Governance Matters: School Autonomy, Leadership, Heterogeneity, and Student Outcomes *

C. Kirabo Jackson
Northwestern University

July 16, 2025

*Jackson: kirabo-jackson@northwestern.edu. **Mail:** Northwestern University, School of Education and Social Policy, 2120 Campus Drive, Room 204, Annenberg Hall, Evanston, IL 60208. **Phone:**: (847) 467-1803. 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.

Abstract

This paper studies a Chicago policy that expanded school-level autonomy by granting principals greater control over budgeting and operations. A meta-analysis of similar reforms shows modest average effects but sizable heterogeneity across contexts. Drawing on theories from public finance, contract design, and psychology, I argue the returns to autonomy depend on principal quality and alignment between school and district priorities. Event-study estimates show increased autonomy improved math and English passing rates by 0.1σ, comparable to more resource-intensive interventions but achieved at minimal cost. Deconvolution analysis reveals substantial heterogeneity, with both negative and positive effects. Design-based evidence supports the theoretical predictions: high-performing principals benefit most, reallocating resources effectively (e.g., reducing class sizes), and schools with atypical student populations benefit more and tailor services to local needs. These results highlight that local capacity, aligned incentives, and heterogeneity are key to the success of decentralization reforms.

I Introduction

It has long been theorized that assigning policy responsibility to the lowest feasible level of government promotes efficiency, as lower levels are better equipped to understand and respond to specific needs (Oates 1999, Wallis and Oates 1988), with the greatest benefits in settings with heterogeneous needs (Acemoglu et al., 2007; Neri et al., 2022). However, granting more authority to agents can also lead to poor decisions and decisions misaligned with the principal's objectives (Aghion and Tirole, 1997; Grossman and Hart, 1986; Bishop et al., 2004). As such, the benefits of "*street-level*" agent autonomy may depend on which forces dominate in a given context.

Although this is well-understood theoretically, empirical evidence using design-based approaches is lacking. To address this gap, this study examines a 2016 Chicago policy that shifted decision-making authority from the district (the principal) to certain school principals (the agents) by increasing their control over budgeting and operations. This study investigates the average effect of this policy and provides design-based evidence on when decentralization is more or less effective.

In public education, there is broad consensus that increasing school resources improves student outcomes on average (Jackson et al. (2023)), but effects are heterogeneous (Jackson and Mackevicius (2024)) and vary depending on how money is spent (Biasi et al. (2024)). Furthermore, even holding costs constant, differences in implementing the same basic program can lead to significant variation in effectiveness (Accelerate (2024)). This underscores the critical role of localized knowledge and high-quality decision making in ensuring investments translate into meaningful improvements.

The existing evidence on the average effect of school autonomy on student achievement is inconclusive, and suggests that context plays a crucial role. In the U.S., charter and pilot schools, which have more autonomy than traditional public schools, show no average improvement (Cohodes and Parham, 2021; Abdulkadiroğlu et al., 2011). However, these schools differ in many ways beyond autonomy, limiting their relevance for assessing the **policy** effect of increasing school-level autonomy *per se*. Design-based U.S. studies of policies expanding principal autonomy find limited effects. Stiefel et al. (2003) and Merkle (2022) report generally positive effects of New York City reforms increasing school autonomy, but find small and statistically insignificant estimates in conservative models, while Steinberg (2014) finds imprecise estimates for a similar 2005 Chicago policy. In the U.K., Clark (2009) and Eyles and Machin (2019) find substantial gains when high schools gain autonomy within broader accountability reforms, while Eyles et al. (2017) and Regan-Stansfield (2018) find little impact in primary schools and (Neri and Pasini, 2023) find potential negative effects. A formal meta-analysis (Section II) examines these studies, quantifies this heterogeneity, and highlights the importance of context.

To make sense of heterogeneous treatment effects, I present a framework for identifying schools benefiting most and least from autonomy. Even though 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 *correlated* with better outcomes in states with strong accountability (Loeb and Strunk, 2007) and nations with better institutions and 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 84 public schools where principals transitioned to ISPs between 2016 and 2019.[1] 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 comparison schools based on pre-reform characteristics, stacking the

[1]In Chicago, most elementary schools enroll students from pre-kindergarten through 8th grade. I consider any school that enrolls students in grades 3 through 8 to be an elementary school.

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 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 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. The results show that simply changing the governance structure (i.e., changing *who* gets to make decisions) while holding the budget effectively fixed can have dramatic effects on student achievement.

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 may exhibit true effects differing by six percentage points on average (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 support the idea that effects can be either positive or negative, as theorized, and align with meta-analytic results from existing design-based studies – emphasizing the importance of better understanding this variability.

To *explain* this effect heterogeneity and explore the role of leadership quality, 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 effects are *negative* for less capable principals. Improved autonomy results in approximately 1.8 percentage points higher passing rates (about 0.046σ) for principals one standard deviation above the mean, highlighting the importance of principal quality.

To test whether benefits are larger when schools have heterogeneous needs (Acemoglu et al., 2007), I classified schools as specialized if they had a proportion of bilingual, special education, white, or Hispanic students above the 90thth percentile. Schools that were outliers in any category had larger treatment effects, and 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, 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.

Further analysis of personnel spending data indicates that ISPs reduced class sizes by about 1.4 students without increasing their personnel budgets. This was likely achieved by using greater autonomy to reallocate staff or shift funding in ways that lowered class sizes without additional financial outlays—for instance, by assigning a special education teacher to assist in a core classroom. Although this reallocation appears to be a meaningful channel through which ISPs improved outcomes, it can account for only about a quarter of the total ISP effect on test scores.

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. Importantly, the implied achievement effects attributable to reduced principal turnover are significantly smaller than the ISP effect. This indicates that while reduced principal turnover is important, it is not the primary channel through which ISP generates test score gains. Consistent with providing stability, school climate improves 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 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 and many strong principals are not ISPs. 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 important allocative efficiencies, creating an opportunity to improve outcomes with minimal financial cost.

Because this study is not based on a randomized controlled experiment, one might worry that improved outcomes reflect changes in student composition rather than the effect of the ISP itself. However, tests of observable student characteristics suggest this is unlikely. Moreover, comparing test scores for continuously enrolled students to those transferring from other sites provides no evidence of selective migration based on achievement after ISP designation. Another concern is that ISP principals might have achieved strong outcomes regardless of the policy. While focusing on within-school changes over time rules this out, one may worry that principals were selected based on expected *future* performance. While this cannot be entirely ruled out, (a) future improvement was ***not*** a selection criterion, and (b) research shows policymakers have difficulty accurately predicting teacher and school effectiveness (Jacob and Lefgren (2008); Donaldson et al. (2021); Grissom et al. (2018)). Moreover, randomization tests show that policymakers would have needed virtually perfect foresight in identifying schools with the highest future test score gains—a highly implausible scenario—to spuriously generate benefits of the magnitude observed. Additionally, many non-ISP schools would have had large ISP effects, further evidence that principals were chosen based on past success and *not* to maximize treatment effects. However, the most compelling evidence against this bias is that there are *negative* effects for principals with weak track records, and stronger effects for schools with specific needs. This pattern cannot be explained by selection on future success, suggesting likely true causal impacts that may offer valuable policy insights.

These results contribute to several literatures. First, they provide clear evidence of school autonomy effects in the United States while emphasizing the importance of context and heterogeneity. Second, they advance the principal quality literature by demonstrating the predictive power of value-added measures for treatment effects, despite their limitations (Chiang et al. (2016); Grissom et al. (2012)), and validate established links between survey-based leadership assessments and student outcomes (Bloom et al. (2015); Liu et al. (2014)). More broadly, the findings underscore the importance of effective leadership (Bertrand and Schoar (2003); Jones and Olken (2005); Frick et al. (2007)). Finally, while theoretical and empirical work on decentralized decision-making exists (Acemoglu et al. (2007); Neri et al. (2022); Colombo and Delmastro (2004)), this study provides direct design-based evidence on when such decentralization improves outcomes.

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

II A Meta-Analysis of Existing Design-Based Studies

To inform the analysis, I draw insights from existing work on policies that expand school-level autonomy and their impact on student achievement. I examine studies that estimate **policy effects** using design-based identification strategies, focusing on policies that increased or decreased autonomy within existing schools. This **policy** focus excludes studies that compare schools with differing levels of autonomy in the cross-section, as other key differences across schools may confound results, and this is not policy variation per se.[2] To isolate autonomy effects, I exclude studies of expanded autonomy that include other treatments such as increased spending (e.g. Tuchman et al. (2022)) or implementing particular school models (e.g. Cohodes et al. (2021)).

To obtain a comprehensive set of papers, I follow Jackson and Mackevicius (2024). I identified eight studies as of December 31, 2023, that met the inclusion criteria. These studies became seed papers, which I input into connectedpapers.com to find related papers. Connected Papers uses the Semantic Scholar Paper Corpus (indexing over 200 million academic papers from publisher partnerships, data providers, and web crawls, including unpublished work) to identify papers based on similarity, determined through co-citations and bibliographic coupling rather than direct citations. I evaluated each related paper against the inclusion criteria and made it a seed paper if included. I repeated this process until no new papers met the criteria. This approach identified publicly available design-based studies on the effects of increasing school-based autonomy. While many studies examine school-based autonomy, this method yielded only eight studies that specifically examined **policy** effects of expanding autonomy in pre-existing schools.

I include multiple estimates from the same paper when final results are presented for different subjects, tests, or specifications – yielding 28 estimates from eight papers.[3] Figure 1 presents the forest plot (individual point estimates and 95 percent confidence intervals based on reported standard errors). All effects are reported as the effect of increased autonomy on student-level standard deviations. Papers reporting effects on passing rates are converted to standardized effects using the inverse normal transformation as in Ho (2009).[4]

The plot shows a wide range of estimates across studies (from -0.15 in Steinberg and Cox (2016) to 0.17 in Clark (2009)). However, the spread of study estimates overstates the degree of true heterogeneity because studies may vary due to true heterogeneity or sampling variability. Indeed, the largest estimates are relatively imprecise, and there is considerable overlap in confidence intervals across most estimates. While most estimates are positive (indicative of a positive pooled average), assessing the degree of *true* heterogeneity requires a more rigorous approach.

[2] Such studies (e.g., comparisons between charter or magnet schools and traditional public schools) capture school-specific effects rather than policy effects *per se* and may reflect other differences across schools.

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

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

Measuring Heterogeneity

While it is common to describe disparate study findings as "mixed," this is imprecise and subjective, and the conclusion can only be justified if differences across studies cannot be explained by sampling variability alone. Although observational studies have documented patterns consistent with heterogeneity across nations (Fuchs and Wößmann 2007) and states (Stiefel et al. 2003), heterogeneity has not been documented among design-based studies.

Estimates from non-experimental studies can vary due to differences in local average treatment effects (LATEs), populations, treatment definitions, unmeasured treatment factors, or random sampling variability. For example, the current papers examine different grade levels and policy types (Eyles and Machin (2019) focus on low-performing secondary schools, while Eyles et al. (2017) target high-performing primary schools), autonomy recipients (Stiefel et al. (2003) examines policies increasing principal autonomy, while Clark (2009) examines policies increasing both principal and parent autonomy), and performance measures (some papers use test score value-added, others examine passing rates, and others consider test score levels). All of these factors contribute to heterogeneity. To define terms, I define true heterogeneity as all variability across studies unexplained by sampling.[5] Understanding this broad conception of heterogeneity sheds light on the plausible range of true effects for policies aimed at increasing school autonomy.

To formally assess whether the observed estimates lie above zero on average, and to measure true heterogeneity, I take the 28 estimates (and associated standard errors) from the eight design-based papers and estimate a Bayesian hierarchical model – estimating both the pooled average, Θ, and the spread of true effects across contexts, τ^2. The basic intuition is that observed variability of estimates reflects both heterogeneity and sampling variability. However, because measures of sampling variability are observed (the standard error of the estimated treatment effects), one can infer the extent of true heterogeneity.[6] The dependence across estimates from the same paper is

[5]The small sample precludes my exploring differences by these individual dimensions.

[6]I briefly outline the Bayesian hierarchical model, or Bayesian meta-analysis. See Appendix Section B for further details. There is some grand mean (Θ) representing the pooled effect of all studies. The true effect for any study (θ_j) is a random draw from a normal distribution (justified by the central limit theorem), centered on Θ, with variance τ^2 – which represents true treatment heterogeneity. Using Bayes rule, if one defines the prior distributions for hyperparameters τ and Θ, one can estimate the posterior distributions of all these parameters. Moments (such as the mean) of the posterior distributions of τ,Θ, and θ provide information about the values of these parameters. Moreover, the spread of the posterior distributions sheds light on the uncertainty around the values of these parameters.

To this aim, I assume that the true effect is a random draw from a normal distribution, and that the heterogeneity parameter τ^2 follows an inverse Gamma distribution as in (1) and (2). The inverse Gamma distribution is commonly used to model variance parameters and avoids the non-negative estimates one can obtain from other approaches.

$$\Theta \sim \mathcal{N}(.) \tag{1}$$

$$\tau^2 \sim InvGamma(.) \tag{2}$$

I estimate this model with starting values such that $\tau^2 \sim InvGamma(0.0001, 0.0001)$ and that $\Theta \sim \mathcal{N}(0, 100)$.

accounted for using Bayesian Model Averaging. Note that all of the main conclusions are similar when using frequentist meta-analytic approaches as in Jackson and Mackevicius (2024).

The estimated pooled average effect across studies (Θ) is 0.022σ, with a 95% credibility interval ranging from -0.0081 to 0.057. This interval, represented by the blue area in Figure VII, indicates that while the model estimates a positive average effect of policies aimed at increasing autonomy within schools—raising test scores by 0.022σ—it does not reject the null hypothesis that the pooled average effect is zero.

Importantly, this does *not* imply that all study effects are zero or the same. Indeed, at least one study has a credibility interval entirely below zero (true negative effect), while others have intervals entirely above zero (true positive effects). Consistent with this variation, the model estimates true heterogeneity τ at 0.047σ, with a 95% credibility interval between 0.0001 and 0.0855. This suggests non-uniform effects across contexts— the model rejects the null hypothesis that all effects are identical and estimates that true effects would differ by approximately 0.047σ between two randomly chosen settings. To show this more clearly, the 95 percent prediction interval for what one would expect for a future study (which accounts for both sampling errors and true heterogeneity) is depicted by the grey region. This is estimated based on draws from the simulated posterior distribution. To put this range into perspective, while school-level autonomy is generally associated with higher test scores, the posterior distribution indicates true negative effects about one-fifth of the time and meaningful positive effects (above 0.04σ) roughly ten percent of the time.[7]

In sum, while existing research suggests that greater school autonomy improves student outcomes in many settings, there is no uniform "autonomy effect"—the effect may be positive in some contexts and negative in others. If research is to inform policy, understanding how effects vary across different settings is crucial. This study aims to contribute to that understanding.

III The Policy and Theoretical Framework

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

Chicago Public Schools (CPS) makes operational decisions such as hiring and curriculum at the district or state level, like most U.S. public school districts. CPS organizes its district-run schools into 18 networks (based on location and type) that provide administrative support, strategic direction, and leadership development. Network chiefs serve as intermediaries between district administration and individual principals.

Chicago struggled with high principal turnover, as Chicago Public Education Fund findings

[7] It is important to note that this measure of heterogeneity reflects variation across study contexts. Heterogeneity across schools or districts within a given study context could be substantially larger, highlighting the likely significance of heterogeneity.

highlighted (CPEF, 2017). Research also revealed that increasing autonomy and reducing oversight could significantly enhance job satisfaction among CPS principals (CPEF, 2015); 7 in 10 Chicago principals identified reducing compliance as one of the top three ways to improve job satisfaction, 65% wanted more tailored professional development, and 40-50% felt unable to organize school resources to advance school goals and priorities. In focus group discussions, top performing principals (measured using the same survey data in this study) repeatedly expressed a desire for more flexibility in choosing curriculum and leading teacher professional development.

Informed by these findings, the Independent School Principal (ISP) Program launched in 2016 to retain high-performing school leaders. The program offers high-performing principals increased autonomy through four key components: (1) Exemption from network membership and network chief oversight. (2) Exemption from budget and Work Plan (CIWP) approval.[8] (3) Increased flexibility in managing budgets and purchases. (4) A requirement to remain in the principal role at their current school for at least two years.

Principals had to apply for ISP designation. Because the ISP Program was designed to reward high-achieving principals, the application process was selective. CPS principals underwent a review process (including an application and interview) to earn the designation. Eligibility depends on the principal's school demonstrating strength in at least three categories from the School Quality Rating Policy (SQRP), the district's framework for evaluating annual school performance, which is heavily weighted toward historical academic test score performance (both levels and growth).[9] Nearly all ISP principals were from schools rated either "Exemplary" or "Commendable," reflecting the district's expressed aim to "reward high-performing principals" (Appendix Figure A1). Applicants must also demonstrate, through their application and interview, that they have the internal capacity and a plan to manage the reduced support resulting from their school's independence.[10] Applicants *were not* evaluated based on the quality of their future plans to improve student achievement. Rather, the ISP designation was granted to principals of strong-performing schools who demonstrated an ability to function with reduced district support and oversight. Successful applicants who remain in their school receive the ISP designation the following school year. Focusing on the pre-pandemic period, 84 schools had principals designated ISP between 2016 and 2019.

In a key study of ISP principals, Travlos (2020) found that principals described being "*free from a network structure of oversight and accountability that is fragmented, stressful, and consumes*

[8]The CIWP is the strategic planning process for schools, meeting federal and state requirements for school improvement plans. Without the ISP, principals must have their CIWP approved by network chiefs.

[9]The SQRP weights various indicators, including school growth percentiles in Math and ELA (35%), percentage of students exceeding national growth norms (10%), average school attainment (15%), percentage of English Language Learners making annual progress (5%), average school attendance (20%), school climate (10%), and data quality (5%) (see Appendix Figure A2).

[10]The school's network chief must approve this plan.

valuable leadership time" and reported feeling "*valued and rewarded.*" These findings suggest that although autonomy is central to the reform, its benefits may be realized through increased principal effort, time, and motivation. Expanding on how ISP principals use their autonomy, Travlos (2020) notes that "*principals are deeply mindful of their schools' unique needs,*" "*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.*" These patterns are consistent with descriptive studies of a similar policy implemented in Chicago in 2008 (Steinberg and Cox, 2016; Steinberg, 2014). However, some principals report the increased independence introduced challenges, reporting feeling "isolated" and expressing a greater need for "budget, network, and management supports." This suggests that while the ISP empowered high-performing principals to enact meaningful changes aimed at improving student outcomes, the reduction in district oversight created difficulties for others. I formalize these ideas below.

III.2 Theoretical Framework and Testable Predictions

Drawing on reports from ISP principals and economic theory, I identify two channels through which autonomy may affect outcomes. The first is the **stability channel**, through enhanced principal effort and motivation. Autonomy, when given as a reward, functions as an amenity that increases motivation, increases effort, and reduces burnout (Deci and Ryan, 1985), as confirmed by a meta-analysis of 318 studies linking autonomy to work motivation and reduced strain (Muecke and Iseke, 2019). Also, the program's two-year commitment may improve principal effort and lower turnover, fostering stability and a stronger climate—benefits that extend broadly across schools.

The second channel is the **allocative efficiency channel**. Textbook and training choices can affect achievement (Jackson et al. 2014; Koedel and Polikoff 2017; Kraft et al. 2018; van den Ham and Heinze 2018), and small implementation changes can produce large differences in outcomes, even under a fixed budget Accelerate (2024). As such, granting principals discretion over curriculum, training, and budgets may affect outcomes by enabling better resource allocation. Indeed, before becoming ISPs, many principals expressed frustration with district restrictions on reallocating funds (Travlos, 2020).[11] To fix ideas, I present a framework that formalizes the role of allocation efficiency and highlights how autonomy can generate both positive and negative effects—yielding testable predictions about which schools benefit most from autonomy.

Optimal Choices

Consider a setting in which schools allocate two inputs, x_1 and x_2, priced at p_1 and p_2, respectively. Each school i has a distinct, strictly quasi-concave production function $F_i(x_1, x_2)$ mapping

[11] Liebman et al. (2017) shows that intertemporal non-fungibility due to expiring budgets can lead to inefficient end-of-year spending.

inputs to student achievement. Schools face a common budget B, and the constraint $p_1x_1 + p_2x_2 = B$ binds. Substituting for x_2 yields the indirect production function:

$$f_i(x_1) = F_i\left(x_1, \frac{B - p_1x_1}{p_2}\right).$$

Each school maximizes $f_i(x_1)$ over its feasible input range, leading to an optimal choice:

$$x^*_{1i} = \arg \max_{x_1 \in [0, B/p_1]} f_i(x_1).$$

By quasi-concavity, achievement is maximized at x^*_{1i} and falls monotonically as x_1 moves away from this optimum. This is illustrated in Figure 2.

Centralization

Under centralization, the district imposes a uniform input level x_{1d} across schools—reflecting how "*central office leaders can limit school leaders' decision-making power*" (Wong et al., 2020) and provide "*a one size fits all type of support*" (Travlos, 2020). Formally, schools are constrained to $x_{1i} = x_{1d}$. Since $x^*_{1i} \neq x_{1d}$ for most schools, achievement generally declines, capturing the cost of "*misalignment between school and district priorities*" (Steinberg and Cox 2016; Travlos 2020).

Principal Quality

Principals differ in their objectives. *Aligned* principals aim to maximize student achievement, while *misaligned* principals pursue alternative goals—such as athletics, staff preferences, or socioemotional outcomes—or lack the capacity to make optimal decisions. Because the analysis focuses on test-based achievement, I define alignment as principal quality. Under autonomy, aligned principals choose x^*_{1i}, while misaligned principals select $x_{1p} \neq x^*_{1i}$.

Results

Proposition 1: Achievement gains from autonomy are (weakly) larger for aligned principals, holding holding the centralized input choice x_{1d} fixed. This aligns with research showing that autonomy is most strongly associated with achievement gains in settings where principals are capable and accountable (Galiani et al., 2008; Stiefel et al., 2003; Fuchs and Wößmann, 2007).

Proof. Let the achievement gain of autonomy for an aligned principal be $\Delta_{\text{aligned}} = f_i(x^*_{1i}) - f_i(x_{1d})$, and that for a misaligned principal to be $\Delta_{\text{misaligned}} = f_i(x_{1p}) - f_i(x_{1d})$. Then, $\Delta_{\text{aligned}} - \Delta_{\text{misaligned}} = f_i(x^*_{1i}) - f_i(x_{1p})$. By strict quasi-concavity, $f_i(x^*_{1i}) > f_i(x_{1p})$, so $\Delta_{\text{aligned}} > \Delta_{\text{misaligned}}$. □

Proposition 2: The benefits of autonomy are larger for schools whose optimal input x^*_{1i} is further from the centralized input x_{1d}. This reflects Acemoglu et al. (2007)'s insight that decentralization yields greater gains when individual needs are heterogeneous.

Proof. Let the maximum potential benefit from autonomy be $\Delta_{\text{aligned}}(x_{1d}) = f_i(x^*_{1i}) - f_i(x_{1d})$. Since f_i is strictly quasi-concave and maximized at x^*_{1i}, Δ_{aligned} increases as x_{1d} moves away from x^*_{1i} in either direction. $\square$

Proposition 3: The effects of autonomy on student achievement are heterogeneous and may be positive or negative. Autonomy improves outcomes when the principal is better aligned with the optimum x^*_{1i} than the district, and worsens outcomes when the district is better aligned.

Proof. Let the gains from autonomy for any principal be $\Delta_i = f_i(x_{1p}) - f_i(x_{1d})$. Since f_i declines as x_1 moves away from x^*_{1i}, if x_{1p} is closer to x^*_{1i} than x_{1d} (on the same side of x^*_{1i}), then $\Delta_i > 0$; if farther, $\Delta_i < 0$. If x_{1p} and x_{1d} lie on opposite sides of x^*_{1i}, the sign of Δ_i depends on which input level yields higher output. $\square$

While the stability channel likely benefits all schools, the framework shows that the allocative efficiency channel can raise or lower achievement—making the overall effect of autonomy theoretically ambiguous. Given this ambiguity and the heterogeneous effects documented in Section II, I estimate the *average* effect of ISP in Section VI.1 and examine the causal mechanisms outlined here in Section VI.4—highlighting implications for policy debates around school autonomy.

IV Data

I collect data from several public sources. I obtained a list of ISP principals and their designation years from the (CPS website) and matched each school principal to a database of all CPS schools in 2014–2015 (the year preceding the first ISP designations). Achievement data come from the state assessment and accountability measures for Illinois public school students: Illinois Standards Achievement Test (ISAT) (2001-2014), the Partnership for Assessment of Readiness for College and Careers (PARCC) (2014-2020), and school-level passing rates from the Common Core of Data (CCD) (2010-2020).[12] To focus on data straddling ISP implementation, I use the PARCC data when available, followed by the CCD and ISAT data. The dataset reports the percentage of grades 3-8 students who meet the state proficiency standard at each school from 2001 through 2019.

Using the CCD, I extract information on total enrollment, the percentage of students eligible for free or reduced-price lunch, racial composition, bilingual status, and special education services for 2001-2019. I then match these demographic and performance data to the list of ISP principals by school and year.

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

Additional data on school climate and organization come from the 5Essentials for 2014-2020, which measure school climate in five domains: (1) Effective Leaders, (2) Collaborative Teachers, (3) Involved Families, (4) Supportive Environment, and (5) Ambitious Instruction. Publicly available data report each school's rating from 1 to 5 across all domains.[13] 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). I use the effective leadership domain scores as a key measure of principal quality. This metric identifies schools where leadership aligns people, programs, and resources around a clear improvement vision.[14] The leadership score summarizes four dimensions: Teacher Influence, Program Coherence, Instructional Leadership, and Trust.

These data are also linked to personnel files from 2010-2020.[15] The personnel data include each CPS position and associated salary, allowing me to code principal turnover and spending on personnel categories (instruction, support staff, special education). I also obtain official measures of teacher (one-year) and principal (six-year) turnover from the Illinois Report Card.

The final dataset includes all public schools in CPS between 2010 and 2019. A total of 84 schools were designated as ISPs between 2016 and 2019. I exclude data from 2020-2022 to avoid conflating results with the effects of school shutdowns associated with COVID-19. As discussed above, the schools that became ISPs were not typical of other schools in the district. Table 1 presents summary statistics for ISP schools (column 2) and non-ISP schools (column 1) along with the p-value associated with the test of equality of means for the two groups of schools.

On average, ISP schools had higher achievement levels and more advantaged student populations than non-ISP schools. Proficiency rates in ISP schools were 43.99 percent in ELA and 43.52 percent in Math, compared to 31.92 percent and 30.84 percent, respectively, in non-ISP schools. ISP schools also scored higher on the 5Essentials survey, averaging 4.42 versus 3.81—a 0.6-point difference corresponding to a sizable 0.45σ gap in school climate. Demographically, ISP schools had a higher share of White students (15 percent vs. 6 percent), a lower share of Black students (25 percent vs. 57 percent), and fewer students eligible for free or reduced-price lunch (76 percent vs. 85 percent). Each of these differences is statistically significant at the 1 percent level.

While some non-ISP schools resemble ISP schools, non-ISP schools differ substantially from ISP schools *on average*. This motivates two strategies: leveraging within-school variation for identification and using matching methods to construct a comparison group of schools that may have followed similar pre-treatment trajectories. I detail these approaches below.

[13] Obtained from the SQRP reports.

[14] From this perspective, effective leaders practice shared leadership, set high goals for instructional quality, maintain mutually trusting and respectful relationships, support the professional growth of faculty and staff, and manage resources for sustained program improvement.

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

V Empirical Strategy

To address potential bias from differences between ISP and non-ISP schools, I employ a design that compares outcomes before versus after ISP designation among ISP schools, eliminating the influence of time-invariant disparities such as differences in student composition or pre-ISP performance. I then isolate the ISP effect from other underlying changes over time by comparing changes within ISP schools to those for a carefully-selected comparison group of observationally-similar schools that never became ISPs during the sample period (eliminating the negative weight problem associated with naive two-way fixed effects models).

The model assesses whether ISP schools had differentially improved outcomes after ISP designation than non-ISP schools, only among schools that were observationally similar (thus likely to have common time shocks) before program introduction. The key identifying assumptions are that the carefully-selected non-ISP schools would have had similar outcome trajectories as ISP-designated schools, and that the timing of ISP designation is unrelated to other contemporaneous changes at ISP schools. I present empirical tests indicating the likely validity of both assumptions.

To form a comparison group of schools, I match each ISP school to a set of non-ISP schools with the most similar pre-ISP characteristics, using Mahalanobis distance to compare demographics (free lunch status and poverty rates), school quality ratings, number of teachers, school climate scores two years before designation, and math and ELA scores both two and three years before designation.[16] Each ISP school is put in a data set with a fixed number of the best non-ISP matches, which constitutes group g. All of these group g datasets are appended to create a stacked dataset that includes a mini-dataset for each treatment-school group.

Using the stacked-matched dataset, I estimate models as below for outcome Y for each school s, in treated stack group g, in each year t, Y_{sgt}.

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

Here, ISP_s equals one if school s was designated as an ISP between 2016 and 2019. For treated schools, τ denotes the year relative to ISP designation (i.e., event time); for non-treated schools, τ is set to -10. The indicator 1_τ equals one for observations in ISP schools during year τ. The coefficients β_τ trace the evolution of outcomes in ISP schools relative to comparison schools before and after ISP designation.

To (a) account for differential time shocks across school types and (b) ensure that only never-

[16]Mahalanobis distance measures the distance between two points while accounting for the variance and correlation structure of the data. It is calculated as $D_M(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}$, where x and y are vectors of covariates for two units, and S^{-1} is the inverse of the covariance matrix of the covariates.

treated schools form the comparison group, I include school-group-by-year fixed effects, γ_{gt}. The term ε_{sgt} is a school-level error that varies by group, since some control schools are reused across treated units. Standard errors are clustered at the school level. This approach follows the stacked difference-in-differences framework used by Cengiz et al. (2019) and Deshpande and Li (2019).

A key aspect of this estimation approach is the number of matches. Fewer matches generally yield higher match quality but reduce precision, and while there is no fixed rule, using between 3 and 10 matches is generally recommended as a balance between bias and precision (Imbens and Rubin, 2015). To allow the data to inform this choice, I estimate the ISP effect on math passing rates and use the standard error as a measure of precision. Figure 3 plots the estimated standard error against the number of matches. The standard error is highest with just two matches at 1.385 and decreases to about 1.08 with 30 or more matches. However, the improved precision is non-linear, with minimal gains beyond 7 matches. Given that bias is likely lower with fewer matches, 8 matches appears optimal, aligning with the rule of thumb. Informed by these patterns, I match each ISP-treated school with its 8 closest non-ISP schools to define the comparison group. To assuage any concerns that this choice affects the results, Section VI.2 shows that **the main results are robust to using as few as one and as many as 20 matches**.

To assess match quality, Table 1 presents summary statistics for ISP schools (column 2) and matched non-ISP schools (column 4), along with *p*-values from tests of equality of means. Consistent with good covariate balance, all student demographic variables and several outcomes—attendance rates, mobility, and ELA and Math state assessment passing rates—are statistically indistinguishable between groups. While some school climate measures and math percentile scores show significant differences, these are outcomes directly affected by the treatment itself. Importantly, Appendix Table A1 demonstrates that these outcomes were balanced before 2015, before any ISP designation – confirming that the differences emerge only *after* the ISP was implemented.

VI Results

VI.1 Average Effects

Figure 4 presents the event study plots for the matched-stacked sample, with effects on ELA and Math passing rates shown in the top and lower panels, respectively. Each event study estimate is displayed relative to the year prior to ISP designation, along with its 95 percent confidence interval.

The first notable pattern is that for both ELA and Math passing rates, there is no evidence of differential pre-trends. While scores are mechanically matched in years $t-2$ and $t-3$, outcomes in years $t-4$ through $t-8$ do not differ between ISP schools and matched non-ISP schools, suggesting that restricting the comparison group to similar schools produced comparison schools with very similar outcome trajectories over time (and thus likely exposed to the same time-varying shocks) as ISP schools. Note that, after matching, this model does not require additional covariates for the common trends assumption to hold, avoiding potential problems associated with models that rely on time-varying covariates for credible identification (Caetano and Callaway, 2023).

Looking at the post-ISP years, $t = 0$ and later, pass rates in both subjects improve by between one and two percentage points immediately in year zero, and increase by between five and seven percentage points three years later. To directly address any concerns that results are driven by method choice, Appendix Figure A5 presents the Callaway and Sant'Anna (2021) estimator—an alternative approach to addressing the negative weight problem—which yields very similar results with no differential pre-trending and improved passing rates of about five percentage points after ISP designation.

To summarize these average results, I estimate a simple before-versus-after effect using the treated and matched schools (Table 2, lower panel). On average, following ISP designation, ELA and math passing rates increase by 4.144 and 3.141 percentage points, respectively (both with $p - values < 0.01$). After two years of ISP designation, these effects grow to 7.079 and 5.410 percentage points, respectively (both with $p - values < 0.01$). Across both subjects, the estimated increase in passing rates averages approximately 3.6 percentage points (about 0.09σ) initially, increasing to about 6.25 percentage points (roughly 0.16σ) after three years under the program.

Effect on Different Margins

Because proficiency represents scoring above a single performance level, observed gains might simply result from moving more students across the proficiency threshold rather than reflecting broader improvements across the achievement distribution. If the proficiency threshold is set relatively low, achieving proficiency may not indicate meaningful overall improvement. Moreover, strong incentives to raise proficiency rates can lead to a focus on students near the threshold—the so-called "bubble kids" effect (e.g., Neal and Schanzenbach (2010))—while performance for other students could be unchanged or even decline.

To assess effects across different achievement levels, I examine changes in the proportion of students at various performance tiers. The PARCC test reports the share of students who are below, partially meeting, approaching, meeting, or exceeding the grade-level state standard. If schools focus their efforts primarily on students near the passing threshold, one would expect to see fewer students exceeding the standard and minimal improvement among those performing well below it. I estimate the effects of ISP designation on the percentage of students scoring at each performance level, with results for treatment years 1, 2, and 3+ plotted in Figure 5.

Figure 5 shows that after ISP designation, fewer students did not meet, partially meet, or approach the standard, while more students met and exceeded the state proficiency standard. This pattern is consistent with improvements throughout the achievement distribution rather than gains concentrated solely among students near the proficiency threshold. This broad improvement is most evident for math, where the largest increases in years 3+ appear in the ”exceeded” category and the largest reductions are in the ”did not meet” category—rather than just in categories near the passing threshold. However, ELA improvements are most pronounced in the ”met” category, suggesting that while there were across-the-board improvements in ELA, gains were greater among students near the threshold.

The second way to assess improvement margins is by examining average performance. Although I do not have an average performance measure for the state test, the district accountability data include each school's average percentile score for grades 3 through 8 on the district test (the NWEA), available beginning in 2014. The estimates for years one through three are reported in columns 3 (math) and 4 (ELA) of Table 1. Math performance on the district test increased by 2.45 percentile points in the first year of ISP designation ($p-value < 0.1$) and by 4.85 percentile points by year three ($p-value < 0.01$). ELA effects are somewhat smaller, with increases of 0.86 percentile points in the first year (not significant) and 2.5 percentile points by year three ($p-value < 0.05$). With normally distributed test scores, these increases correspond to about 0.125 and 0.06 standard deviations, respectively. These results indicate non-trivial gains on the district test that align closely with the observed increases in passing rates on the state test. Pooling both subjects, the before versus after effect is 2.2 percentile points ($p-value < 0.05$, about 0.055σ), increasing to 3.7 percentile points after two years ($p-value < 0.01$, about 0.093σ).

It is worth noting that the effects on passing rates for the state tests are larger for ELA than for Math, whereas the opposite is true on the district tests. This may reflect a greater focus on marginal students for the state tests (as suggested by the patterns in Figure 5) or differences in what the state and district assessments capture. Either way, ISP designation is associated with sizable statistically significant performance gains on standardized tests in both subjects.

Putting The Average Effects Estimates in Context

To put these average effects in perspective, I consider the pooled results from the meta-analysis. The simple before versus after pooled effect is a 3.64 percentage point increase (se=0.77) in the passing rate, corresponding to an increase of roughly 0.09σ. This is larger than the pooled average of 0.022σ from other studies, but similar to and within the range of many reported estimates from the literature. Because 0.09σ is a noisy estimate, I use the existing literature to provide an improved prediction of the true Chicago ISP effect. Taking a Bayesian perspective, I borrow strength from other studies (Efron and Morris 1973; Morris 1983) to inform the true effect in this context. This is analogous to creating empirical Bayes estimates for teacher effects (as in Kane and Staiger (2008); Jackson and Bruegmann (2009)) by using a weighted average of the noisy estimate (0.09σ) and the pooled average (0.022σ), where noisier estimates are weighted closer to the pooled average.

Formally, if 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 (4) where $B = (\sigma_j^2)/(\sigma_j^2 + \tau^2)$.

$$E(\theta_j|\hat{\theta}_j, \sigma_j, \tau) = B \times \Theta + (1 - B) \times \hat{\theta}_j \tag{4}$$

Replacing σ_j, τ, and Θ with their estimates, (4) yields the Best Linear Unbiased Prediction (BLUP) of the true effect for study j. The BLUPs ($\tilde{\theta}_j$), also called Empirical Bayes estimates, are weighted averages of individual estimates and the pooled average, where more precise estimates receive greater weight. Constructing an Empirical Bayes estimate yields a predicted true Chicago ISP effect (given both the raw estimate and information from the extant literature) of 0.081σ.

To put the magnitude of this effect in perspective, increasing teacher quality by one-half of a standard deviation increases test scores by less than this amount (0.06σ) and raises lifetime earnings by \$3,500 (Chetty et al. 2014). From Jackson and Mackevicius (2024), this effect is similar to increasing school spending by about \$3,000 per pupil and is above the 95thth percentile of the distribution of effects from increased school spending by \$1000 per pupil, comparing very favorably to effective uses of school resources. However, the ISP program costs only about \$30,000/800 pupils = \$37 per pupil, implying a remarkable cost-benefit ratio of over 1 to 100.

The high cost-effectiveness is unsurprising because benefits arise not from moving along the production possibilities frontier (e.g., hiring more or better teachers) but from shifting toward the frontier through increased allocative efficiency and/or effort. While the effects indicate very high cost-effectiveness, the magnitude of the effect is modest so the ISP cannot bring all children to proficiency (average proficiency rate: 47 percent). However, the gains are economically meaningful, cutting the non-proficiency rate by around a tenth at low cost within just a few years.

VI.2 Threats to Validity

The validity of the estimates in Section VI.1 relies 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 ISP schools at the same time as ISP designation. In this section, I provide evidence supporting both of these identifying assumptions.

Common Trends in Outcomes and Predictors

To show that treated schools were on a similar trajectory as comparison schools before ISP designation, I first examine the event study models. Figure 4 reveals no visual evidence of pre-trends in ELA or Math passing rates. This is confirmed by formal tests comparing effects in years $t-3$ through $t-8$ against those in $t-1$, yielding p-values of 0.6 and 0.67 for ELA and Math, respectively.

I also test for pre-trends in *predicted* outcomes—constructed using the linear projection of pre-2016 covariates (i.e., school enrollment, student mobility rate, poverty rate of students' census blocks, percent of bilingual students, free lunch percentage, and racial composition) on passing scores—and find no significant differences between effects in years $t-3$ through $t-8$ relative to those in $t-1$ ($p = 0.35$ for Math; $p = 0.27$ for ELA).[17] Indeed, the event-study models on predicted outcomes, shown in red in Figure 4, show no evidence of differential pre-trending in predicted passing rates for either subject—supporting the assumption of common trends in both observed and predicted outcomes.

No Confounding

To assess potential confounding, I test whether ISP designation affected predicted outcomes. If changes in school size, neighborhood characteristics, or student composition influenced achievement, such effects would likely appear in *predicted* outcomes. However, despite sizable improvements in actual outcomes, there is no change in predicted passing rates in either subject in the before-versus-after models (see lower panel of Table 2). Similarly, the lower panel of Table 3 shows no statistically significant or economically meaningful changes in key observable predictors of school-level student achievement.

As an additional 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 (see Table 4). If the main results were driven by confounding, one might expect them to disappear when focusing on the closest match. On the contrary, the results are stronger. Using only the best matched school yields negligible effects on predicted outcomes and large effects on actual outcomes—suggesting minimal bias in the reported results.[18]

Student Selection

Because I use aggregate school-level data, improved outcomes could reflect motivated parents sending their children to ISP schools rather than true ISP effects. While Table 3 shows no changes in observable student demographics, this does not rule out shifts in *unobserved* student composition. However, elementary enrollment is largely determined by residential location, so implausibly large residential changes—limited only to unobserved dimensions—would be required to account for the observed effects.

To provide a partial test of this, I use the fact that the district test (NWEA) reports RIT scores for all tested students as well as for the subset linked to pretest data.[19] The difference between these groups reflects differences between students continuously enrolled at CPS (stayers) and those who

[17] The R-squared in above 0.6, indicating strong predictive power within schools.
[18] Estimated effects on individual covariates in Appendix Table A3 are consistent with this.
[19] The RIT scale is independent of grade level.

entered their school from private schools, out of state, or during the second half of the school year (movers). A straightforward test for differential selection following ISP designation is whether the gap between movers and stayers changes after treatment. To make the test meaningful, I focus on grades and subjects with the most pronounced improvements on the district test.

Table 4 presents ISP designation effects on average RIT scores for elementary (grades 3–6) and middle grades (grades 7–8) in both subjects—showing clear gains in elementary math and reading but no significant effects in middle school. If selective student movement were driving the results, we would expect large and statistically significant shifts in the gap between movers and stayers, particularly in the elementary grades. Yet, there is no evidence of such a shift. The estimated effects on this difference are statistically insignificant and too small to account for the observed gains. Even assuming an unrealistically high share of movers—20 percent—the effect on aggregate scores would be only 0.084 for math and 0.056 for ELA, far below the observed effects.

Selection of School Based on Future Gains

The final key concern is that district leaders may have selected principals they believed would generate future gains, resulting in selection on anticipated trends. This is unlikely for several reasons. First, ISP designation was explicitly designed to reward demonstrated past excellence and assessed principals based on their capacity for autonomous leadership—not their predicted future performance. Second, existing research shows that policymakers are generally poor at forecasting teacher and school effectiveness (Jacob and Lefgren, 2008; Donaldson et al., 2021; Grissom et al., 2018). It is therefore implausible that district leaders could have reliably identified schools primed for substantial improvement. Nonetheless, to formally assess the level of predictive accuracy that would be required to generate the observed gains, I conduct a permutation test.

In this exercise, ISP treatment assignments were randomly reassigned to schools within the same matched school group, and treatment effects were re-estimated. Figure VII shows the distribution of average effects on math and ELA passing rates across 1,000 replications. None of the placebo assignments produced gains—averaged across both subjects—as large as those actually observed. This implies that, to generate the observed results through selection alone, district leaders would have needed near-perfect foresight to identify, among otherwise similar schools, those most likely to improve. To rule out the possibility that this finding is driven by restricting permutations within matched groups, I also conduct permutations *across* matched groups. Again, none of the 1,000 replications produced gains as large as those observed.

Taken Together

The results above strongly suggest that the estimated ISP effects are causal. Additional evidence presented in this paper reinforces this conclusion. Specifically, Section VI.6 shows that many non-ISP schools would have experienced large ISP effects, indicating that principals were selected based on past performance, not expected future growth. Moreover, Section VI.3 presents patterns predicted by the theoretical model that are inconsistent with both selection-based and confounding-based explanations.

VI.3 Heterogeneity

The results indicate that the ISP effect is positive, *on average*. However, I show that while the pooled average is clearly above zero—with a tight confidence interval—negative true effects may nonetheless occur. As suggested by the framework in Section III.2, ISP effects are likely to be heterogeneous (i.e., they vary beyond sampling error), and some schools may experience negative impacts. I test this directly by exploiting the stacked data structure. Because each treated school is matched to its own set of comparison schools, I estimate treatment effects separately for each ISP school using the following difference-in-differences specification, applied within each matched group g:

$$Y_{sgt} = \beta_g(ISP_s \times 1_\tau) + \gamma_s + \gamma_{t,g} + \varepsilon_{sgt} \tag{5}$$

This model yields individual ISP effect estimates ($\hat{\beta}_g$) for each school designated between 2016 and 2019. Because these estimates are noisy—where the noise is approximated by the squared standard error se_g^2—the raw dispersion of $\hat{\beta}_g$ overstates the true heterogeneity. To recover the distribution of true effects, I apply an approach motivated by hierarchical Bayesian modeling.

Each school has a true ISP effect θ_g, and due to sampling error, the estimated effect follows the distribution in (7). This normality assumption is justified by the central limit theorem:

$$\hat{\beta} g \sim \mathcal{N}(\theta_g, \sigma g^2) \tag{6}$$

The true school-level effects θ_g deviate from the overall mean effect Θ due to underlying heterogeneity, with variance τ^2, governed by some unknown distribution $g(\tau)$. The empirical challenge is to recover both the extent of this heterogeneity (τ^2) and the shape of the distribution $g(\tau)$.

Assuming that the sampling variance for each estimate (σ_g^2) is well approximated by the squared standard error (se_g^2), one can estimate the extent of true heterogeneity—i.e., variation in the estimates not attributable to sampling error. To estimate τ^2, I follow the method of DerSimonian and Laird (1986), which computes a precision-weighted variance of the raw estimates and subtracts the variance expected due to sampling variability (based on the observed estimation errors). This yields an estimate of the variance of true effects, τ^2. Notably, this approach does not require any distributional assumptions about the heterogeneity. Estimates are reported in Table 6, which also presents a precision-weighted average ISP effect across schools.[20] This provides a more efficient estimate of the average treatment effect than the equal-weighted averages reported in Table 2.

[20]Each estimate is weighted by $1/(se_g^2 + \hat{\tau}^2)$.

Pooling both subjects (and accounting for clustering at the school level), the estimated heterogeneity (τ) is 6.3, while it is 6.378 for ELA and 5.914 for math. One cannot reject equality of the two subject-specific distributions at the 5 percent level. These estimates imply that for any two randomly selected 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 3.76 percentage points ($p - value < 0.01$), this reflects substantial heterogeneity: some true effects are likely negative, while others are large and positive. I present evidence on this below.

The Distribution of True Effects

To estimate the distribution of true effects, $g(\tau)$, based on the observed estimates, I apply a deconvolution kernel density estimator (Carroll and Hall, 1988), as implemented by Kato and Sasaki (2018). This method identifies the distribution of an unobserved variable X_i using two noisy measurements, X_{1i} and X_{2i}.[21] In this context, I use the estimated ISP effects on math and ELA passing rates for each school as two noisy measures of the same underlying treatment effect. To assess the plausibility of this assumption, I test whether the math and ELA effects have the same mean (using a two-sample t-test) and the same distribution (using the Kolmogorov–Smirnov test). In both cases, I fail to reject equality, supporting the use of these two estimates as separate measures of a common underlying effect.

This approach does not require the two measures to be independent and allows for the estimation of a uniform confidence band for the density. To balance local and global features of the estimated distribution, I follow Efron and Tibshirani (1996) in selecting the tuning parameter (i.e., the kernel bandwidth) to match moments in the data. Specifically, following Walters (2022) and Jackson and Mackevicius (2024), I choose the bandwidth such that the standard deviation of the deconvolved distribution matches the unbiased estimate of the standard deviation of true effects obtained above. Figure 7 displays the distribution of ISP effects: the deconvolved density (and its 95 percent confidence interval), the implied normal distribution based on the variance and mean, and the raw estimates for math and ELA.[22]

The first notable pattern is that the estimated distribution of true effects is well approximated by a normal distribution, with the normal curve falling within the 95 percent confidence band of the estimated density. following Wang and Lee (2020), I implement a formal test which fails to reject that the distribution is normal (see Figure A7). The figure also shows that the distribution of estimated effects spans a wide range—from approximately –15 to +24 percentage points. However, this spread may partly reflect sampling variability rather than true treatment heterogeneity. Indeed,

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

[22] The deconvolved mean is computed by integrating over the estimated density and equals 2.968. This is slightly below the meta-analytic mean because the deconvolved distribution places less mass in the lower and upper tails compared to the raw estimates.

the deconvolved distribution of true effects is somewhat narrower than the distribution of raw estimates. Even so, it spans values well below and well above zero, indicating that true ISP effects can be meaningfully negative or strongly positive. I now turn to whether this heterogeneity aligns with 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 allocative efficiency as a key channel through which increased autonomy (or decentralization) can affect outcomes. Two dimensions emphasized in the model are principal–school alignment (a proxy for principal quality, narrowly construed) and school–district alignment (reflecting underlying heterogeneity). This section presents evidence consistent with both mechanisms mediating the allocative efficiency channel.

Principal Alignment/Quality

The theoretical framework predicts that principals more strongly aligned with improving test scores will generate larger ISP effects on math and ELA passing rates. I test this prediction using two distinct measures of principal alignment toward academic outcomes, both constructed in the year *prior* to ISP designation.

- The first measure is the residual passing rate in ELA and math, net of school-level characteristics including poverty rate, the fraction of students eligible for free lunch, enrollment, and the share of Black, Hispanic, and special education students. This captures the extent to which a school outperformed expectations based on student demographics. 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.

- The second measure is based on teacher survey responses and serves as a proxy for principal quality. Although it captures leadership broadly oriented toward "sustained improvement," academic achievement is a key component of the district's success criteria, and prior work shows this measure correlates with principal test-score value-added (Laing et al., 2016).

To create a single measure of alignment, I construct a composite index of principal quality by averaging the standardized (z-score) values of the two components, following the approach of Kling et al. (2007). For transparency, I also report results using each measure separately.

Using these measures, I regress the ISP effect (the estimated β_gs from Equation (5)) 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,[23] 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. For improved precision, I pool estimates from both subjects and account for correlated errors at the school level. 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)).

The regression results are presented in Table 7. Both measures reveal the same basic pattern: schools with "better" principals—those more aligned with improving achievement—experience larger ISP effects. Examining the two dimensions separately, both the leadership score and the residual pass rate measure predict larger ISP effects for both subjects. A one standard deviation increase in the leadership score (based on survey data) is associated with an ISP effect on passing rates that is 0.073 percentage points higher (p-value < 0.1), while a one standard deviation increase in the residual passing rate is associated with an ISP effect that is 1.374 percentage points higher (p-value < 0.01). Looking at the combined measure, a one standard deviation increase in overall principal alignment is associated with an ISP effect on passing rates that is 1.868 percentage points higher (p-value < 0.01). For context, this means the ISP effect is 1.868 percentage points larger for a principal at the 85th percentile of the alignment distribution than for one at the mean.

Given that the pooled average effect is 3.67 percentage points, this implies ISP effects remain positive for principals with alignment z-scores above approximately -1.96. In other words, ISP improves outcomes for the vast majority of principals, except those with very low measured alignment (a proxy for quality) within the ISP group. The fact that small or negative effects are concentrated among principals with historically poor outcomes or low leadership ratings suggests these negative ISP effects likely arise because weaker principal (when given more autonomy) make decisions misaligned with improving student achievement.

Heterogeneity (District-School Alignment)

Another 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 align well with the generic district position by having a student population demographically quite different from the typical district school. Specifically, I code schools as outliers (i.e., more likely to have specific needs) if they exceed the 90^{th} percentile for percent white, percent Hispanic, percent free and reduced-price lunch, percent bilingual, or percent special education. I do not use percent black because many schools exceed 90 percent black, while this is less true for other ethnic categories. Schools are coded as outliers if they fall into any category, and I also code up the number of categories. I then

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

regress the ISP effect on an indicator for outlier status.

For both subjects, there is evidence that outlier schools have larger ISP effects. The coefficient on the outlier indicator in column 4 is 2.868 ($p < 0.05$). This indicates that, *all else equal*, the ISP effect is considerably larger for schools that are outliers in at least one demographic category. That is, schools that are outliers in some demographic category (a proxy for having specific needs that may not align well with district choices) have ISP effects that are 2.86 percentage points larger than non-outliers. Note that 2.86 percentage points corresponds to an increase of about 0.07σ—a modest but economically significant effect. 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)).

Both Channels Together

In models that include both principal quality and outlier status, the basic results remain largely unchanged—implying that schools with outlier student populations were neither more nor less likely to have strong principals. Column 5 of Table 7 shows that, with both variables included, a one standard deviation increase in principal quality raises the ISP effect on passing rates by 1.75 percentage points (p-value < 0.01), while outlier schools have ISP effects on passing rates that are 2.56 percentage points higher (p-value < 0.1).

A methodological concern with this approach is that it requires that the standard errors for individual school ISP effects are valid. However, with only a single treated unit per group, the standard errors may not be consistent, which could lead to inaccurate inference in the random effects meta-regression. To address this, I also present results from a simple OLS model that ignores standard errors entirely (column 6), and from the precision-weighted approach in Card and Krueger (1992) (column 7). Both alternatives yield very similar results, indicating that the conclusions are robust to potential mismeasurement of school-level standard errors.

To visualize these mechanisms, Figure 6 plots the standardized principal alignment measure against the raw estimated ISP effects. As the regression results show, there is a clear positive relationship between the two. The right panel presents a box plot of ISP effects by the number of outlier categories. ISP effects increase from 0 to 1, from 1 to 2, and from 2 to 3 categories—suggesting that being an outlier in more categories is associated with larger effects – reinforcing the heterogeneity result. Appendix Figure A6 shows analogous figures based on different numbers of matches, with very similar patterns. In the meta-regression, both the alignment measure and outlier status explain about 12 percent of the true heterogeneity in treatment effects.[24] That is, even with relatively crude proxies for the extent to which autonomy increases allocative efficiency, these measures account for more than a tenth of the variation in effects across schools.

[24] In a model with no covariates versus one including these two predictors, τ^2 declines from 39.75 to 34.95.

Testing the Stability Channel

The other mechanism discussed in Section III.2 is the stability channel. A key motivation for the ISP program was to retain effective principals by making the job more rewarding through increased autonomy, while also requiring a two-year commitment to remain at the current school. For both reasons, a decline in turnover is expected during the first two years. However, if the reduction were driven solely by the commitment, turnover would have returned to pre-treatment levels afterward. In contrast, if principals value the job more due to greater autonomy, the reduction in turnover would persist beyond the initial two years.

To assess whether ISP designation improves principal retention, I estimate its effect on the likelihood that a school has a new principal in each year following adoption. Results are reported in Table 8. Column 1 shows that ISP schools are 3.57 and 3.02 percentage points less likely to have a new principal in the first and second years, respectively. I also examine an alternative turnover measure reported by the district: the fraction of principals in a given year who were not at the school six years earlier. As shown in column 2, this measure also shows a notable reduction in turnover during the first two years following ISP adoption, indicating that some principals who might otherwise have left chose to remain. These reductions are consistent with the program's two-year commitment requirement.

Strikingly, for both measures, the reduction continues into year 3—with a 3.2 percentage point decline in the likelihood of having a new principal three years after designation ($p-value < 0.01$). This suggests that ISP principals may have valued the increased autonomy and chose to stay beyond the required period. This interpretation is supported by qualitative evidence: principals reported that before becoming ISP, they "*started looking for other jobs*," but later "*described independence as a motivational boost*" and "*felt rewarded for their demonstrated success*" Travlos (2020).

To put this reduction in perspective, approximately 1 in 10 principals in Chicago leave their school in a given year Sartain (2023). Turnover rates are lower in primary schools (closer to 1 in 8) and even lower in high-achieving schools (as low as 1 in 5). Relative to this baseline, a 3.57 percentage point reduction in the likelihood of having a new principal in the first year represents a 30 to 60 percent decrease in turnover.

While this is a large relative effect, it remains modest in absolute terms and cannot explain the observed test score gains. Although meaningful, a 3.57 percentage point reduction is far too small to generate achievement gains of 0.081σ or more. To gauge how much reduced turnover could plausibly affect test scores, I multiply 0.0357 by the estimated effect of principal turnover on achievement from other studies. Some studies find no significant effect Weinstein et al. (2009), while others report modest impacts ranging from 0.007σ Béteille et al. (2012) to 0.01σ Henry and Harbatkin (2019), with the largest estimates around 0.04σ Miller (2013). Even using the the largest estimate (a likely upper bound), principal stability would imply an effect of only 0.0014σ ($0.0357 \times 0.04\sigma$)—more than an order of magnitude smaller than the BLUP estimate of 0.081σ. This suggests that while reduced turnover may contribute to the program's success, it cannot be the primary mechanism through which the ISP improves test scores.

VI.5 Further Evidence on Mechanisms

By design, schools are expected to pursue a range of strategies when granted increased autonomy, leading to substantial heterogeneity in the specific changes implemented. This variation poses a challenge for identifying the mechanisms behind the ISP's effects. Nevertheless, certain variables can still offer meaningful insight into those underlying processes.

School Climate: Theoretically, reduced principal turnover and increased effort should foster greater stability and a more positive school climate. In addition, changes in principal decision-making—such as implementing reforms or adopting practices valued by teachers and students—are likely to translate into improved climate. To assess this, I estimate the average effect of ISP designation on school climate, measured on a 1-to-5 scale. Consistent with this hypothesis, ISP schools experienced a 0.285-point increase ($p-value < 0.01$), equivalent to an effect size of approximately 0.22σ. This suggests that both teachers and students viewed ISP schools as better learning environments after ISP-designation.

To evaluate whether the observed improvement in school climate aligns with the effects on test scores, I draw on two relevant studies. Porter et al. (2023) finds that a one-standard-deviation (1SD) increase in school climate corresponds to a 0.8SD increase in school effectiveness, while Jackson et al. (2023) estimates that a 1SD increase in effectiveness yields a 0.08σ gain in test scores. Based on these estimates, the observed 0.22σ improvement in school climate implies a test score gain of approximately 0.0145σ ($0.22 \times 0.8 \times 0.08$)—about 18 percent of the BLUP of the ISP effect. While it is unclear whether improved school climate caused the test score gains, vice versa, or whether both reflect broader improvements under the ISP, the results clearly indicate that the program facilitated meaningful enhancements in school climate—improvements that can plausibly explain *some* of the observed achievement gains.

Given the improvement in school climate, one might expect corresponding gains in student attendance and teacher retention. I assess this using the regression model presented in Table 8, columns 4 through 6. There is no evidence that ISP schools experienced a reduction in chronic truancy or an increase in student attendance. Likewise, there is no evidence of a meaningful improvement in teacher retention. These results suggest that the observed gains are not driven by increased student presence (a dosage mechanism), but rather by improvements in the learning environment conditional on attendance.

Inputs: As theory suggests, schools have unique needs, so the specific policies they adopt should vary. I assess this empirically using detailed personnel spending data, which—while imperfect—provides useful insights. Although it does not capture shifts between personnel and non-personnel spending, it is informative given that public schools spend roughly 70 percent of their budgets on personnel (Jackson et al., 2016). Moreover, the spending categories are more detailed than in most prior studies, allowing for more granular analysis.

Using personnel data linked to individual schools, I calculate spending by category—principals, teachers, and other staff. The "other" category includes spending on teaching assistants, nurses, social workers, counselors, bilingual teachers, custodial staff, bus staff, lunch staff, and miscellaneous personnel. I estimate the ISP effect on the natural log of spending in each category and report the results in the top panel of Table 9.

First, I examine changes across all ISP schools on average, then I explore differences across ISP schools with potentially different needs. Across all ISP schools, results from the top panel of Table 9 show no meaningful change in overall personnel spending—the coefficient on logged total spending is 0.0182 (p-value>0.1)—suggesting that ISP primarily schools reallocated existing budgets rather than increased overall expenditures. I explore this below.

Spending on principal salaries increases mechanically by about 10 percent as part of the ISP package. However, there is no detectable change, on average, in teacher or non-teaching staff spending. This is somewhat surprising, given that ISP schools experienced a statistically significant reduction of 1.4 students per class (p-value<0.01), as shown in column 7 of Table 8. While this may appear contradictory, the ISP granted principals the autonomy needed to reallocate staff and schedules to achieve smaller classes. For example, instead of assigning specialist teachers (e.g., art, PE, or reading intervention) to pull out students, principals may have reassigned those teachers to core instruction. They might also have shifted funds from contracted services to classroom teachers, adopted greater departmentalization, or implemented creative scheduling to avoid combining grades—all of which could reduce class size without increasing teacher spending.

This reduction in class size is meaningful. Prior research suggests that a 1.4-student decrease per class raises test scores by 1 to 3 percent of a standard deviation. While not sufficient to explain the full ISP effect, smaller classes enabled by flexible resource use could account for between one-tenth and one-third of the total impact.

Consistent with reallocation, breaking down the "other" personnel category reveals notable shifts. On average, ISP schools reduced spending on special education personnel by 6.33 percent (p-value<0.05), increased spending on bilingual education teachers by 10 percent (p-value<0.05), and increased spending on school counselors by 5.68 percent (p-value<0.10). To help make sense of these shifts, I now examine differences for ISPs with different student populations.

To assess whether these reallocations reflect targeted responses to local needs, I interact the ISP indicator with outlier flags for a school's share of bilingual or special education students. While I cannot directly observe pre-ISP spending shortfalls, these flags likely identify schools that were likely under-resourced in key areas. Although this strategy cannot rule out suboptimal allocation, it can provide evidence consistent with targeted spending increases.

The lower panel of Table 9 presents these results. Both types of outlier schools increased spending on "other" personnel, but the effect is statistically significant only for special education outlier schools, which saw a 14 percent increase (p-value<0.01). Looking within the "other" category, schools with high shares of bilingual students increased bilingual education spending by 12.23 percent (0.106 + 0.0223), while schools with high shares of special education students increased special education spending by 15.7 percent (0.241 – 0.084). The bilingual education effect is significant on average but not significantly different from the ISP average. In contrast, the special education effect is not significant on average but is significantly different from the ISP average —suggestive but not conclusive regarding targeted reallocation to meet specific needs.

Focusing on schools with large shares of special education students, the results in columns 8 and 9 show a reduction in spending on teaching assistants and guidance counselors, which, when considered with the increase in special education spending, indicates a shift away from general in-class support toward more specialized instructional services. These patterns are consistent with schools using added flexibility to prioritize inputs better aligned with their student populations. While not conclusive, these findings provide suggestive evidence that ISP schools—especially those with more specialized needs—used autonomy to make allocatively efficient resource decisions.

VI.6 Extrapolating to All Schools

Because the principals (and schools) that opted into greater autonomy likely faced lower costs or anticipated higher benefits, the returns to voluntary adoption—such as through the ISP program or similar initiatives in the UK (e.g., Clark (2009))—may overstate the benefits for the average school. To explore this, I present estimates of both the predicted treatment effect on the treated (ATT) and the predicted average treatment effect (ATE). The difference between these estimates sheds light on the broader external validity of the results.

To estimate the predicted ISP effect for all schools, I use observable measures of principal alignment and outlier status. Specifically, (using the IPS schools) I run a meta-regression of the observed ISP effects on principal ratings, residual passing scores, and the number of outlier categories. I then and use the fitted model to generate predicted ISP effects for all schools. This approach facilitates a like-with-like comparison of predicted effects for both treated and untreated schools.[25] I also explore how the predicted effects vary for early versus late adopters.

[25] I do not compare actual ISP effects for ISP schools to predicted effects for non-ISP schools, as this would not constitute a like-for-like comparison.

Figure 9 presents kernel density plots of the predicted ISP effect for non-ISP schools, alongside the distributions for each ISP cohort. Consistent with the idea that effects may be larger for earlier adopters, the distribution for non-ISP schools lies to the left of those for schools designated in 2016, 2017, 2018, and 2019. The largest gap appears for the 2016 cohort, suggesting that schools that applied for and received ISP designation in that year were stronger candidates than those never designated. Although predicted effects for later cohorts are somewhat smaller, statistical tests do not reject equality across the 2016–2019 cohorts.

The average predicted effect for ISP schools is a 3.67 percentage point increase in passing rates, compared to 3.2 percentage points for non-ISP schools. Even this smaller predicted effect corresponds to a 0.082σ gain—still a meaningful improvement in academic outcomes. One might have expected a larger gap between the ATT and ATE. In fact, nearly 49 percent of untreated schools have a predicted ISP effect equal to or greater than the average effect for treated schools, suggesting that many high-potential schools had not been designated by 2019.

The relative similarity between the ATT and ATE reflects two offsetting forces. On one hand, ISP schools tend to have principals with higher teacher ratings and stronger pre-treatment performance, which would tend to raise the ATT above the ATE. On the other hand, ISP schools were less likely to serve outlier student populations—such as those with high proportions of English learners or students with disabilities—schools that may benefit most from greater autonomy. While the explicit selection of high-performing principals contributed to larger treatment effects among ISP schools, it did not fully target schools with the greatest potential gains.

Predicted ISP effects are weakly positive for most schools in the district, with strong predicted effects for many schools that were not designated ISP. A simple calculation suggests that extending greater autonomy to the top 75 percent of schools with the largest predicted effects—those led by exemplary principals or serving unique student populations—could raise average districtwide passing rates by about 2.5 percentage points (roughly 0.063σ). While gains of this size are not transformative, the cost-effectiveness of the policy makes it attractive. However, this extrapolation assumes that the observed predictors of the ISP effect apply similarly across all schools and does not account for potential general equilibrium effects if autonomy were extended districtwide. As such, these predictions should be interpreted as suggestive rather than definitive.

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 provides evidence of a significant impact of principal autonomy in a large urban district in the United States, with findings that align with results from highly competitive settings. However, I find considerable variation in the effects across schools that cannot be explained by sampling variability alone. This heterogeneity is consistent with the theoretical framework, which predicts larger gains for schools with high-quality principals and student populations requiring atypical policy responses. The ISP also promoted greater stability, as evidenced by reduced principal turnover and improved school climate—suggesting that enhanced stability contributed meaningfully to the program's success. An analysis of school personnel spending reveals patterns consistent with schools using increased autonomy to (a) reallocate resources to reduce class sizes and (b) tailor spending to meet the specific needs of their student populations—consistent with predictions from canonical work in public finance (e.g., Oates (1999)).

The observed patterns highlight the benefits of increased school autonomy, particularly in heterogeneous settings (Acemoglu et al., 2007). However, they also suggest that autonomy should be granted selectively to effective and motivated school leaders, as it may lead to worse outcomes in contexts with agency problems or low principal capacity—consistent with descriptive comparative evidence across countries (Fuchs and Wößmann, 2007) and U.S. states (Loeb and Strunk, 2007). These results underscore the importance of leadership and management quality (Bloom et al., 2015; Branch et al., 2012), and reinforce the need to account for local context when evaluating policy effects (Jackson and Mackevicius, 2024). Finally, the findings suggest that granting greater autonomy to high-quality school leaders—particularly in schools with unique needs—can substantially improve student outcomes at minimal cost.

References

Abdulkadiroğlu, A., J. Angrist, S. Dynarski, T. J. Kane, and P. Pathak (2011, 5). Accountability and flexibility in public schools: Evidence from boston's charters and pilots. *The Quarterly Journal of Economics 126*, 699–748.

Accelerate (2024, 11). Quarterly research note: Research studies and program design characteristics.

Acemoglu, D., P. Aghion, C. Lelarge, J. V. Reenen, and F. Zilibotti (2007, 11). Technology, information, and the decentralization of the firm. *The Quarterly Journal of Economics 122*, 1759–1799.

Aghion, P. and J. Tirole (1997). Formal and real authority in organizations. *Journal of Political Economy 105*(1), 1–29.

Berkey, C. S., D. C. Hoaglin, F. Mosteller, and G. A. Colditz (1995, 2). A random-effects regression model for meta-analysis. *Statistics in Medicine 14*, 395–411.

Bertrand, M. and A. Schoar (2003, 11). Managing with style: The effect of managers on firm policies. *The Quarterly Journal of Economics 118*, 1169–1208.

Beuermann, D. W. and C. K. Jackson (2022, 5). The short- and long-run effects of attending the schools that parents prefer. *Journal of Human Resources 57*, 725–746.

Beuermann, D. W., C. K. Jackson, L. Navarro-Sola, and F. Pardo (2023, 1). What is a good school, and can parents tell? evidence on the multidimensionality of school output. *The Review of Economic Studies 90*, 65–101.

Biasi, B., J. Lafortune, and D. Schönholzer (2024). What works and for whom? effectiveness and efficiency of school capital investments across the u.s. *Social Science Research Network*.

Bishop, J., L. Wossmann, J. Bishop, and L. Wossmann (2004, 4). Institutional effects in a simple model of educational production. *Education Economics 12*, 17–38.

Bloom, N., R. Lemos, R. Sadun, and J. V. Reenen (2015, 5). Does management matter in schools? *The Economic Journal 125*, 647–674.

Branch, G. F., E. A. Hanushek, S. G. Rivkin, and T. Schools (2012, 2). Estimating the effect of leaders on public sector productivity: The case of school principals.

Béteille, T., D. Kalogrides, and S. Loeb (2012). Stepping stones: Principal career paths and school outcomes. *Social Science Research*.

Caetano, C. and B. Callaway (2023). Difference-in-differences with time-varying covariates in the parallel trends assumption. *Papers*.

Callaway, B. and P. H. Sant'Anna (2021, 12). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*, 200–230.

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

Carroll, R. J. and P. Hall (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association 83*, 1184–1186.

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer (2019, 8). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics 134*, 1405–1454.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014, September). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review 104*(9), 2593–2632.

Chiang, H., S. Lipscomb, and B. Gill (2016, 7). Is school value added indicative of principal quality? *Education Finance and Policy 11*, 283–309.

Clark, D. (2009, 8). The performance and competitive effects of school autonomy. *Journal of Political Economy 117*, 745–782.

Cohodes, S. R. and K. S. Parham (2021, 2). Charter schools' effectiveness, mechanisms, and competitive influence.

Cohodes, S. R., E. M. Setren, and C. R. Walters (2021, 2). Can successful schools replicate? scaling up boston039;s charter school sector. *American Economic Journal: Economic Policy 13*, 138–67.

Colombo, M. G. and M. Delmastro (2004, 3). Delegation of authority in business organizations: An empirical test. *Organizational Behavior 52*, 53–80.

CPEF (2015). Chicago's fight to keep top principals. Technical report, Chicago Public Education Fund.

CPEF (2017). School leadership in chicago: A baseline report. Technical report, Chicago Public Education Fund.

Deci, E. L. and R. M. Ryan (1985). Intrinsic motivation and self-determination in human behavior. *Intrinsic Motivation and Self-Determination in Human Behavior*.

DerSimonian, R. and N. Laird (1986, September). Meta-analysis in clinical trials. *Controlled Clinical Trials 7*(3), 177–188.

Deshpande, M. and Y. Li (2019, 11). Who is screened out? application costs and the targeting of disability programs. *American Economic Journal: Economic Policy 11*, 213–48.

Donaldson, M. L., M. Mavrogordato, P. Youngs, S. Dougherty, and R. A. Ghanem (2021, 1). "doing the 'real' work": How superintendents' sensemaking shapes principal evaluation policies and practices in school districts. *https://doi.org/10.1177/2332858420986177 7*.

Efron, B. and C. Morris (1973, March). Stein's Estimation Rule and its Competitors—An Empirical Bayes Approach. *Journal of the American Statistical Association 68*(341), 117–130.

Efron, B. and R. Tibshirani (1996, 12). Using specially designed exponential families for density estimation. *https://doi.org/10.1214/aos/1032181161 24*, 2431–2461.

Eyles, A. and S. Machin (2019, 8). The introduction of academy schools to england's education. *Journal of the European Economic Association 17*, 1107–1146.

Eyles, A., S. Machin, and S. McNally (2017, 11). Unexpected school reform: Academisation of primary schools in england. *Journal of Public Economics 155*, 108–121.

Frick, B., R. Simmons, B. Frick, and R. Simmons (2007). The impact of managerial quality on organizational performance: Evidence from german soccer.

Fuchs, T. and L. Wößmann (2007, 5). What accounts for international differences in student performance? a re-examination using pisa data. *Empirical Economics 32*, 433–464.

Galiani, S., P. Gertler, and E. Schargrodsky (2008, 10). School decentralization: Helping the good get better, but leaving the poor behind. *Journal of Public Economics 92*, 2106–2120.

Grissom, J. A., R. S. Blissett, and H. Mitani (2018, 6). Evaluating school principals: Supervisor ratings of principal practice and principal job performance. *https://doi.org/10.3102/0162373718783883 40*, 446–472.

Grissom, J. A., D. Kalogrides, and S. Loeb (2012). Using student test scores to measure principal performance. *Educational Evaluation and Policy Analysis 37*, 3–28.

Grossman, S. J. and O. D. Hart (1986, 8). The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy 94*, 691–719.

Henry, G. T. and E. Harbatkin (2019). Turnover at the top: Estimating the effects of principal turnover on student, teacher, and school outcomes. EdWorkingPaper 19-95, Annenberg Institute at Brown University.

Ho, A. D. (2009, 6). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics 34*, 201–228.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.

Jackson, C. K. (2018, 5). What do test scores miss? the importance of teacher effects on non-test score outcomes. *Journal of Political Economy*.

Jackson, C. K. and E. Bruegmann (2009, 10). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics 1*, 85–108.

Jackson, C. K., R. C. Johnson, and C. Persico (2016, 2). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *The Quarterly Journal of Economics 131*, 157–218.

Jackson, C. K., S. Kiguel, S. C. Porter, and J. Q. Easton (2023, 2). Who benefits from attending effective high schools? *Journal of Labor Economics*.

Jackson, C. K. and C. Mackevicius (2024). What impacts can we expect from school spending policy? evidence from evaluations in the u.s. *American Economic Journal: Applied Economics*, p.43.

Jackson, C. K., C. Persico, K. Kelly, and P. Decker (2023, 9). Point column on school spending: Money matters. *Journal of Policy Analysis and Management 42*, 1118–1124.

Jackson, C. K., S. C. Porter, J. Q. Easton, A. Blanchard, and S. Kiguel (2020, 12). School effects on socioemotional development, school-based arrests, and educational attainment. *American Economic Review: Insights 2*, 491–508.

Jackson, C. K., J. E. Rockoff, and D. O. Staiger (2014). Teacher effects and teacher-related policies. *Annual Review of Economics 6*(1), 801–825.

Jacob, B. A. and L. Lefgren (2008, 1). Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of Labor Economics 26*, 101–136.

Jones, B. F. and B. A. Olken (2005, 8). Do leaders matter? national leadership and growth since world war ii. *The Quarterly Journal of Economics 120*, 835–864.

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

Kato, K. and Y. Sasaki (2018, 11). Uniform confidence bands in deconvolution with unknown error distribution. *Journal of Econometrics 207*, 129–161.

Kling, J. R., J. B. Liebman, and L. F. Katz (2007, 1). Experimental analysis of neighborhood effects. *Econometrica 75*, 83–119.

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

Kraft, M. A., D. Blazar, and D. Hogan (2018, 8). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research 88*, 547–588.

Laing, D., S. G. Rivkin, J. C. Schiman, and J. Ward (2016). Decentralized governance and the quality of school leadership. *NBER Working Papers*.

Liebman, J. B., N. Mahoney, L. :, and J. F. Kennedy (2017, 11). Do expiring budgets lead to wasteful year-end spending? evidence from federal procurement. *American Economic Review 107*, 3510–49.

Liu, K., D. Stuit, J. Springer, J. Lindsay, and Y. Wan (2014, 11). The utility of teacher and student surveys in principal evaluations: An empirical investigation — american institutes for research. *AIR Report*.

Loeb, S. and K. Strunk (2007). Accountability and local control: Response to incentives with and without authority over resource generation and allocation. *Source: Education Finance and Policy 2*, 10–39.

Merkle, J. (2022). School-level autonomy and its impact on student achievement. *Peabody Journal of Education 97*, 497–519.

Miller, A. (2013, 10). Principal turnover and student achievement. *Economics of Education Review 36*, 60–72.

Morris, C. N. (1983, March). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association 78*(381), 47–55.

Muecke, S. and A. Iseke (2019, 8). How does job autonomy influence job performance? a meta-analytic test of theoretical mechanisms. *https://doi.org/10.5465/AMBPP.2019.145*.

Neal, D. and D. W. Schanzenbach (2010, 5). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics 92*, 263–283.

Neri, L. and E. Pasini (2023, 6). Heterogeneous effects of school autonomy in england. *Economics of Education Review 94*, 102366.

Neri, L., E. Pasini, and O. Silva (2022). The organizational economics of school chains. *IZA Discussion Papers*.

Oates, W. E. (1999). An essay on fiscal federalism. *Journal of Economic Literature 37*, 1120–1149.

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

Regan-Stansfield, J. (2018, 4). Does greater primary school autonomy improve pupil attainment? evidence from primary school converter academies in england. *Economics of Education Review 63*, 167–179.

Sartain, A. L. Smith, C. G. B. M. (2023). New principals in chicago public schools: Diversity and their prior experiences. Technical report, NORC at the University of Chicago.

Steinberg, M. P. (2014, 1). Does greater autonomy improve school performance? evidence from a regression discontinuity analysis in chicago. *Education Finance and Policy 9*, 1–35.

Steinberg, M. P. and A. B. Cox (2016, 1). 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.

Stiefel, L., A. E. Schwartz, C. Portas, and D. Y. Kim (2003). School budgeting and school performance: The impact of new york city's performance driven budgeting initiative. *Journal of Education Finance 28*, 403–24.

Travlos, J. (2020). A phenomenological study of chicago's independent school a phenomenological study of chicago's independent school principals principals.

Tuchman, S., B. Gross, and L. Chu (2022, 8). Weighted student funding and outcomes: Implementation in 18 school districts. *Peabody Journal of Education 97*, 479–496.

van den Ham, A. K. and A. Heinze (2018, 12). Does the textbook matter? longitudinal effects of textbook choice on primary school students' achievement in mathematics. *Studies in Educational Evaluation 59*, 133–140.

Wallis, J. J. and W. E. Oates (1988). Decentralization in the public sector: An empirical study of state and local government. *NBER Chapters*, 5–32.

Walters, P. K. K. E. R. C. R. (2022, 9). Systemic discrimination among large u.s. employers. *The Quarterly Journal of Economics 137*, 1963–2036.

Wang, C.-C. and W.-C. Lee (2020, March). Evaluation of the Normality Assumption in Meta-Analyses. *American Journal of Epidemiology 189*(3), 235–242.

Wang, X.-F. and B. Wang (2011, March). Deconvolution Estimation in Measurement Error Models: The R Package decon. *Journal of Statistical Software 39*(10), i10.

Weinstein, M., A. E. Schwartz, R. Jacobowitz, T. Ely, and K. Landon (2009). New schools, new leaders: A study of principal turnover and academic achievement at new high schools in new york city. Condition Report 2011–09, Education Finance Research Consortium.

Wong, L. S., C. E. Coburn, and A. Kamel (2020, 8). 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.

Tables and Figures

Table 1: Descriptive Statistics for ISP and Non-ISP Schools: Full Panel and Matched Sample

	(1) **Panel Non-ISP**	(2) **Panel ISP**	(3) ***p*-value (Panel)**	(4) **Matched Non-ISP**	(5) ***p*-value (Matched)**
Percent White	0.06	0.15	0.000	0.17	0.538
	(0.14)	(0.20)		(0.21)	
Percent Black	0.57	0.25	0.000	0.25	0.845
	(0.42)	(0.31)		(0.35)	
Percent Hispanic	0.33	0.52	0.000	0.51	0.764
	(0.36)	(0.35)		(0.34)	
Percent Special Ed.	0.16	0.13	0.008	0.13	0.830
	(0.11)	(0.11)		(0.09)	
Percent Bilingual	0.15	0.22	0.000	0.23	0.636
	(0.17)	(0.19)		(0.18)	
Enrollment	689.27	741.00	0.277	767.33	0.666
	(745.18)	(430.95)		(496.58)	
Attendance Rate	0.93	0.95	0.000	0.95	0.852
	(0.05)	(0.02)		(0.02)	
Mobility Rate	0.18	0.11	0.000	0.12	0.308
	(0.14)	(0.09)		(0.08)	
Chronic Truancy Rate	0.32	0.20	0.000	0.19	0.730
	(0.24)	(0.17)		(0.17)	
Percent Free or Reduced-Price Lunch	0.85	0.76	0.001	0.76	0.871
	(0.20)	(0.26)		(0.26)	
ELA Passing Rate (%)	31.92	43.99	0.000	42.79	0.625
	(22.15)	(23.44)		(22.97)	
Math Passing Rate (%)	30.84	43.52	0.000	42.47	0.668
	(24.85)	(25.42)		(25.58)	
ELA Percentile (NWEA)	49.50	71.63	0.000	67.63	0.166
	(27.13)	(20.98)		(21.79)	
Math Percentile (NWEA)	46.02	72.27	0.000	65.99	0.036
	(27.70)	(21.79)		(23.05)	
Instructional Leadership Score	56.93	63.03	0.000	59.16	0.023
	(20.58)	(16.68)		(18.15)	
FiveEssentials Score	3.81	4.42	0.000	4.22	0.015
	(1.34)	(0.92)		(1.06)	

Notes: Means and standard deviations (in parentheses) are shown for both the full panel and matched sample. *p*-values test the null hypothesis of equality of means between ISP and non-ISP schools.

Table 2: Estimated Effects of ISP Treatment on Test Scores and Passing Rates

	(1) **ELA Passing Rate**	(2) **Math Passing Rate**	(3) **ELA Percentile (2014–2019)**	(4) **Math Percentile (2014–2019)**	(5) **Predicted Math Passing Rate**	(6) **Predicted ELA Passing Rate**
ISP Year 1	3.009***	2.198***	0.864	2.455*	0.229	0.196
	[0.817]	[0.829]	[1.012]	[1.320]	[0.355]	[0.354]
ISP Year 2	3.827***	3.010***	1.299	3.115*	0.860*	0.946*
	[0.952]	[0.988]	[1.242]	[1.844]	[0.472]	[0.482]
ISP Year 3+	7.079***	5.410***	2.516**	4.857***	0.769	0.697
	[1.528]	[1.515]	[1.181]	[1.825]	[0.576]	[0.559]
ISP (Pooled)	4.144***	3.141***	1.302	3.099**	0.537	0.531
	[0.835]	[0.827]	[0.972]	[1.390]	[0.356]	[0.357]
Observations	6,470	6,468	3,895	3,895	6,737	6,737

Notes: Robust standard errors are shown 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 *designation*, while the lower panel reports the simple before versus after comparison.

Passing rate outcomes are based on state standardized tests. Percentile score outcomes are based on district benchmark assessments.

Table 3: Estimated Treatment Effects on Other Characteristics by Year

	(1) Percent White	(2) Percent Black	(3) Percent Hispanic	(4) Percent Special Ed.	(5) Percent Bilingual	(6) Mobility Rate	(7) Percent Free Lunch	(8) Enrollment
ISP Year 1	0.0060	-0.0013	-0.0067	-0.0018	-0.0117**	0.0018	-0.0020	20.25
	[0.0044]	[0.0039]	[0.0053]	[0.0025]	[0.0058]	[0.0038]	[0.0060]	[16.90]
ISP Year 2	0.0085*	-0.0042	-0.0044	-0.0052	-0.0135*	0.0048	-0.0206***	38.25*
	[0.0047]	[0.0052]	[0.0064]	[0.0035]	[0.0072]	[0.0042]	[0.0078]	[22.85]
ISP Year 3+	0.0076	-0.0106	-0.0000	-0.0030	0.0002	0.0047	-0.0112	34.28
	[0.0081]	[0.0071]	[0.0069]	[0.0040]	[0.0076]	[0.0047]	[0.0103]	[35.55]
ISP (Pooled)	0.0071	-0.0043	-0.0045	-0.0031	-0.0098*	0.0033	-0.0095	28.69
	[0.0047]	[0.0043]	[0.0052]	[0.0026]	[0.0052]	[0.0034]	[0.0058]	[21.08]
Observations	6,703	6,703	6,703	5,239	5,009	6,463	6,399	6,501

Notes: Robust standard errors are shown in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Each column reports estimated coefficients for the indicated outcome variable. 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. All models are based on the stacked-matched sample.

Table 4: Estimated ISP Effects on Test Outcomes (One-to-One Matching)

	(1) ELA Passing Rate	(2) Math Passing Rate	(3) ELA Percentile (2014–2019)	(4) Math Percentile (2014–2019)	(5) Predicted Math Passing Rate	(6) Predicted ELA Passing Rate
ISP Year 1	3.682***	3.541***	1.039	4.182**	-0.126	-0.192
	[1.048]	[1.108]	[1.345]	[1.717]	[0.572]	[0.590]
ISP Year 2	4.434***	3.945***	2.133	5.340**	0.205	0.517
	[1.109]	[1.454]	[1.754]	[2.342]	[1.097]	[1.114]
ISP Year 3+	8.477***	6.051***	3.694**	6.469***	0.027	0.232
	[1.947]	[2.200]	[1.816]	[2.425]	[1.028]	[1.006]
ISP (Pooled)	4.976***	4.221***	1.858	4.947***	0.006	0.113
	[1.022]	[1.193]	[1.257]	[1.714]	[0.675]	[0.681]
Observations	1,406	1,402	792	792	1,500	1,500

Notes: Robust standard errors are shown in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Each column shows estimated ISP effects on state test passing rates, district percentile scores, and predicted outcomes for ELA and Math using one-to-one matched samples. The top panel reports dynamic treatment effects by year, and the bottom panel reports average (pooled) effects across years.

Table 5: Estimated ISP Effects on NWEA RIT Scores by Grade Level and Subject

	(1) **Elementary Math**	(2) **Middle Math**	(3) **Elementary Reading**	(4) **Middle Reading**	(5) **Elementary Math Difference**	(6) **Elementary Reading Difference**
ISP Year 1	0.955***	0.330	0.451*	0.151	0.399	0.198
	[0.333]	[0.536]	[0.259]	[0.302]	[0.330]	[0.263]
ISP Year 2	1.096**	0.351	0.417	0.171	0.388	0.362
	[0.538]	[0.744]	[0.383]	[0.402]	[0.472]	[0.256]
ISP Year 3+	1.048**	1.024	0.446	0.493	-0.142	0.512
	[0.526]	[0.904]	[0.418]	[0.479]	[0.700]	[0.433]
ISP (Pooled)	1.014***	0.474	0.440*	0.225	0.337	0.284
	[0.354]	[0.564]	[0.262]	[0.298]	[0.276]	[0.212]
Observations	4,564	4,216	4,564	4,216	3,906	3,906

Notes: Robust standard errors are shown in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Each cell reports the estimated effect of ISP *designation* on average NWEA RIT scores by grade level and subject. The top panel reports dynamic treatment effects by year, and the bottom panel reports average (pooled) effects across years.

Table 6: Precision-Weighted Average ISP Effects on Passing Rates

	(1) **ELA** **Passing Rate**	(2) **Math** **Passing Rate**	(3) **Both Subjects** **Passing Rate**
Precision-Weighted Average ISP Effect	4.259***	3.275***	3.764***
	[0.769]	[0.748]	[0.706]
Observations	79	79	158
τ **(Between-School SD)**	6.378	5.914	6.304

Notes: Robust standard errors are shown in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Each column shows the estimated precision-weighted average effect of ISP *designation* on state test passing rates for ELA, Math, and both subjects combined. The bottom panel reports the number of schools and the estimated standard deviation (τ) of true effects across schools, following DerSimonian and Laird (1986).

Table 7: Heterogeneity of ISP Effects by Principal Quality and Outlier Status

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	ISP Effect on Both Subjects (Passing Rate)						
	Random Effects Meta-Regression					**OLS**	**Precision-Weighted**
Principal Quality (Index)	1.868***	—	—	—	1.751***	1.624***	2.034***
	[0.558]				[0.572]	[0.571]	[0.666]
Leadership Score	—	0.0737*	—	—	—	—	—
		[0.0409]					
Residual Scores	—	—	1.374***	—	—	—	—
			[0.425]				
Outlier	—	—	—	2.868**	2.567*	2.216*	3.018**
				[1.371]	[1.326]	[1.326]	[1.408]
Observations	158	158	158	158	158	158	158

Notes: Robust standard errors are shown in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Each column reports the estimated effect of ISP *designation* on combined ELA and Math passing rates, based on alternative measures of principal alignment and outlier status. Leadership scores and residual scores are measured the year prior to ISP designation. Outlier status is based on the extent to which a school's 2015 student demographics differed from district averages.

Table 8: Estimated ISP Effects on Principal Turnover, School Climate, and Staff Outcomes

	(1) **New Principal**	(2) **Principal Turnover**	(3) **FiveEssentials Score**	(4) **Attendance Rate**	(5) **Chronic Truant Rate**	(6) **Teacher Retention Rate**	(7) **Average Class Size**
ISP Year 1	-0.0357***	-0.384***	0.240**	0.00014	-0.00566	0.00122	-1.798***
	[0.0104]	[0.0843]	[0.104]	[0.000890]	[0.0104]	[0.00630]	[0.608]
ISP Year 2	-0.0302***	-0.448***	0.318**	0.00001	-0.00549	0.00171	-1.336*
	[0.0117]	[0.121]	[0.123]	[0.00133]	[0.0124]	[0.00811]	[0.722]
ISP Year 3+	-0.0324**	-0.495***	0.357**	0.00104	-0.0125	0.0113	-0.555
	[0.0161]	[0.176]	[0.155]	[0.00150]	[0.0126]	[0.0140]	[0.814]
ISP (Pooled)	-0.0333***	-0.426***	0.285***	0.00031	-0.00716	0.0035	-1.413***
	[0.0107]	[0.101]	[0.0916]	[0.000955]	[0.0100]	[0.00700]	[0.473]
Observations	6,737	5,041	4,375	6,481	6,272	5,035	4,869

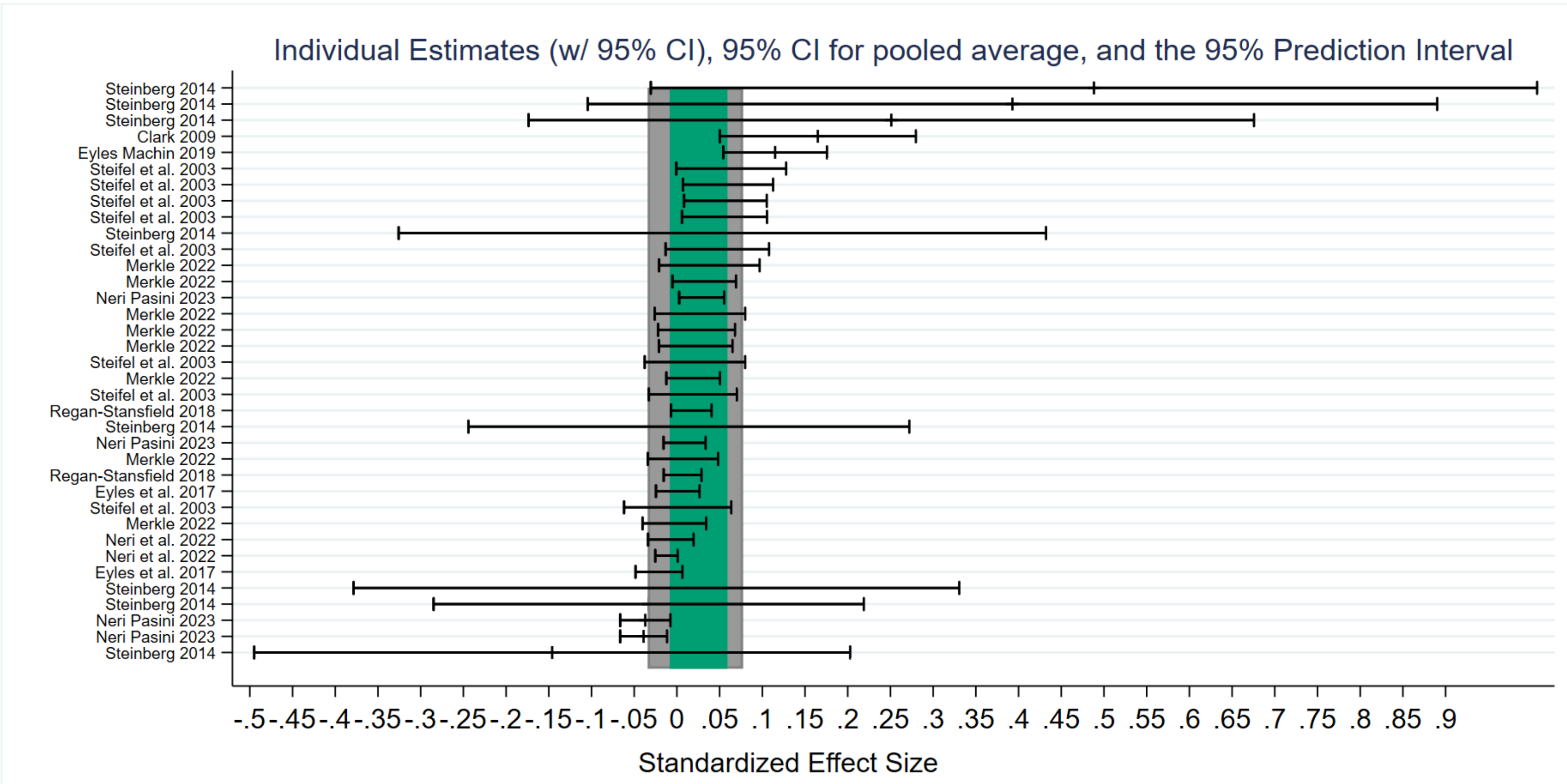
Notes: Robust standard errors are shown in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Each column reports the estimated effect of ISP *designation* on principal turnover, school climate (FiveEssentials), student attendance, chronic truancy, teacher retention, and class size. The top panel reports the dynamic treatment effects for the first three years after ISP designation, while the bottom panel reports the pooled pre–post effect. Estimates are based on a stacked sample using a difference-in-differences framework.

Table 9: Estimated ISP Effects on Log Personnel Expenditures by Staff Category

	(1) **Total Personnel**	(2) **Teacher Salaries**	(3) **Principal Salaries**	(4) **Other Staff Salaries**	(5) **Special Education Teacher Salaries**	(6) **Bilingual Teacher Salaries**	(7) **School Security Salaries**	(8) **Teacher Assistant Salaries**	(9) **School Counselor Salaries**	(10) **School Clerk Salaries**
Panel A: Main Effect										
ISP (Pooled)	0.0182	0.0117	0.101***	0.0170	-0.0633**	0.104**	0.0163	0.0653	0.0568*	0.0425
	[0.0147]	[0.0165]	[0.0255]	[0.0202]	[0.0276]	[0.0471]	[0.0294]	[0.0714]	[0.0290]	[0.0332]
Panel B: Heterogeneous Effects										
ISP (Pooled)	0.00699	0.00401	0.0794***	-0.00366	-0.0840***	0.106*	0.00200	0.101	0.0527	0.0191
	[0.0175]	[0.0199]	[0.0287]	[0.0224]	[0.0293]	[0.0572]	[0.0313]	[0.0785]	[0.0324]	[0.0354]
ISP × Outlier in % Special Education	0.0916**	0.0580	0.206**	0.143***	0.241*	-0.119	0.112	-0.399**	-0.211*	0.138
	[0.0423]	[0.0604]	[0.0830]	[0.0327]	[0.138]	[0.102]	[0.122]	[0.198]	[0.120]	[0.135]
ISP × Outlier in % Bilingual	0.0328	0.0236	0.0580	0.0668	0.0442	0.0229	0.0412	-0.0914	0.0661	0.0815
	[0.0303]	[0.0336]	[0.0460]	[0.0434]	[0.0691]	[0.0703]	[0.0675]	[0.146]	[0.0609]	[0.0745]
Observations	4,331	4,328	4,331	4,331	4,318	3,195	3,615	2,503	4,319	4,331

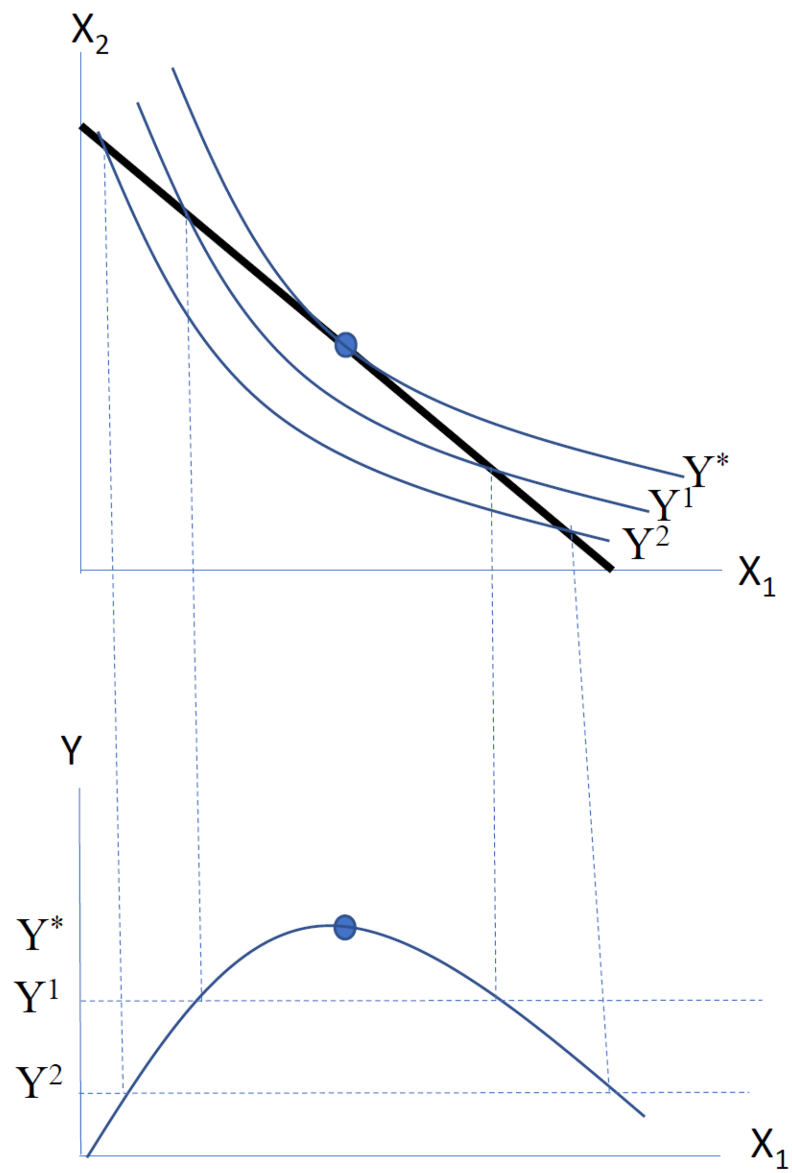
Notes: Robust standard errors are shown in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Panel A reports the estimated average effect of ISP designation on the natural log of personnel spending by staff category. Panel B adds interactions for schools with high shares of special education and bilingual students. All models use a difference-in-differences framework on the stacked sample.

Figure 1. *Summary of Design-Based Studies*



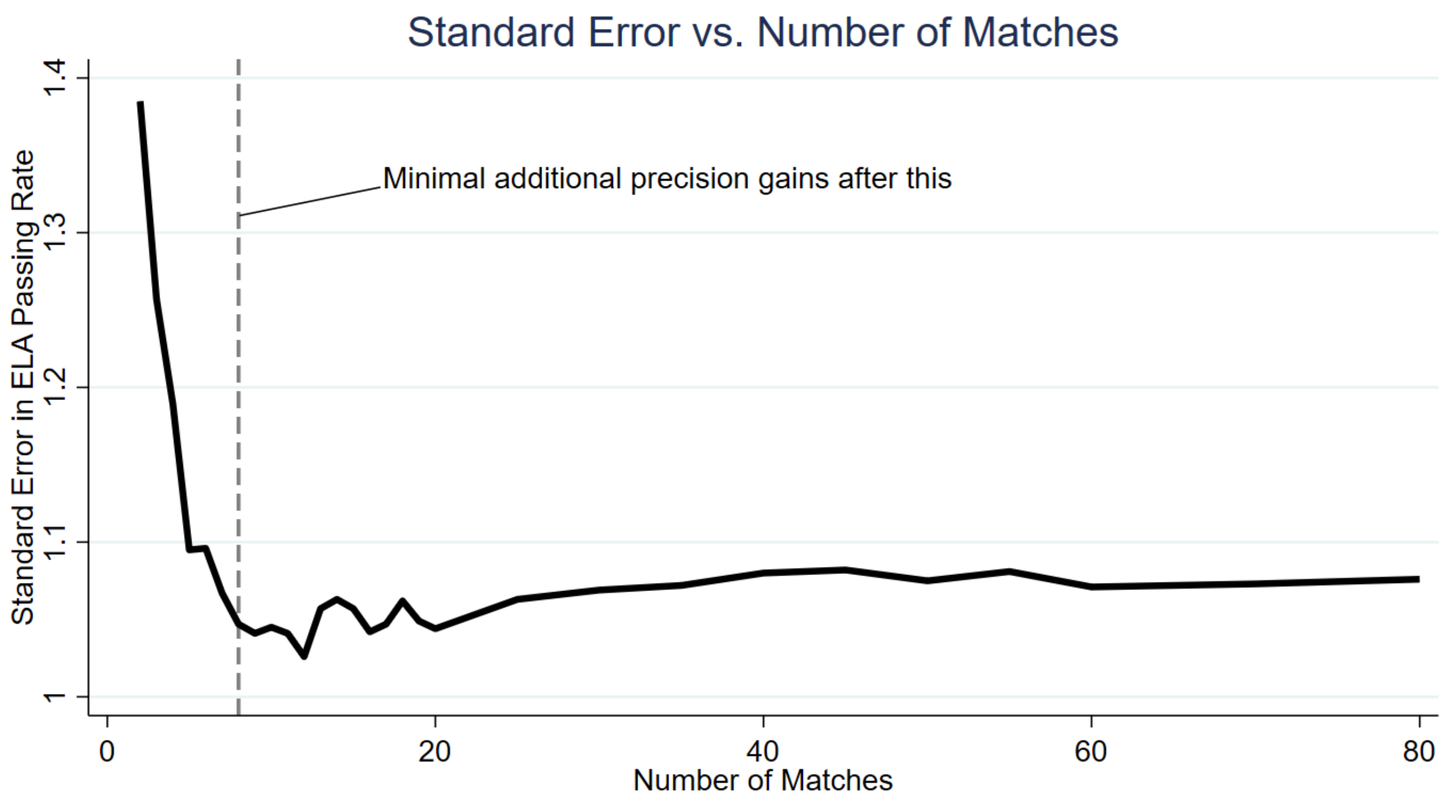
Notes: This forest plot displays the estimates from all design-based studies discussed in Section II, each shown with its corresponding 95% confidence interval. The green shaded area represents the 95% credibility interval for the pooled average effect from the Bayesian meta-analysis. The grey shaded area indicates the 95% prediction interval, reflecting the expected range of true treatment effects across settings.

Figure 2. *Sketch of Theory*



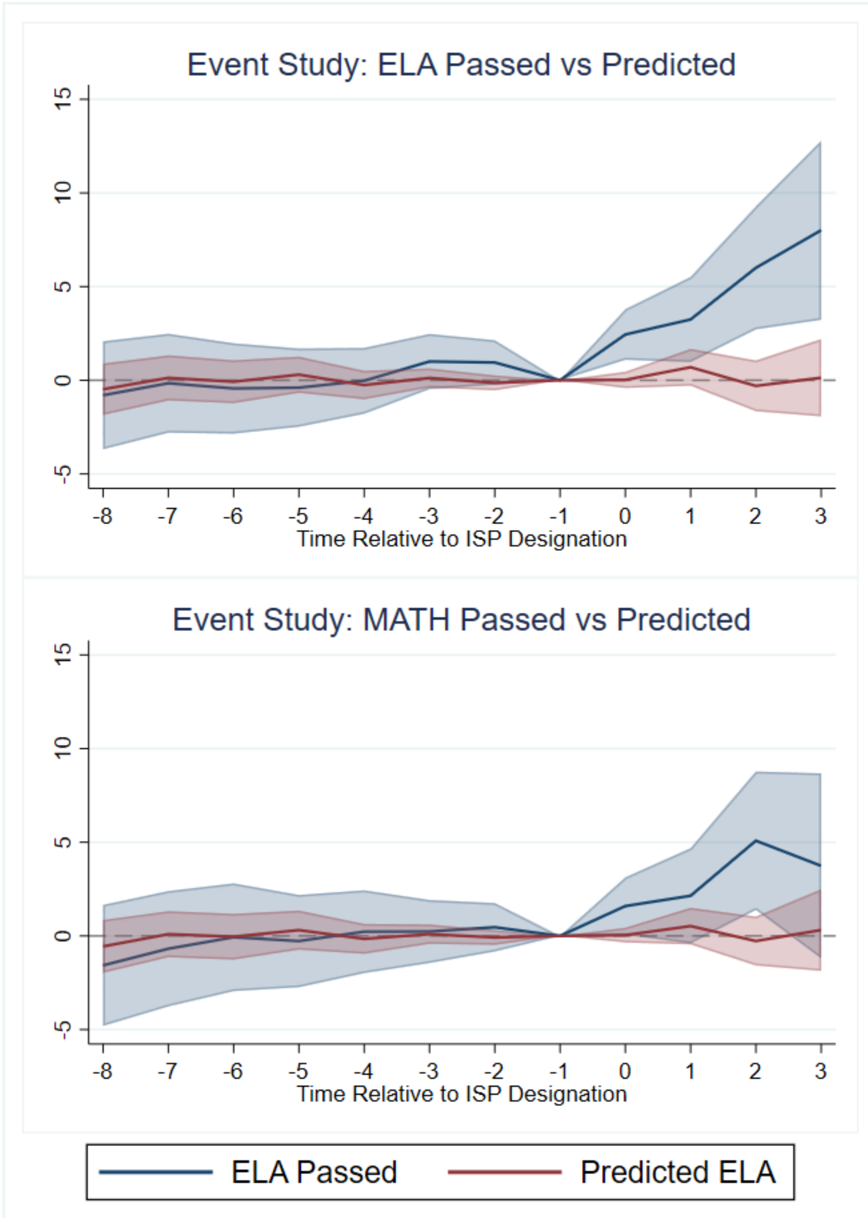
Notes: With any smooth twice-differentiable concave production function, the production with respect to any single output (spending all the budget) will be inverse U-shaped, with a maximum at the output maximizing level of input 1.

Figure 3. *Precision and Number of Matches*



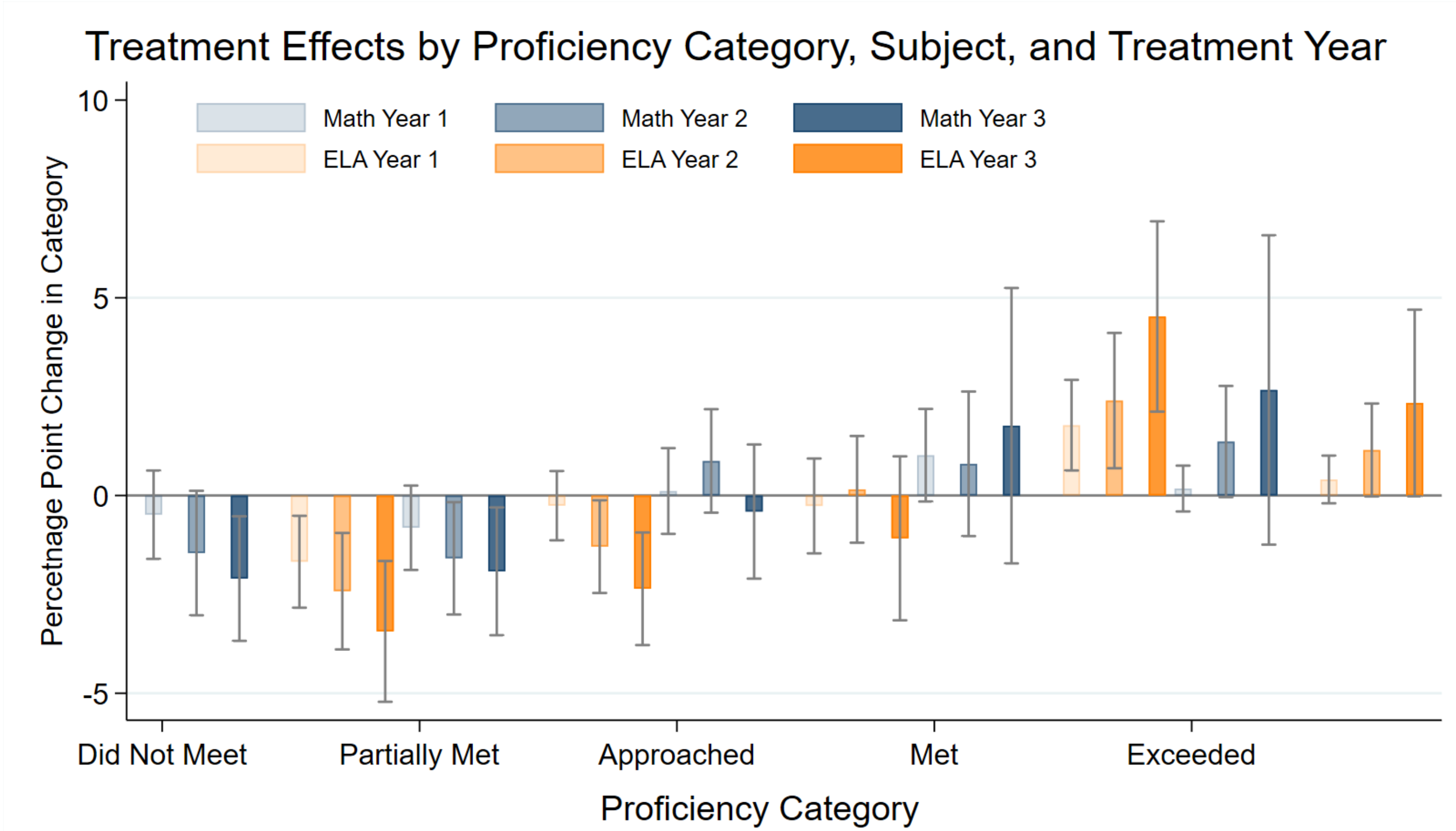
Notes: This figure plots the estimated standard error of the ISP effect on math passing rates against the number of matches used in the analysis. The standard error serves as a measure of statistical precision, allowing the data to inform the optimal number of matched control schools.

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



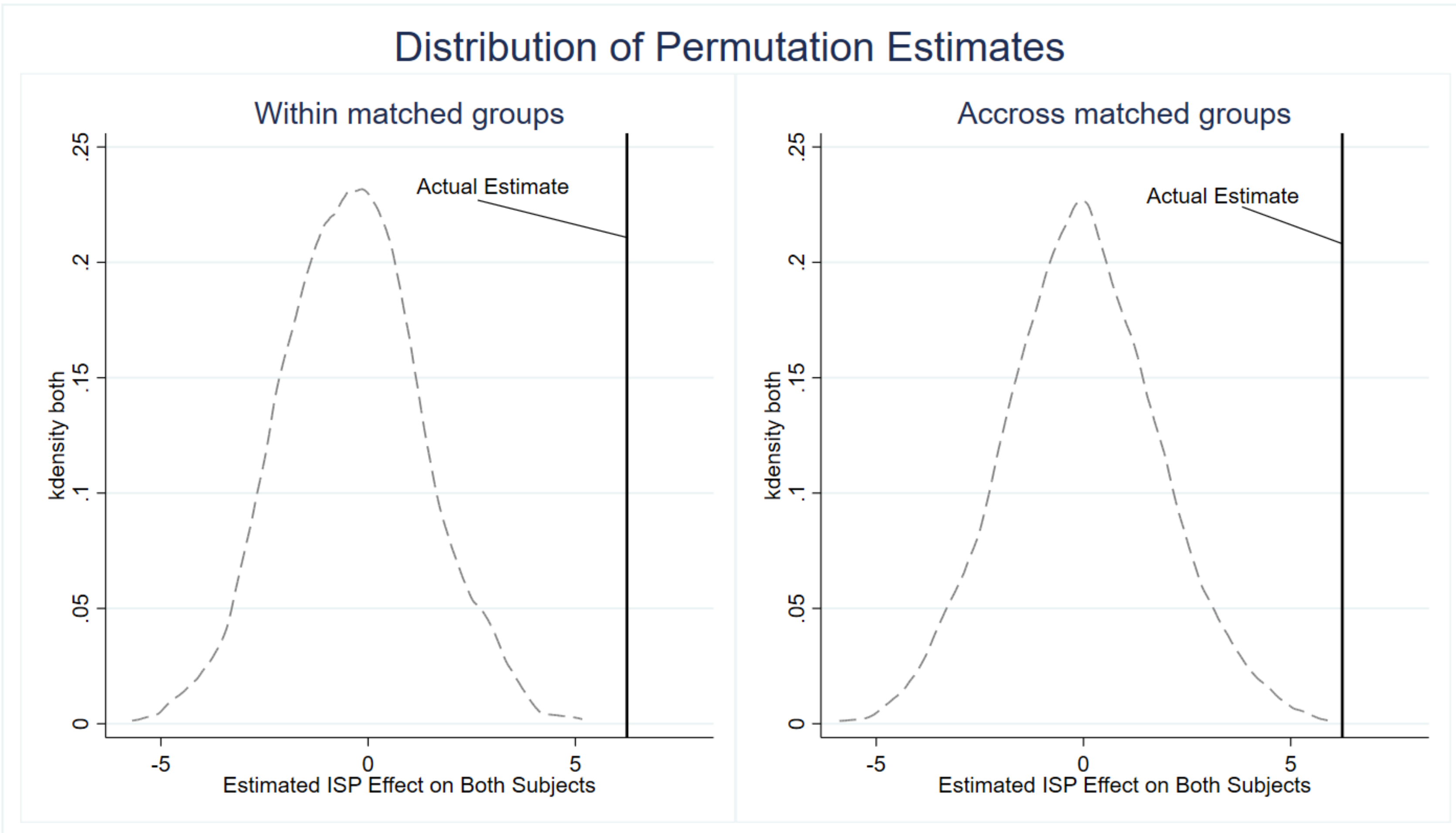
Notes: These event-study estimates are based on stacked difference-in-differences models applied to the matched sample, following the approach in Cengiz et al. (2019) and Deshpande and Li (2019). No additional covariates are included. The figure shows two event-study series: actual passing rates (in blue) and predicted passing rates (in red), where predicted outcomes are based on an outcome-weighted average of all observable predictors of school passing rates. Year 0 denotes the first year after ISP designation—i.e., the earliest year in which treatment effects may appear.

Figure 5. *Effect on Different Margins*



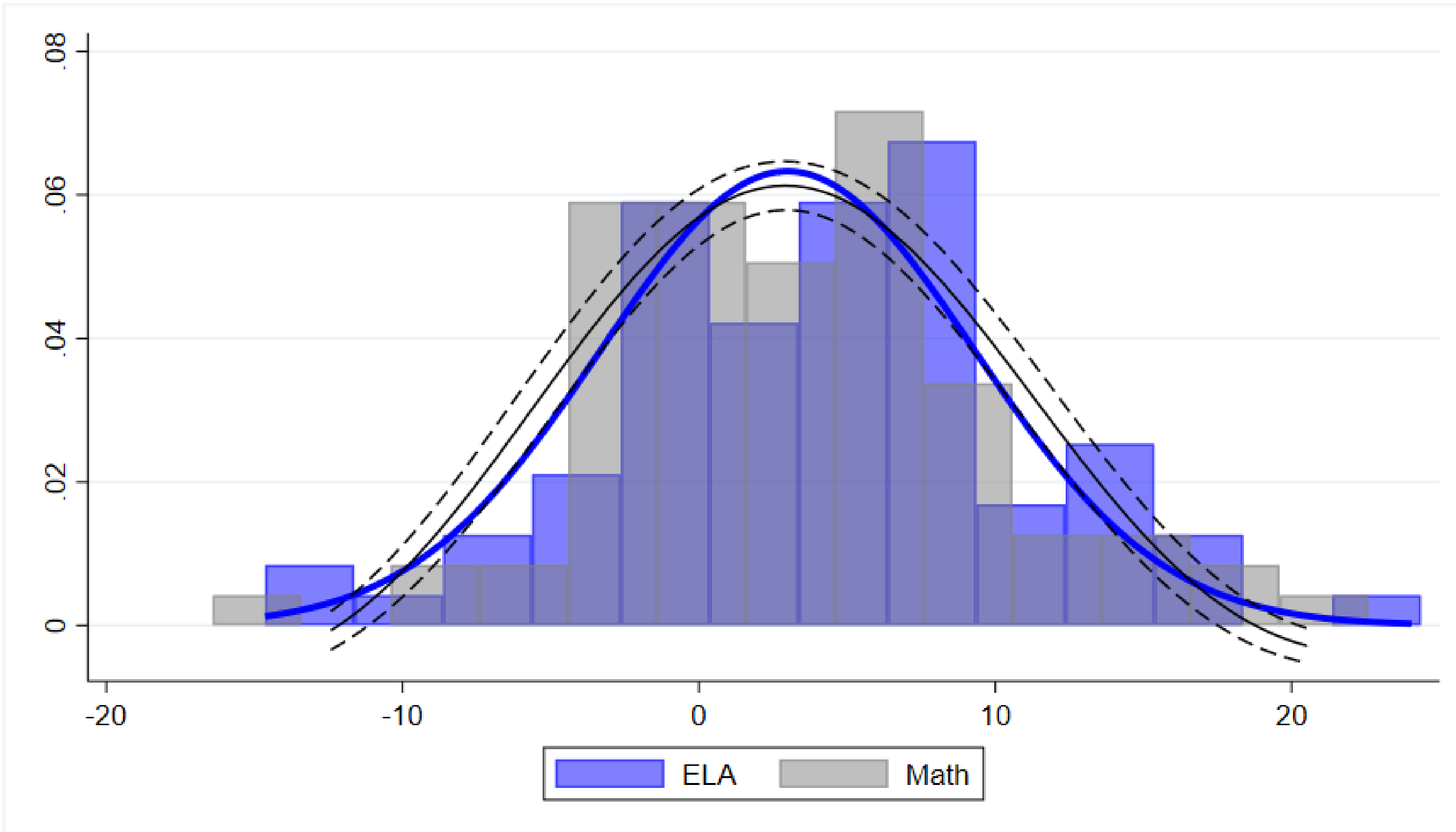
Notes: This figure displays the estimated effects of ISP designation on the percentage of students scoring at each of the five PARCC performance levels: below, partially meeting, approaching, meeting, and exceeding grade-level standards. Separate estimates are shown for ISP Year 1, Year 2, and Year 3+, allowing for an examination of how impacts evolve over time. Math results are shown in blue and ELA results in orange, with darker shades representing later treatment years. Each bar represents the point estimate, and error bars denote the 95% confidence interval for that estimate.

Figure 6. *Distribution of Estimates Under Permutation Test*



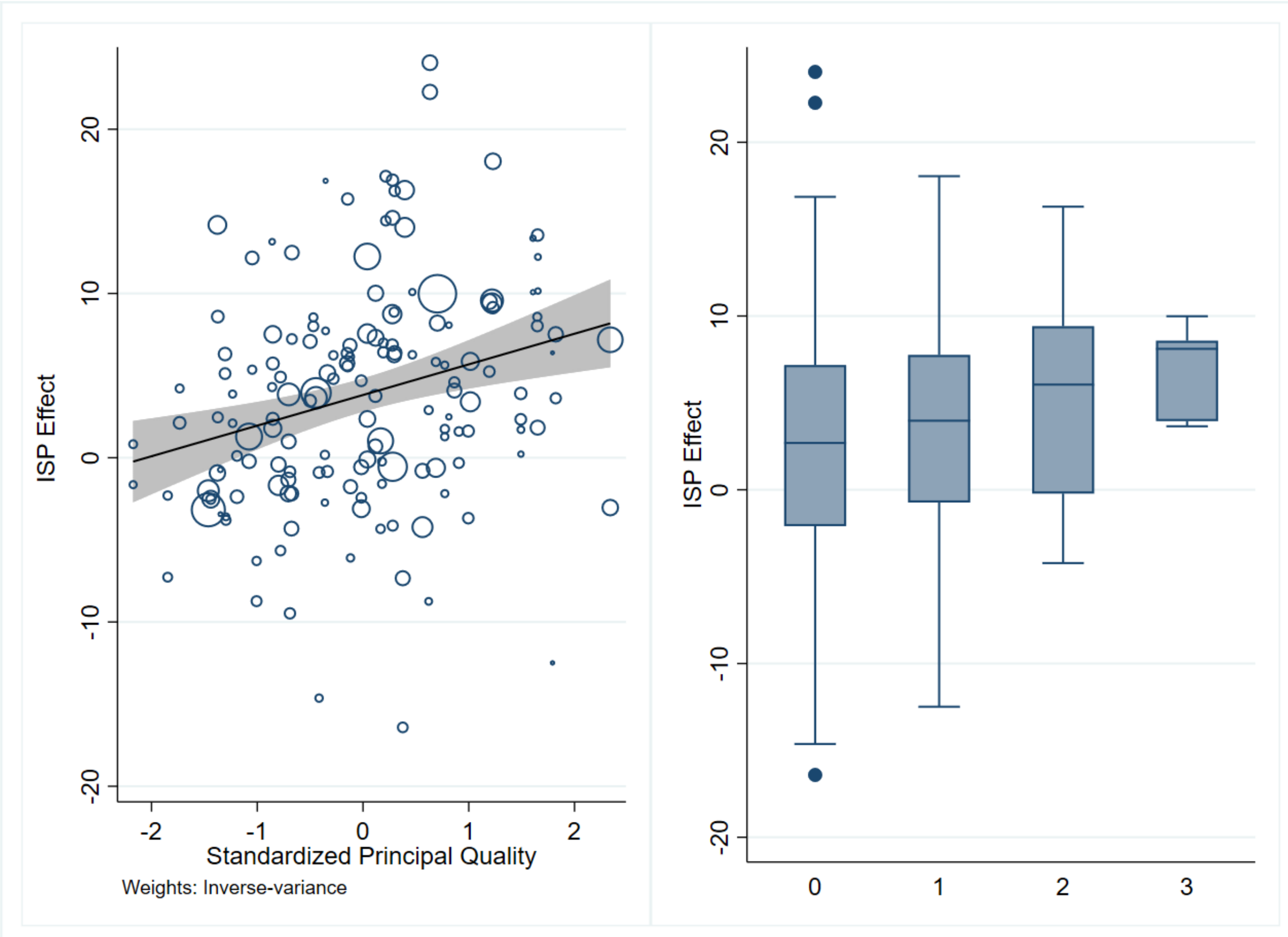
Notes: This figure presents the distribution of placebo treatment effects from 1,000 random reassignments of ISP treatment status. In the left panel, treatment assignments were permuted within matched school groups, preserving group-level characteristics. In the right panel, assignments were permuted across matched groups to test robustness. Each histogram shows the distribution of average effects on math and ELA passing rates across placebo replications. In both panels, the vertical line indicates the actual observed average effect.

Figure 7. *Distribution of Individual ISP Effects*



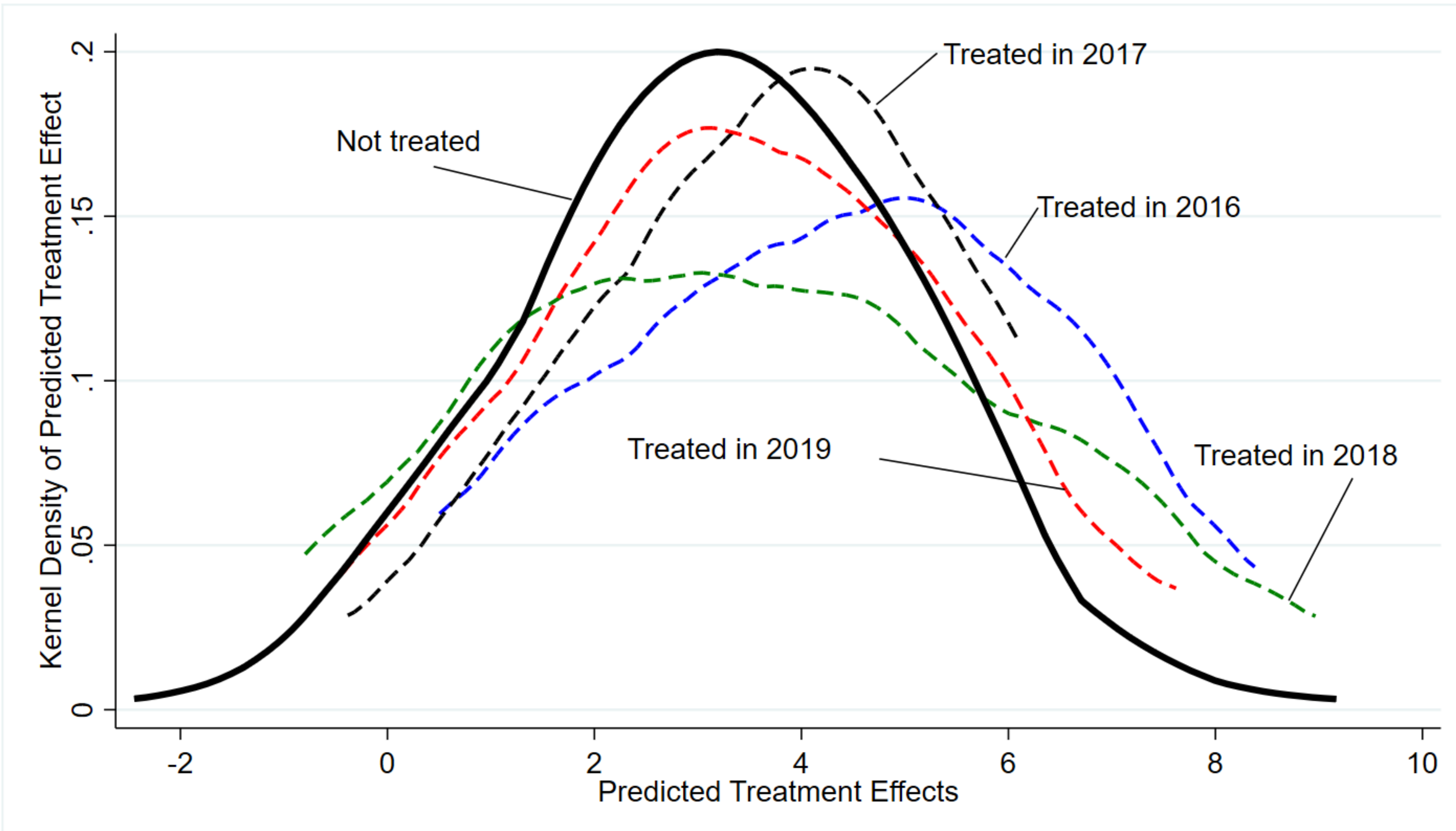
Notes: This figure displays a histogram of the estimated impact of the ISP on passing rates in ELA (blue) and Math (grey). The deconvolved density distribution is depicted, along with its 95 percent confidence interval (black). Additionally, a normal distribution (blue) is included for reference, with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

Figure 8. *Individual ISP Effects by Measures of Principal Quality and Outlier Status*



Notes: 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, two, and three categories.

Figure 9. *Predicted ISP Effects for All Schools*



Notes: This figure presents kernel density plots of predicted ISP effects for both ISP-designated schools (colored dashed lines) and non-ISP-designated schools (solid black line). Predicted effects are obtained from a meta-regression estimated among treated schools, with fitted values applied to the full sample. Dashed lines represent the distribution of predicted effects by ISP designation year: blue for 2016, black for 2017, green for 2018, and red for 2019.

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.

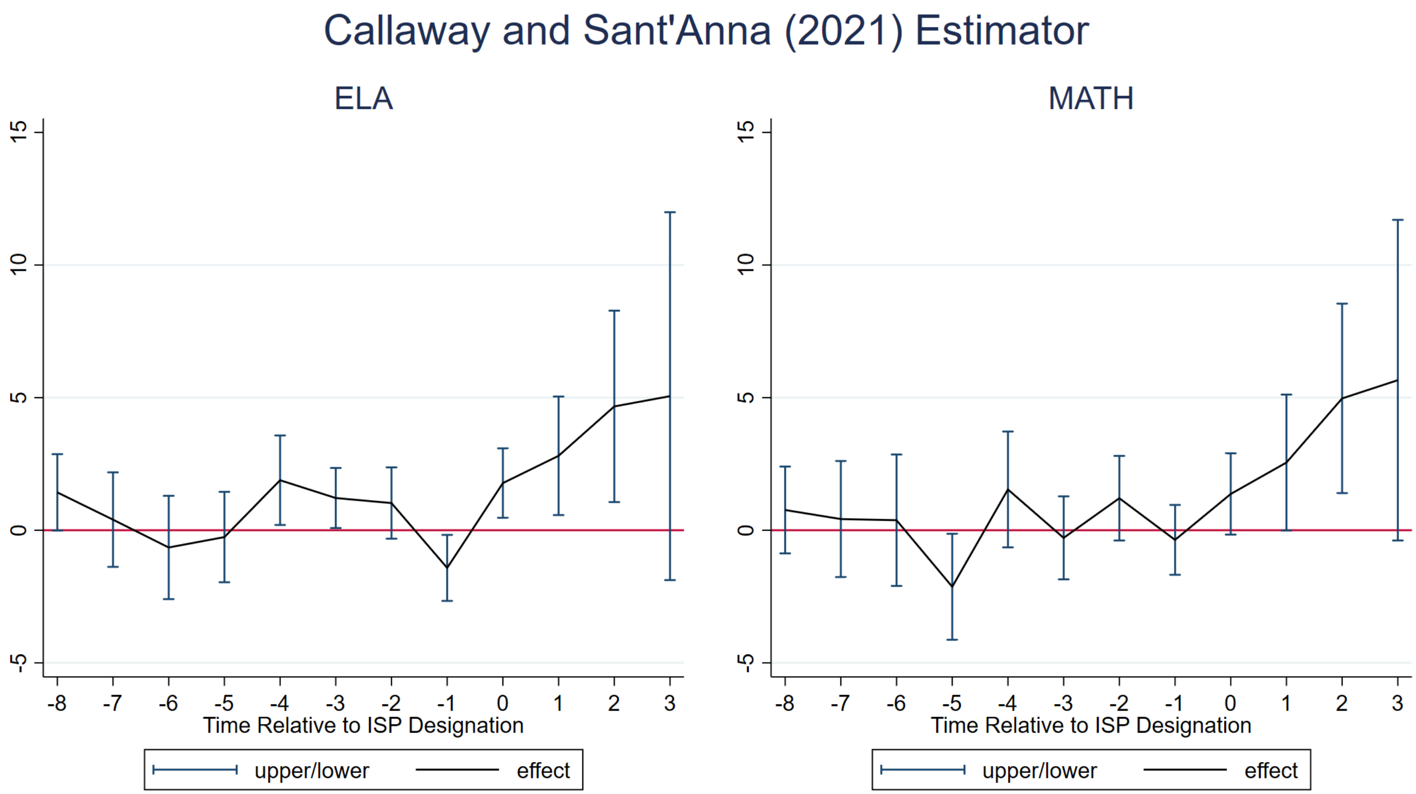
Figure A3. *List of ISPs (Pages 1 and 2)*

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)*

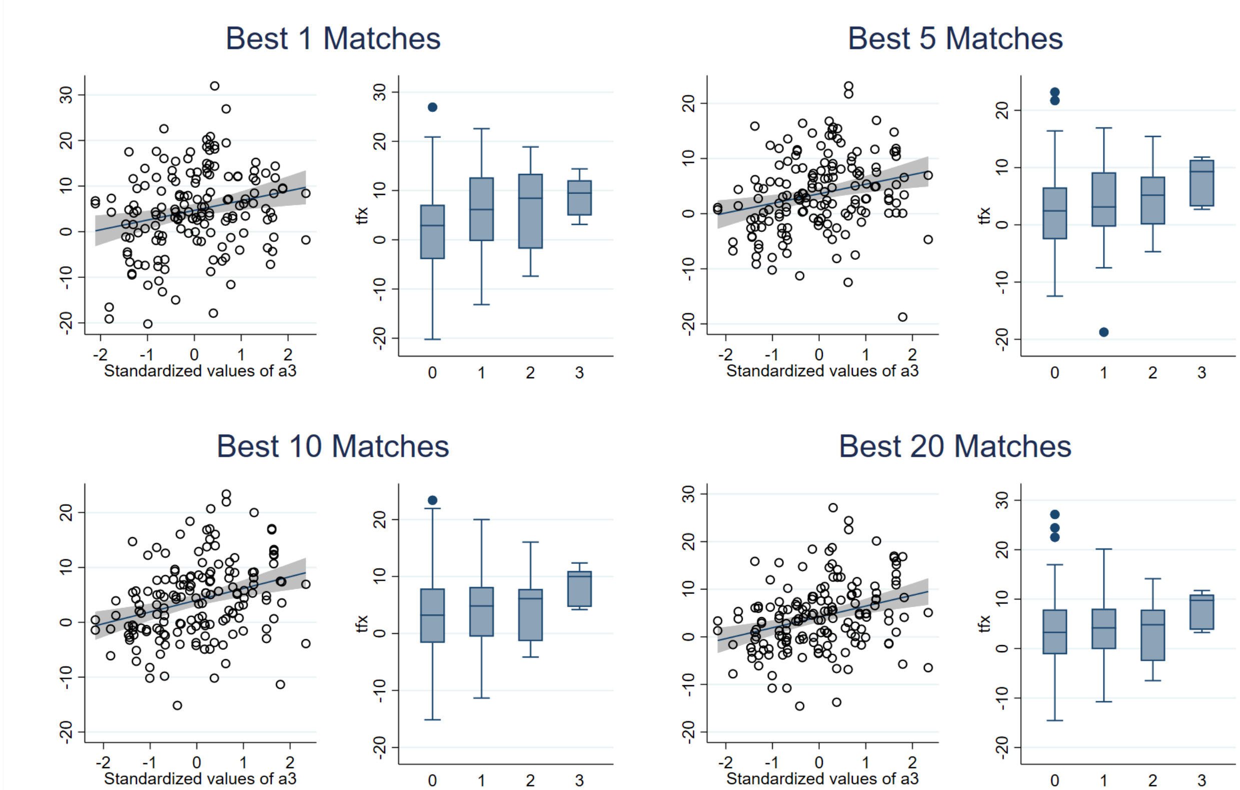
Lillian Lazu	LITTLE VILLAGE	ES	2018
Jay Thompson	LLOYD	ES	2016
July Cyrwus	LORCA	ES	2018
Erin Galfer	MARINE LEADERSHIP AT AMES HS	HS	2018
Jose Juan Torres	MARSH	ES	2020
Joseph Shoffner	MCCLELLAN	ES	2018
Jo Easterling-Hood	MCDOWELL	ES	2017
Karime Asaf	MOOS	ES	2016
Catherine Reidy	MOUNT GREENWOOD	ES	2017
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*



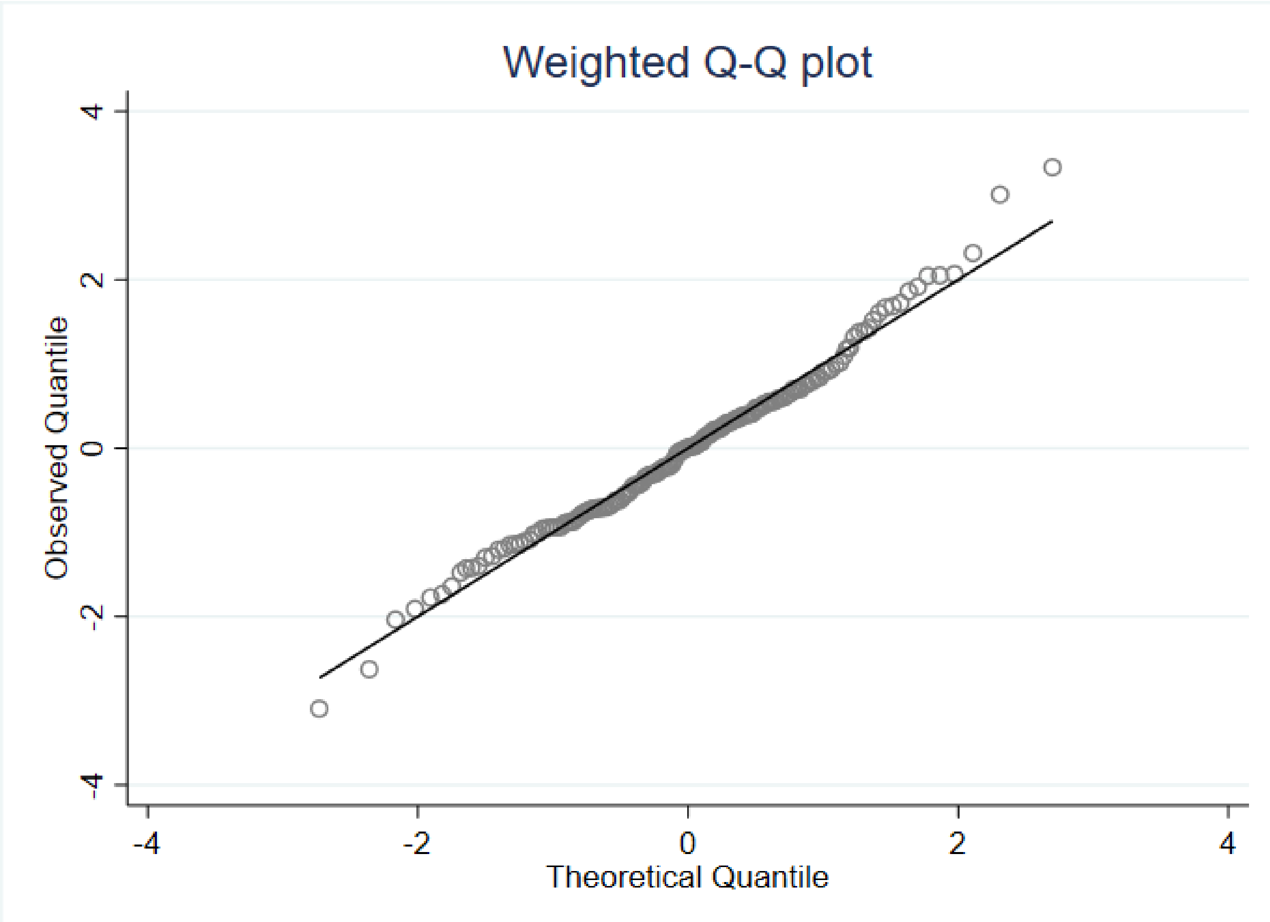
Notes: The event-study estimates depict the difference-in-difference models using the methodology of Callaway and Sant'Anna (2021). This model is estimated without covariates. Year 0 is the first year after ISP designation (where there could be an effect). Effects on ELA are on the left and effects on math on the right.

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



Notes: Each panel presents a scatter 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. Note that these plots are not precision weighted. **Left:** a scatter plot displaying the raw ISP effects plotted against the standardized principal alignment measure. Each observation has equal weight. The plot includes a 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, two, and three categories.

Figure A7. *Normality of ISP Effects*



Notes: 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 0.22 – indicative of the distribution of true effects being approximately normal.

Table A1: Comparison of School Characteristics **Before 2015**: Full Panel and Matched Sample

	(1) **Panel: ISP = 0**	(2) **Matched: ISP = 0**	(3) **Matched: ISP = 1**	(4) ***p*-value (Panel)**	(5) ***p*-value (Matched)**
Percent White	0.06	0.17	0.15	0.000	0.440
	(0.14)	(0.21)	(0.19)		
Percent Black	0.59	0.26	0.25	0.000	0.932
	(0.41)	(0.35)	(0.31)		
Percent Hispanic	0.31	0.50	0.52	0.000	0.739
	(0.36)	(0.34)	(0.35)		
Percent Special Ed.	0.15	0.13	0.12	0.032	0.928
	(0.11)	(0.09)	(0.11)		
Percent Bilingual	0.12	0.21	0.21	0.000	0.973
	(0.16)	(0.18)	(0.18)		
Enrollment	689.27	767.33	741.00	0.277	0.666
	(745.18)	(496.58)	(430.95)		
Attendance Rate	0.93	0.95	0.95	0.000	0.913
	(0.05)	(0.02)	(0.02)		
Mobility Rate	0.21	0.14	0.13	0.000	0.387
	(0.15)	(0.08)	(0.10)		
Chronic Truancy Rate	0.35	0.22	0.23	0.000	0.607
	(0.26)	(0.21)	(0.21)		
Percent Free or Reduced-Price Lunch	0.87	0.78	0.78	0.002	0.891
	(0.19)	(0.26)	(0.25)		
ELA Passing Rate (%)	36.23	45.75	46.01	0.000	0.916
	(22.67)	(23.37)	(24.18)		
Math Passing Rate (%)	36.74	47.51	47.72	0.000	0.934
	(25.74)	(25.72)	(26.09)		
ELA Percentile (Grades 3–8)	41.65	61.58	64.28	0.000	0.465
	(30.61)	(25.90)	(25.32)		
Math Percentile (Grades 3–8)	41.76	63.98	67.38	0.000	0.367
	(31.66)	(26.47)	(26.23)		
Instructional Leadership Score	53.11	57.59	60.62	0.006	0.252
	(20.94)	(19.25)	(18.91)		
FiveEssentials Score	3.28	3.81	3.93	0.000	0.524
	(1.46)	(1.30)	(1.27)		

Notes: Means and standard deviations (in parentheses) are shown for the full panel and the matched sample. *P*-values test the null hypothesis of equality of means between ISP and non-ISP schools.

Table A2: Estimated ISP Effects on Test Outcomes: Comparison of 1-to-1 Matching and 20 Matches

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	1 Match						**20 Matches**					
	ELA Passing Rate	**Math Passing Rate**	**ELA Percentile (2014–2019)**	**Math Percentile (2014–2019)**	**Predicted Math Passing Rate**	**Predicted ELA Passing Rate**	**ELA Passing Rate**	**Math Passing Rate**	**ELA Percentile (2014–2019)**	**Math Percentile (2014–2019)**	**Predicted Math Passing Rate**	**Predicted ELA Passing Rate**
ISP Year 1	3.682***	3.541***	1.039	4.182**	-0.126	-0.192	3.223***	2.559***	0.708	2.707**	0.253	0.177
	[1.048]	[1.108]	[1.345]	[1.717]	[0.572]	[0.590]	[0.813]	[0.864]	[0.951]	[1.242]	[0.334]	[0.339]
ISP Year 2	4.434***	3.945***	2.133	5.340**	0.205	0.517	4.093***	3.277***	1.539	3.170*	0.756	0.864*
	[1.109]	[1.454]	[1.754]	[2.342]	[1.097]	[1.114]	[1.021]	[1.076]	[1.287]	[1.814]	[0.473]	[0.483]
ISP Year 3+	8.477***	6.051***	3.694**	6.469***	0.027	0.232	7.596***	6.267***	3.371***	5.464***	0.716	0.609
	[1.947]	[2.200]	[1.816]	[2.425]	[1.028]	[1.006]	[1.680]	[1.687]	[1.235]	[1.840]	[0.465]	[0.439]
ISP (Pooled)	4.976***	4.221***	1.858	4.947***	0.006	0.113	4.440***	3.585***	1.450	3.364**	0.506	0.478
	[1.022]	[1.193]	[1.257]	[1.714]	[0.675]	[0.681]	[0.891]	[0.898]	[0.956]	[1.337]	[0.340]	[0.339]
Observations	1,406	1,402	792	792	1,500	1,500	15,133	15,129	9,165	9,165	15,690	15,690

Notes: Robust standard errors are shown in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Each column reports the estimated effect of ISP designation on student outcomes under two alternative matching strategies: one-to-one matching (left panel) and 20 matches per treated school (right panel). The top panel reports estimates for the first year, second year, and third year and beyond under the ISP, all from the same regression model. In the lower panel, each coefficient for ISP(Pooled) reflects the before-versus-after ISP designation comparison from a separate regression model.

Table A3: Estimated ISP Effects on Demographic and School Characteristics (One-to-One Matching)

	(1) **Percent White**	(2) **Percent Black**	(3) **Percent Hispanic**	(4) **Percent Special Ed.**	(5) **Percent Bilingual**	(6) **Mobility Rate**	(7) **Percent Free Lunch**	(8) **Enrollment**
ISP Year 1	0.00801	0.00110	-0.0132**	0.00032	-0.0163**	-0.00038	-0.00146	19.83
	[0.00496]	[0.00501]	[0.00637]	[0.00301]	[0.00772]	[0.00584]	[0.00766]	[17.51]
ISP Year 2	0.00423	-0.00057	-0.00744	-0.00477	-0.0150*	0.00851	-0.00972	22.42
	[0.00560]	[0.00732]	[0.00805]	[0.00801]	[0.00903]	[0.00656]	[0.0102]	[24.12]
ISP Year 3+	-0.00084	0.00408	-0.00982	-0.00450	0.00380	0.0170	-0.00229	3.123
	[0.00822]	[0.0150]	[0.0153]	[0.00492]	[0.00999]	[0.0116]	[0.00935]	[37.09]
ISP (All Years)	0.00489	0.00128	-0.0108	-0.00219	-0.0117*	0.00623	-0.00401	16.69
	[0.00509]	[0.00643]	[0.00722]	[0.00373]	[0.00703]	[0.00631]	[0.00678]	[22.04]
Observations	1,492	1,492	1,492	1,174	1,106	1,426	1,406	1,442

Notes: Panel A shows year-specific ISP effects on demographic composition and school characteristics using one-to-one matched samples. Panel B reports the average effect across all years. Robust 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.

Appendix B: Bayesian Estimates

This section outlines the logic of the Bayesian meta-analysis model, drawing closely on the formulation in Jackson and Mackevicius (2024). Each study is assumed to have a true effect θ_j, with the corresponding estimate $\hat{\theta}_j$ subject to sampling error. By the central limit theorem, the sampling distribution of estimates follows:

$$\hat{\theta}_j \sim \mathscr{N}(\theta_j, \sigma_j^2) \tag{7}$$

The true effects θ_j vary across studies due to heterogeneity and are centered around a grand mean Θ, with variance τ^2. Following convention, I assume the true effects are normally distributed:

$$\theta_j \sim \mathscr{N}(\Theta, \tau^2) \tag{8}$$

Let the set of true effects be $\theta = [\theta_1, \theta_2, \theta_3, ...\theta_J]'$. The observed estimates of these true effects are $\hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, ...\hat{\theta}_J]'$. The corresponding sampling standard deviations are $\sigma = [\sigma_1, \sigma_2, \sigma_3, ...\sigma_J]'$ which is approximated by $se = [se_1, se_2, se_3, ...se_J]'$.

Because the probability of observing the estimated effects ($\hat{\theta}$) is a function of true effects (θ), the probability of which is determined by τ, the likelihood of observing estimates ($\hat{\theta}$) and sampling standard deviations (se) can be computed for any given value of τ, Θ, and θ – that is, $\mathscr{L}(\tau, \Theta, \theta)$. Frequentist approaches, such as Maximum Likelihood, solve for the values of τ, Θ, and θ that maximize this likelihood.

Bayes' rule says that the joint posterior probability for the parameters (i.e., $p(\tau, \theta, \Theta|\hat{\theta}, se)$, is proportional to the likelihood of the data given certain parameter values ($\mathscr{L}(\tau, \Theta, \theta)$) multiplied by the prior probability of those parameters ($\pi(\tau, \Theta, \theta)$). As such, using Bayes rule, given some prior distribution, one can compute the posterior distribution of the true effects Θ, θ, and τ. Moments (such as the mean) of the posterior distributions of τ,Θ, and θ provide information about the values of these parameters. Moreover, the spread of the posterior distributions sheds light on the uncertainty around the values of these parameters.

The Bayesian model works as follows:

1. One chooses a probability density — i.e., prior distribution — that expresses beliefs about the distribution of each parameter *before seeing any data*.

2. One defines a statistical model $p(\hat{\theta}, se|\tau, \theta, \Theta)$ that reflects our beliefs about the data given the parameters.

3. After observing data $\hat{\theta}$ and se, the model updates our beliefs using Bayes rule and calculates the joint posterior distribution for the parameters of interest $p(\tau, \theta, \Theta|\hat{\theta}, se)$.

4. The model takes random draws of τ, Θ, and θ from the posterior distributions and reports moments (in our case, the mean) of the posterior distribution of the parameter estimates.

   - Note that $p(\theta, \Theta, \tau|\hat{\theta}, se)$ can be written as $p(\theta|\Theta, \tau, \hat{\theta}, se)p(\Theta, \tau|\hat{\theta}, se)p(\tau, |\hat{\theta}, se)$. As such, the model will draw the hyperparameters τ, then Θ, from their marginal posterior distributions and then draw θ from its posterior distribution conditional on the drawn values of τ and Θ.

Under this approach, one must define the prior distributions for τ and Θ. To this aim, I assume that the true effect is a random draw from a normal distribution (justified by the central limit theorem), and that the heterogeneity parameter τ^2 follows an inverse Gamma distribution as in (9) and (10).

$$\Theta \sim \mathcal{N}(.) \tag{9}$$

$$\tau^2 \sim InvGamma(.) \tag{10}$$

The inverse Gamma distribution is commonly used to model variance parameters and avoids the non-negative estimates one can obtain from method of moments approaches. I estimate this model with starting values such that $\tau^2 \sim InvGamma(0.0001, 0.0001)$ and that $\Theta \sim \mathcal{N}(0, 100)$. The resulting Θ and τ from this model is similar to those using frequentist methods.