORTIZ DE DOMINGUEZ	ES	2020
Jennifer K. Dixon	PALMER	ES	2020
Gerardo Trujillo	PASTEUR	ES	2018
Timothy Devine	PAYTON HS	HS	2016
Brigitte Swenson	PEACE AND EDUCATION HS	HS	2017
Okab Hassan	PECK	ES	2016
Lorainne Zaimi	PEIRCE	ES	2020
Ferdinand Wipachit	PHOENIX MILITARY HS	HS	2019
Rigo Hernandez	PICKARD	ES	2019
Nathan Manaen	RAVENSWOOD	ES	2019
Michael Biela	RICKOVER MILITARY HS	HS	2018
Christine Jabbari	ROGERS	ES	2019
Lourdes Jimenez	SALAZAR	ES	2019
Christine Munns	SAUGANASH	ES	2019
John O'Connell	SHERIDAN	ES	2019
Alice Buzanis	SHERWOOD	ES	2019
Deborah Clark	SKINNER	ES	2016
Jerry Travlos	SMYSER	ES	2017
Tara Shelton	SOUTH LOOP	ES	2016
Joshua Long	SOUTHSIDE HS	HS	2018
Maria McManus	STEM	ES	2019
Olimpia Bahena	TALCOTT	ES	2017
Jacqueline Medina	TALMAN	ES	2017
MaryKay Richardson	THOMAS	ES	2018
Efren Toledo	THORP O	ES	2018
Gerardo Arriaga	TONTI	ES	2017
Sabrina Boone Jackson	TURNER-DREW	ES	2020
Renee Mackin	VON LINNE	ES	2018
Ekaterini Panagakis	WACKER	ES	2018
Rashid Shabbazz	WADSWORTH	ES	2019
Karen Anderson	WARD J	ES	2018
Antigoni Lambrinides	WEST RIDGE	ES	2019
Joyce Kenner	YOUNG HS	HS	2016
Ruth Garcia	ZAPATA	ES	2016

Figure A5. *Alternative Estimator*



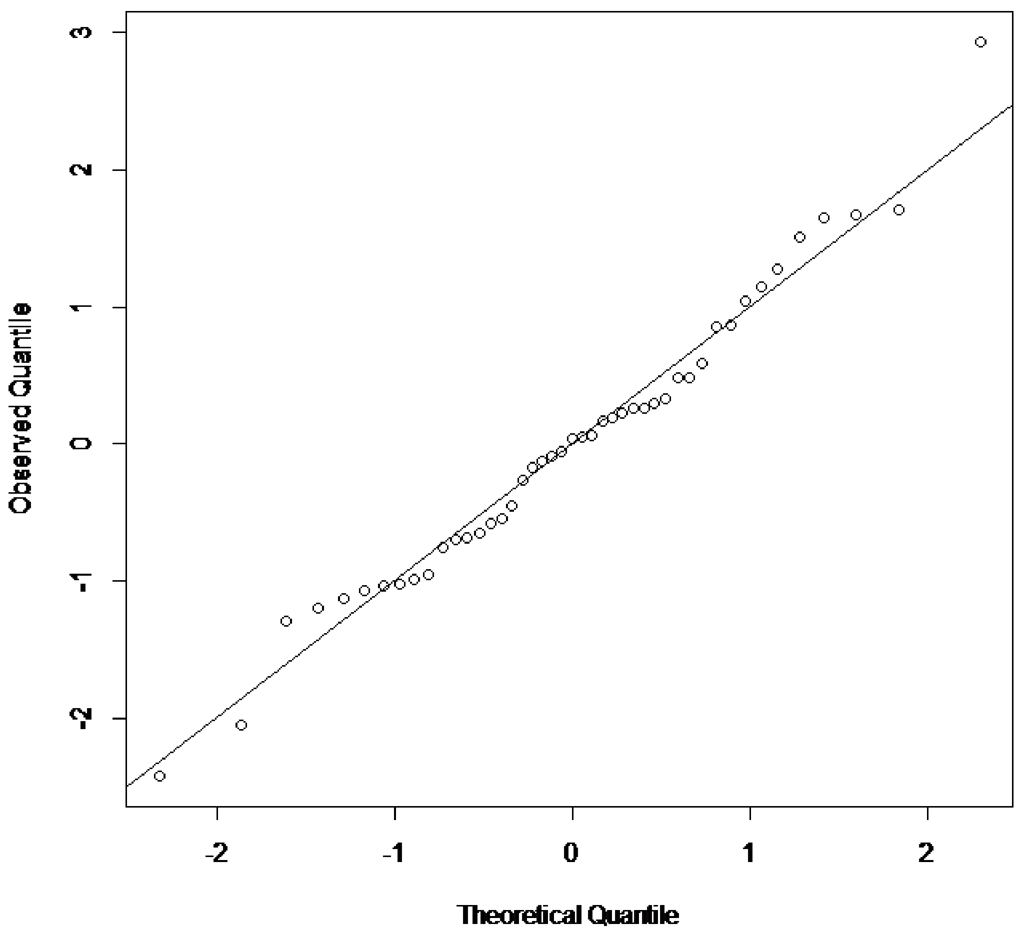
The grey event-study estimates depict the difference-in-difference models using the methodology of Callaway and Sant'Anna (2021). The black event-study estimates show the preferred stacked difference-in-difference models using the matched sample, following Cengiz et al. (2019) and Deshpande and Li (2019) with matching. Both models are estimated without covariates. Year 0 is the first year after ISP designation (where there could be an effect).

Figure A6. *Pattern Of Heterogeneity using More or Fewer Matches*



Note: Each panel presents a bubble plot on the left and a box plot on the right, as detailed below, using a different number of matches for the matched individual-level school ISP estimates. **Left:** a bubble plot displaying the raw ISP effects plotted against the standardized principal alignment measure. More precise estimates (which receive greater weight) are presented as larger bubbles. The plot includes a precision-weighted line of best fit, along with the 95% confidence interval for the line. **Right:** This is a box plot showing the raw ISP estimates for schools categorized as outliers in zero, one, and two categories.

Figure A7. *Normality of ISP Effects*



Following Wang and Lee (2020), I report The p-value associated with the Shapiro-Wilk test of normality on the appropriately shrunken estimates. This yields a value of 0.71 – indicative of the distribution of true effects being approximately normal.

Table A1: Estimate ISP Effect Using a Different Number of Matches

	All Non-Treated Elementary Schools						40 Most Similar Non-Treated Elementary Schools					
	Passed ELA	Passed Math	ELA Percentile	Math Percentile	Predicted Passed Math	Predicted Passed ELA	Passed ELA	Passed Math	ELA Percentile	Math Percentile	Predicted Passed Math	Predicted Passed ELA
Treated (All)	4.880***	3.700***	-0.938	-0.283	0.305	0.770*	4.262***	3.151***	0.745	2.467*	0.329	0.225
	[1.235]	[1.342]	[1.355]	[1.640]	[0.523]	[0.418]	[1.087]	[1.036]	[1.027]	[1.491]	[0.413]	[0.403]
Observations	198,419	198,466	118,927	118,927	225,694	225,694	16,056	16,056	10,549	10,549	16,823	16,823
	20 Most Similar Non-Treated Elementary Schools						15 Most Similar Non-Treated Elementary Schools					
Treated (All)	4.060***	2.919***	1.069	2.657*	0.223	0.127	4.207***	3.113***	1.407	3.136**	0.366	0.266
	[1.069]	[1.033]	[0.986]	[1.466]	[0.412]	[0.406]	[1.072]	[1.034]	[0.970]	[1.448]	[0.416]	[0.399]
Observations	8,354	8,351	5,510	5,510	8,657	8,657	6,402	6,402	4,220	4,220	6,599	6,599
	10 Most Similar Non-Treated Elementary Schools						5 Most Similar Non-Treated Elementary Schools					
Treated (All)	3.793***	2.528**	1.212	2.653*	0.243	0.182	4.689***	3.315***	1.960*	3.518**	0.162	0.169
	[1.078]	[1.020]	[0.968]	[1.420]	[0.434]	[0.403]	[1.142]	[1.101]	[1.030]	[1.474]	[0.497]	[0.469]
Observations	4,393	4,393	2,883	2,883	4,535	4,535	2,406	2,406	1,576	1,576	2,477	2,477
	2 Most Similar Non-Treated Elementary Schools						Single Most Similar Non-Treated Elementary School					
Treated (All)	5.600***	3.588***	2.316**	5.040***	-0.085	-0.057	6.475***	5.083***	3.656***	6.264***	0.00729	-0.0391
	[1.214]	[1.200]	[1.085]	[1.659]	[0.512]	[0.478]	[1.232]	[1.439]	[1.228]	[2.032]	[0.584]	[0.602]
Observations	1,189	1,189	760	760	1,242	1,242	796	796	504	504	828	828

Standard errors in brackets
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Each reported coefficient is the before versus after ISP-designation effect from a separate regression models. All models are based on stacked-matched sample. The number of non-ISP schools used to create the comparison group (ranging from all non-treated schools to the single best match) is indicated in the table.

Table A2: Estimated ISP Effect on Outcomes and Covariates: Single Best Match Sample

	Passed ELA	Passed Math	ELA Percentile	Math Percentile	Free-Lunch %	Enrollment	% Black	% White	% Hispanic	% Spec. Ed.	% Bilingual	Mobility Rate	Predicted Passed Math	Predicted Passed ELA
Treated (Year 1)	**4.562*** **	**3.496*** **	**2.700* **	**5.208*** **	-0.000366	4.616	-0.00218	0.00999	-0.00316	0.000881	-0.00622	0.00367	-0.511	-0.465
	[1.087]	**[1.175]**	**[1.377]**	**[1.702]**	[0.00909]	[22.80]	[0.00649]	[0.00658]	[0.00796]	[0.00296]	[0.0103]	[0.00473]	[0.602]	[0.594]
Treated (Year 2)	**6.293*** **	**4.596*** **	**4.081** **	**6.803*** **	-0.0164	4.593	-0.00238	0.00847	-0.000668	-0.00404	-0.0103	-0.00338	0.265	0.241
	[1.275]	**[1.425]**	**[1.642]**	**[2.396]**	[0.0130]	[24.38]	[0.00733]	[0.00721]	[0.00912]	[0.00375]	[0.00989]	[0.00584]	[0.671]	[0.761]
Treated (Year 3)	**9.105*** **	**7.675*** **	**4.409** **	**6.996** **	-0.0072	-14.09	-0.00995	0.00345	0.00372	-0.00652	0.00575	-0.00548	0.337	0.145
	[2.136]	**[2.324]**	**[1.970]**	**[3.009]**	[0.00880]	[38.07]	[0.00644]	[0.00927]	[0.0110]	[0.00490]	[0.00979]	[0.00840]	[0.762]	[0.724]
Treated (All)	**6.475*** **	**5.083*** **	**3.656*** **	**6.264*** **	-0.00813	-0.824	-0.00447	0.00758	-0.000316	-0.00294	-0.00454	-0.0015	0.00729	-0.0391
	[1.232]	**[1.439]**	**[1.228]**	**[2.032]**	[0.00846]	[25.78]	[0.00622]	[0.00692]	[0.00844]	[0.00329]	[0.00827]	[0.00518]	[0.584]	[0.602]
Observations	796	796	504	504	774	794	826	826	826	640	618	780	828	828

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

The top and lower panels of this table report coefficients from separate regression models. All models are based on the stacked-matched sample using only the best single untreated match for each ISP-designated school. The top panel reports the dynamic treatment effects for the first three years after ISP designation, while the lower panel reports the simple before versus after comparison.