Welfare Added?
Optimal Teacher Assignment with Value-Added Measures

Tanner S. Eastmond* Michael David Ricks†

Julian Betts‡ Nathan Mather§

July 2026

Abstract

We study how teacher "value added" should inform optimal teacher-assignment policy. Our welfare-theoretic framework illustrates (1) how theoretically optimal assignments leverage variation in teachers' impacts both across student types and across different outcomes, and (2) how empirically optimal assignments trade off improved targeting from estimating richer student heterogeneity against a growing winner's curse. In practice, optimal assignments use limited student types (only lagged achievement) and multiple outcomes (not just math). Even after correcting for policy overfitting, assignments raise average present-value earnings by $3,000 and increase lower-achieving students' earnings by 66–142% more than benchmark policies using homogeneous effects, single-subject heterogeneity, or teacher deselection.

*Department of Economics, Brigham Young University: tanner.eastmond@byu.edu

†Department of Economics, University of Nebraska–Lincoln: mricks4@unl.edu

‡Department of Economics, University of California, San Diego, and NBER: jbetts@ucsd.edu

§Secretariat: nmather@secretariat-intl.com

This research is the product of feedback and comments from many people including Garrett Anstreicher, Andrew Bacher-Hicks, Sandy Black, Ash Craig, Gordon Dahl, William Delgado, Itzik Fadlon, Julie Cullen, Amy Finkelstein, Yifan Gong, Jim Hines, Nathan Hendren, Caroline Hoxby, Peter Hull, Brian Jacob, Erzo Luttmer, Lars Lefgren, Ross Milton, Jonah Rockoff, Jesse Rothstein, Andrew Simon, Rich Patterson, Nolan Pope, and researchers at the Education Policy Initiative, Youth Policy Lab, and SANDERA, as well as with seminar and conference participants at the University of California, San Diego, the University of Michigan, Brigham Young University, the CAL Labor Summit, Boston University, the ASSA Annual Meetings, Simon Fraser University's Vancouver Applied Microeconomics Conference, NBER Education Spring Program Meeting, and NBER Public Spring Program Meeting. Thanks also to Andy Zau, who facilitated data access, and to Wendy Ranck-Buhr, Ron Rode, and others at the San Diego Unified School District for their interest and feedback. The views expressed in this paper are those of the authors and do not reflect the views of Secretariat Advisors or its clients. `Refine.ink` was used to check the paper for consistency and clarity.

1. Introduction

Each year school districts assign teachers to classes, distributing scarce instructional skill across hundreds of millions of students. Research on teacher effects, especially "value-added" measures, reveals that teacher assignments shape students' lives for decades (e.g., Chetty et al., 2014b), motivating a broader policy question: How should teacher effects be used to make socially optimal assignments? Our paper studies this public-service allocation problem and quantifies the gains from optimal policy.

We propose a welfare-theoretic framework addressing two key complications with using value added for teacher assignments. First, because teacher effects vary across student subgroups (e.g., Delgado, 2025) and outcomes (e.g., Jackson, 2018; Petek and Pope, 2023), there is no single ranking of assignments (Condie et al., 2014). We characterize the (full-information) first-best assignment by aggregating gains with welfare weights and an index for lifetime utility. Second, because value-added measures are estimated with noise, assignments based on estimated effects suffer a "winner's curse" (as in Andrews et al., 2024). Modeling richer heterogeneity can improve targeting but can worsen the winner's curse and cause misallocations. We let the social planner face this tradeoff directly by choosing which model of teacher effects to use. By incorporating match effects, multiple outcomes, model choice, and distributional objectives, the social planner is able to characterize a (feasible) "second-best" solution.

Focusing on the social planner's problem highlights three understudied issues in the teacher value-added literature. First, maximizing welfare requires integrating gains across subgroups and outcomes, but existing work in applied econometrics focuses on only measuring these differences (e.g., Mulhern and Opper, 2021; Gilraine et al., 2022; Graham et al., 2023; Delgado, 2025). Second, optimal policy requires endogenizing the measurement of value added, which papers studying labor markets and education policy typically take as given (e.g., Bau, 2022; Beuermann et al., 2023; Bates et al., 2025). Third, because other teacher assignment counterfactuals focus only on math-score match effects (as in Ahn et al., 2025; Laverde et al., 2026; Bobba et al., 2026), they forego welfare gains from incorporating multiple outcomes, distributional objectives, and model choice.[1] The resulting welfare loss depends on the empirical distance to the optimal second-best frontier.

We test the quantitative importance of optimal policy by estimating teacher value added in the San Diego Unified School District (SDUSD). We construct shrinkage-adjusted value added for 4,000 third- through fifth-grade teachers from 1998 through 2019 on math, read-

[1] Graham et al. (2023) explicitly call counterfactuals focused on match effects a "first-pass" and emphasize the need to consider model selection, outcomes beyond math scores, and distributional objectives.

ing, attendance, and behavior GPA, with heterogeneity by lagged-outcome quantiles, race, gender, and their interactions. Although the resulting value-added measures are highly correlated with standard, homogeneous value added, the average within-teacher difference in value added across groups (i.e., comparative advantage) is still substantial: 26–44% of the standard deviation of standard value added. In addition to passing a battery of robustness and specification checks, our estimates are also forecast-unbiased, quasi-experimentally honest, and stable across changes in class composition.

We use these estimates to trace out the second-best frontier of gains from teacher assignments as follows. Given each set of estimates and welfare weights, we choose assignments that maximize welfare-weighted expected earnings gains (or intermediate outcomes) using linear programming (Bertsimas and Tsitsiklis, 1997). We consider two policies—one reassigning teachers within school-grade cells and the other considering all 10^{690} possible within-grade assignments in the district. We then characterize the optimal second-best frontier by comparing the optima attained under different estimates of value added.

These exercises produce three main results. First, even after corrections for policy overfitting, optimal reassignments produce pronounced gains relative to the status quo. Although feasible shrinkage methods attempt to control the distribution of teacher effects, they do not automatically prevent overfitting or the winner's curse in the assignment problem. To address potential bias from estimation error, we (1) use robust optimization (Gabrel et al., 2014) to select assignments that are less sensitive to estimate uncertainty, (2) deflate predicted gains from reassignments with quasi-experimental forecast accuracy, and (3) evaluate assignments under the joint posterior to identify the expected (rather than the predicted) gains. The resulting gains are conservative but still large. For example, using simple, homogeneous estimates of math value added to place better teachers in larger classes district-wide would increase average scores by about 0.06 standard deviations over third through fifth grade, or +\$1,500 in present-value earnings (after accounting for correlated effects on other outcomes). The gains from the second-best earnings-maximizing policy are nearly double, just under +\$3,000 per student. These are 1.3× and 2.3× the gains from a benchmark 5% teacher "deselection" intervention based on math value added (Hanushek, 2009).

Second, using multiple outcomes expands the frontier of feasible gains and reshapes their incidence. Our optimal second-best policies yield up to 2× larger gains than math-only assignment policies. Interestingly, the gains from considering multiple outcomes disproportionately accrue to lower-achieving students. For example, under a math-score maximizing assignment policy, higher-achieving students gain over +\$3,100 versus +\$1,700 for lower-achieving students. The second-best policy gives lower-achieving students an additional +\$1,100 (68%), with no statistically appreciable difference for higher-achieving students.

Because varying the welfare weights traces out a frontier of optima with different incidence, a redistributive planner considering multiple outcomes could provide over +$5,400 to lower-achieving students without harming higher-achieving students on average.

Third, the second-best policy reveals that it is optimal to model only a modest amount of student heterogeneity loaded on lagged academic achievement. We find that the marginal gains from considering additional heterogeneity quickly diminish—and can reverse as forecast quality deteriorates. In our setting, partitions based on lagged outcomes tend to produce larger gains and a smaller winner's curse than those using demographic characteristics (as in Delgado (2025) or Ahn et al. (2025)),[2] and race-blind implementations of the second-best policy often dominate race-focused assignments for both minority and non-minority students. Furthermore, including non-cognitive outcomes actually reduces expected earnings gains because of their poor out-of-sample performance (as in Petek and Pope (2023)). As a result, the optimal second-best policies use terciles of lagged math and reading scores to maximize earnings and use lagged achievement quartiles to maximize math scores. The preeminence of lagged achievement reflects the role of core pedagogical practices, such as differentiated instruction (see Betts, 2011), that affect education production.

Given the size of the expected gains, we perform additional analyses to probe the real-world feasibility of optimal assignments. For example, we find that smaller, targeted interventions still produce large gains: reassigning 10% (25%) of teachers still attains 33% (60%) of the second-best. Similarly, rather than holding compensation fixed and requiring assignment incentive compatibility, we show that the implied surplus from the second-best is large enough to compensate teachers for accepting welfare-improving reassignments. We find that a $15,000 payment to change schools (as in Glazerman et al., 2013) has a marginal value of public funds (MVPF; Hendren and Sprung-Keyser, 2020) of 2.0, and payments up to $1,200 have an infinite MVPF for within-school class assignments (compare with Kane et al., 2013). Even if optimal policies are not implementable, the potential gains from reassignment in our context illuminate the shadow value of relaxing constraints on reassignment and compensation—roughly $700 million.

Our paper expands broader conversations about education policy, value added in other contexts, and empirical public finance. Regarding education policy, our paper moves the conversation about value added beyond accountability and toward welfare. In this sense, our research on teacher assignment is most similar to other work using teacher- or school-value-added as a primitive to understand other economically important education policy issues such as school competition (Bau, 2022), teacher labor markets (Biasi et al., 2021; Bobba et al., 2026; Bates et al., 2025; Laverde et al., 2026), and school choice (Beuermann et al.,

[2]Or as in the teacher match literature (Dee, 2005; Delhommer, 2022).

2023). In contrast, work on value added has mainly considered it as a tool for accountability and firing (see Jackson et al., 2014), with recent extensions to staffing high-needs schools (e.g., Glazerman et al., 2013; Colas and Fu, 2025; Harris et al., 2026). We find that our optimal assignments outperform these accountability and staffing uses of value added in both efficiency and equity.

Our paper informs the use of value added in settings beyond teacher assignment. Researchers and policymakers directly use measures similar to value added in contexts as diverse as education,[3] healthcare,[4] and governance,[5] and implicitly use them in nearly every judge-IV design.[6] We provide practical and broadly applicable insights into effectively using value added in optimal policy problems. For example, when researchers are interested in a unit's "welfare added," our results highlight the first-order importance of considering multiple outcomes. Similarly, the primacy of lagged outcomes in model choice could inform other models of heterogeneous effects (e.g., Arnold et al., 2022; Dahlstrand, 2022; Einav et al., 2025b). Our paper also complements work studying allocation problems from a mechanism-design perspective while taking value-added measurement as given (e.g., Baron et al., 2024).

Finally, we provide an empirical example of addressing uncertainty and model choice in empirical welfare analyses in two ways. First, by proposing a welfare framework that endogenizes model choice, we allow the social planner to pursue targeting benefits without ignoring misallocation risks.[7] Second, we empirically quantify the costs and benefits of considering increasingly rich heterogeneity, showing that the winner's curse is a quantitatively important consideration in optimal policy. Growing theoretical literatures emphasize how targeting based on treatment effects can improve welfare (Kitagawa and Tetenov, 2018; Athey and Wager, 2021) and how optimization with imperfect estimates can generate regret—especially in highly non-linear settings (Mbakop and Tabord-Meehan, 2021; Andrews et al., 2024). While the first insight motivates recent interventions targeting treatments as varied as social safety programs (Alatas et al., 2016; Finkelstein and Notowidigdo, 2019), energy-efficiency interventions (Ito et al., 2023; Ida et al., 2022), entrepreneurial lending (Hussam et al., 2022), and interventions against gun violence (Bhatt et al., 2024), the regret from winner's curse has received much less empirical attention. In this sense, our paper is an empirical complement

[3]In addition to teachers, consider evaluations of principals (e.g., Branch et al., 2009; Hanushek et al., 2024), counselors (Mulhern, 2023) and schools (Angrist et al., 2023).

[4]Consider papers about doctors (Chan et al., 2022; Dahlstrand, 2022), hospitals (Chandra et al., 2016; Doyle et al., 2019; Hull, 2020), and nursing homes (Einav et al., 2025a,b)

[5]In addition to prosecutors (Harrington and Shaffer, 2023) and public defenders (Abrams and Yoon, 2007; Landon, 2024) there is research on case examiners (Norris, 2019) and case workers (Baron et al., 2024).

[6]For a typical judge IV, the reduced form functions as a value-added measure (e.g., Kling, 2006; Aizer and Doyle Jr, 2015; Dobbie et al., 2018; Bhuller et al., 2020).

[7]Cutting-edge work on value added model choice focuses on in-sample diagnostics of match effects (Ahn et al., 2025), which we extend to the implications of model choice for optimal policy or targeting.

to these papers and contemporaneous theoretical work like Chernozhukov et al. (2025).

This paper contains six sections. Section 2 introduces our framework for welfare and value added. Section 3 presents our data, background, and estimation procedure. Section 4 illustrates our optimal assignments, focusing only on math scores to build intuition. Section 5 extends the analysis to consider multiple outcomes, welfare, and policy. Section 6 concludes.

2. Optimal Teacher Assignment Policy

This section considers the optimal assignment of teachers to classes using value-added measures. We (1) formalize the social planner's problem of assigning teachers to classes; (2) discuss the first-best teacher assignment policy; (3) define a plug-in estimator to predict the welfare of any assignment given value-added estimates; and (4) describe the choice of optimal teacher assignments as a feasible second-best policy. Although the exposition focuses on teacher value added, the theoretical insights apply to any setting in which a social planner can use different empirical estimates to target treatments with heterogeneous impacts.

2.1 Welfare Framework

Consider a policymaker selecting a policy $\mathcal{J}$ from a set of potential policies $\mathscr{J}$ to maximize $\sum_i \phi_i U_i^{\mathcal{J}}$, the weighted sum of students' lifetime utilities. In our application, each policy, $\mathcal{J} : i \to j$, is an assignment that maps all students to their teachers. To focus on the teacher-assignment problem, we consider only policies that hold classes constant,[8] given evidence that tracking and peer effects matter in educational settings.[9]

The gains from an assignment $\mathcal{J}$ depend on student-teacher match effects. We denote the match effect of teacher j on student outcome y_i as $\mu^y_{i,j}$. These match effects are both *heterogeneous*—meaning that each teacher's effect may vary among students, i—and *multidimensional*—meaning that each teacher may affect multiple outcomes, y, in different ways. Unfortunately, as pointed out in Condie et al. (2014), these match effects only partially order assignments in terms of welfare. While finding a Pareto gain would be ideal, this is only possible through a series of swaps such that $\mu^y_{i,j} \leq \mu^y_{i,j'}$ for all outcomes of all students. This condition cannot be met if each teacher has homogeneous effects on all students, and it is nearly impossible in general given the dispersion of student needs within classes and teacher effectiveness within the district.

[8] i.e., if two students share a teacher under assignment $\mathcal{J}$, they must also share a teacher under any other assignment $\mathcal{J}'$: $\mathscr{J} = \{\mathcal{J} : \mathcal{J}(i) = \mathcal{J}(i') \iff \mathcal{J}'(i) = \mathcal{J}'(i')\}$.

[9] To the extent to which manipulating either tracking or peer effects could generate additional gains, our optimal assignments will serve to give a lower bound on the social gains from a more flexible policy.

We use welfare theory to order the welfare gains from different assignments. The social planner can use each student's *ex ante* welfare weight to integrate equity concerns over effect heterogeneity. To integrate over multidimensionality the social planner can define a "score function," $S_i^{\mathcal{J}} = s(\boldsymbol{Y}_i^{\mathcal{J}}, \boldsymbol{X}_i)$, that maps all observable outcomes $\boldsymbol{Y}_i$ (such as test scores or annual earnings) and characteristics, $\boldsymbol{X}_i$ (such as lagged scores or poverty status), into a welfare-relevant scalar.[10] In Appendix B.1 we illustrate how to characterize the expected welfare of any assignment conditional on the scores, $\boldsymbol{S}^{\mathcal{J}}$:

$$\mathcal{W}^{\mathcal{J}} \equiv \sum_{j} \sum_{i:\, \mathcal{J}(i)=j} \omega_i \mu_{i,j}^{\mathcal{S}} \tag{1}$$

where $\mu_{i,j}^{\mathcal{S}}$ is teacher j's effect on student i's score function and ω_i is student i's S-adjusted welfare weight. Although the social planner may seek to maximize average scores ($\omega_i = 1\ \forall i$), revealed preference suggests that policymakers care particularly about gains to certain students. For example, the U.S. No Child Left Behind Act focused on low-achieving students by penalizing schools with subgroups that were not meeting standards.

Equation 1 implies three restrictions. First, welfare gains are linear in score gains, consistent with relatively modest ("local") changes in each student's scores.[11] Second, only student gains affect welfare. This restriction would hold if score gains fully capture family preferences for assignments and if teachers are indifferent between assignments. Alternatively, assignments must satisfy an incentive-compatibility constraint in which families will not re-sort to new schools or classes[12] and teachers are compensated sufficiently to switch classes willingly. In Section 5, we consider the welfare effects of various policies that could ensure teacher incentive compatibility. Third, the welfare formulation excludes any costs from policy implementation. These restrictions relate to prior research on teacher labor market failures, cross-student spillovers, school choice, union concerns, and administrative costs. We abstract from these issues to focus on the optimal assignment problem. Although these considerations matter, large enough gains from optimal assignment could support interventions to alleviate these concerns. Readers unsatisfied with these assumptions could instead consider all welfare gains in partial-equilibrium terms.

[10] One appealing score function could be a surrogate index for expected lifetime utility or earnings, see Athey et al. (2025) and the Remark in Appendix B.1.

[11] Relaxing this restriction to arbitrarily large changes is trivial if one is willing to specify (continuous) parameterizations of utility and welfare weights over the score.

[12] Formally, this relates to the "Policy Independence" condition assumed in Appendix B.1. This implication seems plausible because the vast majority of families do not request specific teachers in the status quo, and even then, not all requests are honored (Jacob and Lefgren, 2007). Additionally, families do not respond to information about value added in school choice (Abdulkadiroğlu et al., 2020) or housing markets (Imberman and Lovenheim, 2016).

2.2 The First-Best (Full-Information) Teacher Assignment Policy

The first-best assignment of teachers to classes maximizes (welfare-weighted) scores. To build intuition about this assignment policy, consider two examples, assuming all match effects are known:

- First, let each teacher j's effect be homogeneous: $\mu^{\mathcal{S}}_{i,j} = \bar{\mu}^{\mathcal{S}}_{j}$. After ranking teachers by their absolute advantage, the first-best assignment would place the best teachers in the classes with the largest (welfare-weighted) number of students. This assignment rule generalizes the first-best policy proposed in Bates et al. (2025).

- Second, let all class sizes be equal and let each teacher j have differentiated effects across two student subgroups (denoted by k): $\mu^{\mathcal{S}}_{i,j} = \bar{\mu}^{\mathcal{S}}_{j,k}$ for all i in group k. After ranking teachers by their comparative advantage at teaching group k, the first-best assignment would place the most specialized teachers in the classes with the largest (welfare-weighted) share of group k students. This assignment rule generalizes the policy proposed in Delgado (2025) and in Ahn et al. (2025) for the two-subgroup case.

In reality, both match effects and class sizes may vary. The first-best assignment, therefore, must trade off marginal gains from placing better (higher absolute-advantage) teachers in larger classes against marginal gains from placing more specialized (higher comparative-advantage) teachers in well-matched classes.

The full-information first-best policy is feasible only if the social planner knows every possible match effect, $\mu^{\mathcal{S}}_{i,j}$. In practice, the social planner must rely on the econometrician to estimate match effects. The following subsections outline the implications of using value-added estimates to predict welfare and present an approach to identify second-best policies by endogenizing value-added model choice.

2.3 Estimating Welfare with Value-Added Measures

To choose between possible teacher assignment policies, the social planner needs estimates of welfare under each policy rule. One approach is to use value-added measures, typically constructed as the (leave-year-out, shrinkage-adjusted) mean residual of test-score gains for students taught by teacher j. Given the value added estimates $\hat{\boldsymbol{\mu}}_m$ from any model, m, we define a plug-in estimator of welfare based on Equation 1 as

$$\begin{aligned}\widehat{\mathcal{W}}^{\mathcal{J}}_{m} &= \sum_{j} \sum_{i:\mathcal{J}(i)=j} \omega_i \hat{\mu}_{m,j,k(i)} \\ &= \sum_{j} \sum_{k \in \mathcal{K}_m} n_{j,k}(\mathcal{J})\ \bar{\omega}_{j,k}(\mathcal{J})\ \hat{\mu}_{m,j,k}\end{aligned}$$

where students are partitioned into $\mathcal{K}_m$ subgroups, $n_{j,k}(\mathcal{J})$ represents the number of type-k students in teacher j's assigned class,[13] $\bar{\omega}_{j,k}(\mathcal{J})$ is their average welfare weight, and $\hat{\mu}_{m,j,k}$ predicts the relevant effect of teacher j (based on model m). When using a standard (homogeneous) value-added model, the welfare estimate is similar to the common welfare approximation of average treatment effect multiplied by average welfare weights (see discussion in Hendren and Sprung-Keyser, 2020). But standard value added will systematically mischaracterize welfare when teacher effects vary with race (Delgado, 2025), poverty (Bates et al., 2025), achievement (Biasi et al., 2021; Aucejo et al., 2022; Graham et al., 2023), or their interactions (Ahn et al., 2025)—as will subgroup value-added estimates when other idiosyncratic student-teacher match effects are present.

Equation 2 decomposes the difference between estimated and realized welfare into five interpretable terms. While the equation is written in general, the text discusses the case of standard value added (ignoring the k subscripts) to build intuition.

$$\mathcal{W}^{\mathcal{J}} - \widehat{\mathcal{W}}^{\mathcal{J}}_m = \sum_j \sum_{k \in \mathcal{K}_m} n_{j,k}(\mathcal{J})\, \bar{\omega}_{j,k}(\mathcal{J}) \Bigg[\underbrace{\tilde{\mu}_{j,k}(\mathcal{J}) - \bar{\mu}_{j,k}}_{\text{Matching Gains}} + \underbrace{\frac{\text{Cov}(\omega_i, \mu_{i,j}; \mathcal{J})}{\bar{\omega}_{j,k}(\mathcal{J})}}_{\text{Distributional Gains}} + \underbrace{\bar{\mu}_{j,k} - \tilde{\mu}_{j,k}(\mathcal{J}_0)}_{\text{External Validity}} + \underbrace{\tilde{\mu}_{j,k}(\mathcal{J}_0) - \mu^{\star}_{j,k}}_{\text{Shrinkage Bias}} + \underbrace{\mu^{\star}_{j,k} - \hat{\mu}_{m,j,k}}_{\text{Estimation Error}} \Bigg] \tag{2}$$

where $\tilde{\mu}_{j,k}(\mathcal{J})$ represents the average match effects of teacher j on type-k students in their assigned class under $\mathcal{J}$, and $\bar{\mu}_{j,k}$ is the teacher's average type-k match effect in the population. $\text{Cov}(\cdot)$ reports the residual within-class-subgroup covariance of welfare weights and match effects in teacher j's assigned class. Finally, $\tilde{\mu}_{j,k}(\mathcal{J}_0)$ is the average match effect of teacher j on type-k students in the class teacher j taught in the estimation sample and $\mu^{\star}_{j,k}$ is the oracle value added estimates (see the full derivation in Appendix B.2).

We will consider each term in turn. The first three terms all reflect fundamental welfare bias generated by model misspecification of an over-simplified target welfare object. When comparing models, these biases will tend to decrease as the number of relevant subgroups rises. The fourth and fifth terms reflect the imperfect nature of estimation. They capture increased bias introduced through shrinkage and also reflect the variability of the value added estimate. As such, only these terms govern the variability of $\widehat{\mathcal{W}}^{\mathcal{J}}_m$ around the target welfare object.

First, focusing on the average (or subgroup-average) effect of each teacher ignores the

[13] The presence of n in the welfare approximation reveals the importance of absolute advantage even when making assignments using match effects—requiring adaptations from assignment rules that abstract from student counts (e.g., Bobba et al., 2026) or absolute advantage (e.g., Ahn et al., 2025).

matching gains from assignments to residually well-matched classes. Because the target model is too simple, it will understate the gains or harms from good or bad matches. This simplification will be less costly when match effects are highly correlated across students (within subgroups), but estimating richer value-added will generally enable assignments based on more information about comparative advantage. As such, this first bias will shrink as long as new subgroups capture additional welfare-relevant information such as heterogeneity in effects or welfare weights.

Second, average effects cannot capture *distributional gains* from targeting good matches to students with large welfare weights. A Diamond-style covariance between within-class match effects and student welfare weights characterizes these gains. Any estimate of welfare ignoring this covariance will be biased unless teachers have homogeneous effects or the social planner uses uniform welfare weights. Models with richer subgroup-specific effects will reduce this welfare bias when based on partitions that capture more residual variation in weights and effects. For example, estimating separate effects on students with more versus less academic preparation will likely reduce the bias more than estimating effects on students with earlier versus later birthdays in a given month.

Third, quality of the plug-in estimate depends on the *external validity* of the estimates. This consideration reflects a structural difference between a teacher's average effect, $\bar{\mu}_{j,k}$ (an "average treatment effect" parameter) and the target parameter of value added estimates, $\tilde{\mu}_{j,k}(\mathcal{J}_0)$ (a "treatment-on-the-treated" parameter for the students teacher j actually taught). As a result, effects estimated in one class will not be externally valid for other classes whenever the distributions of match effects vary from class to class.[14] Richer value-added models can also reduce this bias. Consider switching a teacher into a class with no lower-achieving students. If the teacher is particularly good at helping lower-achieving students but only average at helping higher-achieving students, using subgroup-specific value-added scores would be less likely to overstate the gains from the switch.

The fourth term reflects the *shrinkage bias* caused by typical value-added estimates. Empirical-Bayes methods rein in noisy signals by shrinking all estimated effects toward the population mean and other correlated teacher signals such as the teachers' effects on other subgroups or in other years. Although shrunk estimates are less variable, they are also biased such that $\mathbb{E}[\tilde{\mu}^0_{j,k} - \mu^\star_{j,k}|\mu_{i,j}] \neq 0$. This attenuation reduces the variability of predicted welfare.[15] Increasing model complexity can also improve welfare prediction through this channel by "borrowing strength" across correlated signals as more relevant subgroups are

[14] While nonrandom sorting into classes will produce this problem in the limit, idiosyncratic variation from randomization error will do the same in finite samples even under random assignment.

[15] Whether the attenuation also dampens predicted welfare changes depends on the teacher-level covariance of shrinkage bias with class sizes and welfare weights—see Remark in Appendix B.2.

added (Walters, 2024). Although borrowing strength is only *guaranteed* if (1) the true shrinkage function is known and (2) effects are integrated over the full prior distribution—not estimated given the actual realized teacher effects. In practice, however, it is still likely to decrease the variance of estimated welfare.

Unlike the other terms in Equation 2, the *estimation error* term becomes increasingly problematic when evaluating welfare with more complex models. Even after shrinkage, finite-sample value-added estimates will vary. Under standard conditions, this estimation error will not bias the target welfare object, but it will affect the variability of $\widehat{\mathcal{W}}_m^{\mathcal{J}}$ around that target. In contrast to the oracle estimate $\mu^\star$, feasible estimates use estimated student residuals and shrinkage hyperparameters (like the variance and correlation of teacher effects across student groups). Adding more subgroups reduces the number of effective observations identifying hyperparameters and subgroup residuals, and the shrinkage rule passes that variation into the value-added estimates (see full derivation in Appendix B.3). Because they have countervailing effects on welfare variability, the net effect of shrinkage bias and estimation error is an empirical question. To the extent to which there are diminishing returns to borrowing strength over additional subgroups, however, variability will eventually start to increase with model complexity (as found in Ahn et al., 2025, who discuss increasing estimation error in relation to overfitting match effects).

Taken together, these terms of Equation 2 suggest a possible variance-bias tradeoff in choosing which model m to use to predict welfare. Appendix B.3 also characterizes the mean-squared-error of predicted welfare given any model m or assignment policy $\mathcal{J}$. For evaluating a fixed assignment, this tradeoff between bias and variance may matter little, but as we see in the following subsection, these issues compound when plug-in welfare estimates are used to select optimal assignment policies.

2.4 The Second-Best (Empirical) Teacher Assignment Policy

With a way to predict the welfare from any given assignment, we turn to the optimal policy problem. Let $\hat{\mathcal{J}}_m$ denote the assignment rule that produces the highest predicted welfare under model m. In many settings, $\hat{\mathcal{J}}_m$ would be treated as the optimal policy (e.g., Bates et al., 2025; Ahn et al., 2025). In our setting, however, the social planner need not take estimates from model m as given; instead, she may prefer to make assignments using another model, m', that could generate larger welfare gains. In short, we allow the social planner to endogenize model choice.

Model choice introduces a tradeoff between targeting gains from improved model specification and a growing winner's curse produced by estimation error. Because the first-best policy is infeasible, the social planner seeks an optimal second-best policy, $\hat{\mathcal{J}}_{m^*}$, to maximize

expected welfare. Formally, let $m^* = \arg\max_m \mathbb{E}[\mathcal{W}^{\hat{\mathcal{J}}_m}]$, where the expectation is taken over sampling uncertainty that would change the selected assignment rules $\hat{\mathcal{J}}$ and resulting welfare. The tradeoff characterizing this optimum can be described by regret, the gap between realized welfare under $\mathcal{J}$ and the first best (see Kitagawa and Tetenov, 2018; Athey and Wager, 2021). Equation 3 defines expected regret under any model m and decomposes it into two channels that relate to the concepts of model misspecification and estimation error as discussed in the previous subsection:

$$\begin{aligned}\mathcal{R}_m &= \mathcal{W}^*_{\text{First Best}} - \mathbb{E}[\mathcal{W}^{\hat{\mathcal{J}}_m}] \\ &= \underbrace{\mathcal{W}^*_{\text{First Best}} - \mathcal{W}^{\mathcal{J}^*_m}}_{\text{Misspecification Regret}} + \underbrace{\mathcal{W}^{\mathcal{J}^*_m} - \mathbb{E}[\mathcal{W}^{\hat{\mathcal{J}}_m}]}_{\text{Estimation Regret}}\end{aligned} \tag{3}$$

where $\mathcal{J}^*_m$ denotes the model-m oracle assignment, i.e., the assignment made using perfect information about teacher effects over the subgroups in model m. In general, more complex models reduce expected misspecification regret, but increase estimation regret.

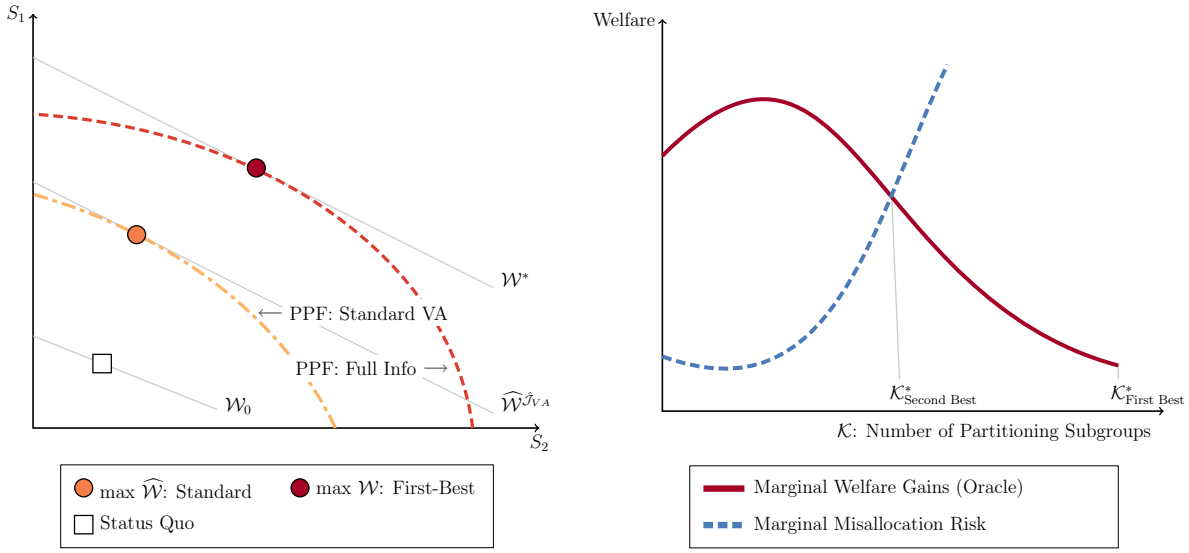
Exploring each component of regret in turn builds intuition for the trade-offs characterizing this second-best policy. First, improving model specification will reduce misspecification regret. To see this, consider a stylized example in which all heterogeneity in effects and weights occurs across two types of students and estimation error is negligible. Panel (a) of Figure 1 presents two curves representing the frontier of possible gains to students of each type—the inner frontier results from assignments based on standard value added, m_{VA}, and the outer frontier from assignments based on heterogeneous value added by subgroup, m^*. Circle markers show the gains from $\hat{\mathcal{J}}_{m_{VA}}$ and $\hat{\mathcal{J}}_{m^*}$. Although an assignment based on m_{VA} increases welfare (from $\mathcal{W}_0$ to $\mathcal{W}^{\hat{\mathcal{J}}_{VA}}$), an assignment based on m^* performs much better because it reduces (or in this stylized example eliminates) misspecification regret. As discussed in Section 2.3, richer models reduce misspecification, and therefore this first component of regret.

Second, increasing estimation error increases estimation regret. Recall that the social planner selects $\hat{\mathcal{J}}_m$ to maximize *predicted* welfare. Because this selection rule conflates true value added with estimation error in $\hat{\boldsymbol{\mu}}_m$, the selected $\hat{\mathcal{J}}_m$ suffers from a "winner's curse": predicted welfare is overly optimistic (see the analogous discussion in Andrews et al., 2024) and produces estimation regret.[16] Increasing the variability of welfare comparisons will, in turn, increase both the probability of overfitting policy to estimation noise and the welfare costs of such policy overfitting. In Appendix B.4, we show that the expected optimism increases convexly with the number of subgroups included in model m. As such, model

[16] While shrinkage does temper the winner's curse relative to using unbiased estimates, feasible Empirical Bayes estimates will still produce positive, increasing optimism—see Remark in Appendix B.4.

complexity raises estimation regret even as it lowers misspecification regret. Furthermore, because this policy overfitting is a distinct issue from predictive overfitting (e.g., see Kitagawa and Tetenov, 2018; Mbakop and Tabord-Meehan, 2021), standard model selection criteria (e.g., AIC, R^2, log likelihood) risk understating these costs from estimation error, leading the social planner to use overly optimistic assignments based on models with too many heterogeneous effects.

Figure 1. Optimal Assignments Trade Off Matching Gains and Misallocation Risk



(a) Infeasible First-Best, Stylized Example **(b)** Feasible Second-Best, Stylized Example

Note: This figure illustrates the welfare gains from using value-added information to make teacher assignments. Panel (a) presents a stylized depiction of the benefits of considering heterogeneity. The two axes present the average outcome of interest, S, for two types of individuals. The graph contains two production possibility frontiers and three iso-welfare curves. The interior production possibility frontier assigns teachers using standard value-added measures and attains social gains of $\mathcal{W}^{\hat{\mathcal{J}}_{VA}}$. By considering both absolute and comparative advantage, the dominant frontier attains the first best, $\mathcal{W}^*$. Panel (b) compares the marginal benefits and costs of considering heterogeneity over the number of subgroup-specific value-added estimates used to make assignments.

Panel (b) of Figure 1 visualizes the trade-off between the costs and benefits of using marginally richer value-added models in another stylized example. The figure plots the marginal welfare effects of assigning teachers using estimates with more and more student subgroups. The benefits come from reduced misspecification regret made possible from making assignments using the (true) subgroup effects. The costs come from increased estimation regret in the form of foregone gains from misallocations based on overfitting increasingly noisy

empirical estimates. The infeasible first-best assignment would use the information from the limiting partition if the match effects were known. In practice, however, the second-best policy must balance the marginal benefits from better matching against the marginal costs from increased misallocation risk.[17]

So long as teacher effects must be estimated, including model selection in the social planner's problem is necessary to achieve the second-best. In a district with tens of thousands of students, the number of possible partitions over which to estimate heterogeneous value added is enormous. Because there are so many ways to estimate teacher effects, previous work on value added often takes this model choice as given—except occasionally as a robustness exercise (e.g., Petek and Pope, 2023; Bates et al., 2025). Ahn et al. (2025) is an important exception, where model choice is carefully considered in the estimation step before exploring counterfactual assignment policies. However, because model overfitting and policy overfitting are conceptually distinct phenomena, endogenizing model choice as part of the counterfactual assignment problem is essential to control the winner's curse and to attain optimal allocations in practice. To this end, we will consider the accuracy and variability of welfare, oracle measures of expected regret across counterfactuals, and quasi-experimental out-of-sample validation rather than relying on in-sample fit to identify optima. As misspecification, winner's curse optimism, and second-best model selection are all empirical questions, we now turn to data and estimation.

3. Estimating Heterogeneous Value Added for Teachers in San Diego Unified

This section sets the groundwork for estimating and characterizing teacher value added. To that end, we describe the data from the San Diego Unified School District, present our estimation strategy and evidence of its validity and robustness, and summarize the descriptive patterns of comparative advantage.

3.1 Background and Administrative Data

We use administrative data on the universe of students in the San Diego Unified School District (2019, SDUSD). The administrative data link teachers to students each semester and contain student demographics and academic outcomes. We identify four outcomes of interest: two cognitive outcomes—standardized scores on math and reading exams[18]—and two noncognitive outcomes—standardized attendance rates and behavior GPAs computed from

[17] This argument is similar in spirit to the idea of Empirical Welfare Maximization (Kitagawa and Tetenov, 2018) and Penalized Welfare Maximization (Mbakop and Tabord-Meehan, 2021), but our setting requires considering high-dimensional treatments (teachers) each with their own heterogeneous effects.

[18] In most years these exams are offered statewide in grades 2–5 or 3–5 (see Appendix Table C.1).

citizenship marks on students' report cards. We provide additional details on measurement and outcome availability in Appendix C.1.

We study the assignment of third- through fifth-grade teachers in 1998–2019. In SDUSD, elementary school principals assign teachers and students to classes, often with the goal of uniformly distributing student social and academic preparation across teachers (as seen elsewhere; e.g., Osborne-Lampkin and Cohen-Vogel, 2014). We sample 3,630 unique teachers who are the main instructors in 18,298 traditional third-, fourth-, or fifth-grade classes from 1998 through 2019.[19] This captures 80.0% of elementary enrollments in SDUSD. We link these teachers to their students using spring-term class rosters and create two samples. Our estimation sample includes students from all years who have two consecutive years of outcomes, and our policy-counterfactual sample includes all students from the 2003–2011 third-grade cohorts (for whom we have second- through fifth-grade test scores). Appendix C.2 contains additional details about sample construction and provides summary statistics.

3.2 Estimation, Identification, and Validation

The administrative data allow us to study optimal teacher assignments. Because optimal policy considerations require estimating different value-added models, this section outlines our general empirical approach. Let each model m be characterized by the $\mathcal{Y}_m$ outcomes it considers and the $\mathcal{K}_{y,m}$ subgroup effects estimated for each outcome—such that value-added estimates will be jointly shrunk across all outcomes and subgroups. For example, standard value added on math scores would have $\mathcal{Y}_m = \mathcal{K}_{1,m} = 1$; math value added between higher- and lower-achieving students would have $\mathcal{Y}_m = 1$ and $\mathcal{K}_{1,m} = 2$, and jointly estimated value added between higher- and lower-achieving students on both math and reading scores would have $\mathcal{Y}_m = \mathcal{K}_{1,m} = \mathcal{K}_{2,m} = 2$. In this paper, we will compare value-added models using different outcomes (math, reading, attendance, etc.) and with different subgroups (race, gender, quantiles of lagged outcomes, and their interactions). To build intuition, the maintained examples in this section discuss models with one outcome and with subgroups split by above- or below-median lagged outcomes.

Our estimation approach is adapted from Delgado (2025) and Bates et al. (2025)—see Appendix C.3 for details.[20] For each outcome y and subgroup k in model m, we estimate

[19] We identify classrooms as traditional if they are not special education classes, mixed-grade classes, and are within the 2.5th to 97.5th percentile of class size (13–35 students). See details in Appendix C.1.

[20] We choose to follow Delgado (2025) and Bates et al. (2025) rather than Gilraine and Pope (2021) and Ahn et al. (2025) because the latter requires pooling observations over years to estimate the higher-dimensional effects and drift in teacher effects over time is a quantitatively important consideration (Chetty et al., 2014a).

the following regression (suppressing parameter dependence on m on intermediate steps)

$$y_{i,t} = \alpha^{y}_{\mathcal{J}(i,t),k,t} + \beta^{y}_{k} X_{i,t} + v^{y}_{i,t} \tag{4}$$

where $\alpha^{y}_{\mathcal{J}(i,t),k,t}$ are outcome-specific teacher-by-subgroup-by-year fixed effects (essentially nuisance parameters at this stage for estimating β), and $X_{i,t}$ are student observables including student ethnicity, gender, age, lagged suspensions and absences, indicators for special education and English language learner status, and grade-specific cubics of all four lagged outcomes or missing flags.[21] Appendix C.5 shows that results are not sensitive to alternative specifications, such as those in Chetty et al. (2014a).

After estimating Equation 4, we compute average residuals in each model by teacher, outcome, subgroup, and year in three steps. First, we subtract estimated effects of observed student characteristics, $\hat{\beta}^{y}_{k} X_{i,t}$ from each outcome $y_{i,t}$. Second, for each outcome we project these intermediate residuals onto teacher fixed effects $\alpha^{y}_{\mathcal{J}(i,t)}$, a teacher-experience profile $f(y, z_{\mathcal{J}(i,t),t})$, school fixed effects $\phi^{y}_{\ell(i,t)}$, and year fixed effects (for normalization). Third, we use $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{f}())$ to obtain average residuals by teacher, outcome, subgroup, and year:

$$\bar{A}_{y,j,k,t} = \frac{1}{n_{j,k,t}} \sum_{i:\mathcal{J}(i,t)=j,k_i=k} \left[y_{i,t} - \hat{\beta}^{y}_{k} X_{i,t} - \hat{\phi}^{y}_{\ell(i,t)} - \hat{f}(y, z_{\mathcal{J}(i,t),t}) \right]$$

We compute shrinkage-adjusted estimates for each teacher's value added using their average residuals in prior years. After stacking the residuals from all outcomes and subgroups from prior years into a vector, $\boldsymbol{A}^{-t}_{j}$, we estimate shrinkage-adjusted value added and add back experience.

$$\hat{\mu}^{y}_{j,k,t} = \hat{\boldsymbol{\psi}}^{y}_{j,k} \boldsymbol{A}^{-t}_{j} + \hat{f}(y, z_{j,t}) \tag{5}$$

where $\hat{\boldsymbol{\psi}}^{y}_{j,k}$ are reliability weights mapping residuals from all outcomes and subgroups into value added on y and k. As such, Equation 5 shrinks each teacher's value-added scores toward their value added in other years, subgroups, and outcomes; toward the value added of other teachers; and allows for drift over time.[22] For each model, we then implement this procedure in 1,000 bootstrapped samples stratified by class to non-parametrically characterize the joint distributions of value added.

We make two extensions to Bates et al. (2025) to deal with multidimensionality and high-dimensional heterogeneity. First, we shrink value-added estimates across both the subgroups

[21] Following Petek and Pope (2023) we use outcomes in year t and lagged scores from year $t-1$ for cognitive outcomes and outcomes from year $t+1$ and lagged outcomes from year $t-1$ for noncognitive outcomes (thus preventing grading stringency from entering value added).

[22] As in Bates et al. (2025) we set a drift limit of 7 years.

and the outcome domains considered in each model. This relaxes the implicitly maintained assumption in Chetty et al. (2014a) and Bates et al. (2025) that there is no covariance between the idiosyncratic components of teacher value added across outcomes. For example, in a model of math value added with two achievement subgroups, we would use residuals from both subgroups to predict the math value added for each subgroup (just as in Bates et al. (2025)); however, unlike other papers, in a model of both math and reading, we would use residuals from (all subgroups of) both math and reading to predict the math and reading value added of each teacher on each subgroup.[23] Second, because we consider much higher-dimensional models than Bates et al. (2025), we add a Ridge-type regularization to keep the reliability weights, $\boldsymbol{\psi}$, well-behaved in high-dimensional cases. Appendix C.4 formalizes this process and shows that measured welfare gains are not sensitive to this decision.[24]

Interpreting our estimates as causal effects requires that individual- and class-type-level shocks are conditionally independent of teacher assignment. Students and teachers are certainly not randomly assigned, but rich information about lagged outcomes seems to capture the key unobserved determinants of outcomes (as confirmed in Chetty et al. (2014a)). This is why every regression flexibly controls for grade-specific cubics in all four lagged outcomes and includes school fixed effects. Note that with school fixed effects in our model, value added is identified for all teachers through a dense network of quasi-experimental changes in subgroup-specific test scores when teachers switch schools.[25]

As additional evidence consistent with conditional independence we document precise cross-sectional forecast unbiasedness. Following Chetty et al. (2014a) and Bates et al. (2025), we regress class-level residuals on the value added predicted from previous years and show that the coefficient is close to 1. Appendix Figure A.1 plots class residuals over ventiles of subgroup-weighted value-added estimates. All outcomes have slopes very close to 1 with tight standard errors (cluster-corrected at the teacher level). Appendix Figure A.2 depicts the full distribution of cross-sectional forecast coefficients across all models. The coefficients range from 0.98 to 1.19, with an average of 1.004—comparable to Chetty et al. (2014a) and Aucejo et al. (2022) and slightly closer to one than both Bates et al. (2025) and Delgado (2025).

Next, because optimal policy considerations require evaluating counterfactual assignment policies where teachers are assigned to different schools, we also examine teacher switches to

[23]It turns out that after this generalization the value-added approach of Mulhern and Opper (2021) becomes a special case of our empirical method.

[24]Regularization does not impact welfare because second-best optima use relatively sparse models whereas regularization has meaningful effects only in relatively higher-dimensional models—that also have larger expected optimism and misallocation risk.

[25]Appendix C.5 replaces teacher-by-year and school fixed effects with teacher effects and cubics of class and school averages of all lagged outcomes, class means of demographics, class size, and grade indicators.

further explore our estimates' external validity. Following Chetty et al. (2014a) and Gilraine and Pope (2021) we exploit the quasi-experimental variation induced by teacher moves across schools. We regress the year-to-year change in average residuals on the predicted change generated by teacher arrivals and departures. Table A.4 reports small quasi-experimental forecast bias for academic outcomes that increases with model complexity, but noncognitive value-added measures tend to perform poorly.[26] This pattern motivates our decision to focus on multidimensionality in cognitive value added as our main specification in Section 5.

Because average forecast unbiasedness could mask distributional differences, we also examine teacher switches to further explore our estimates' external validity. Although our forecast unbiasedness already suggests external validity, Appendix Figure A.3 also compares school-grade-level changes in value added based on the change in class size and composition experienced by the quasi-experimentally switching teacher. On the one hand, if switching to a larger class is more challenging, reallocating high value-added teachers to these classes would causally reduce their value added, leading us to overstate the gains from reassignment. This pattern would suggest a strong negative relationship between absolute advantage and class size and class composition; however, Panel (a) shows no evidence of this. On the other hand, if switching to teach classes with more concentrated composition enables teachers to specialize more effectively, we would understate the gains from reassignments. This pattern would suggest a strong positive relationship between value added and composition; however, Panel (b) shows extremely precise null relationships.[27] Together, these relationships suggest that our estimates reflect fundamental, school-invariant teacher characteristics with strong external validity in counterfactual assignments.

3.3 Heterogeneity Highlights the Importance of Comparative Advantage

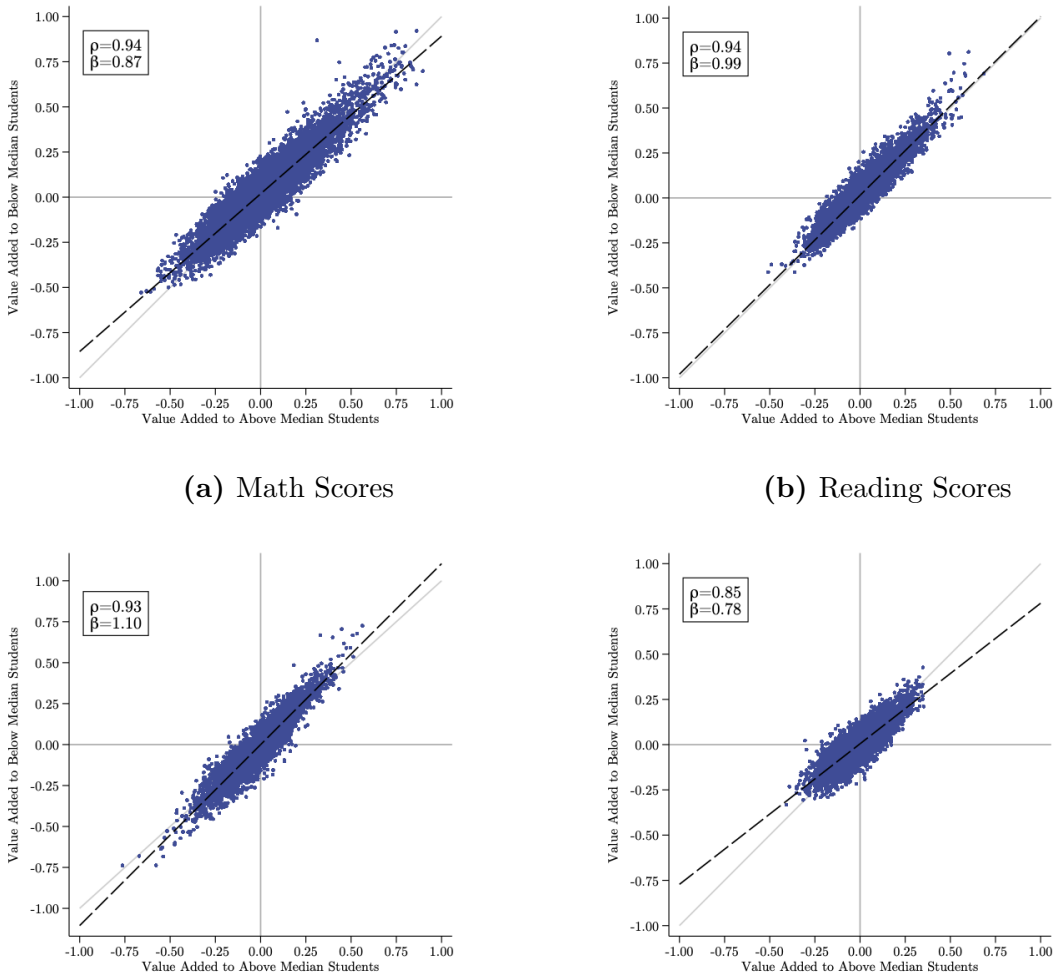
Using the procedure outlined above, we estimate the heterogeneous value added of the 4,000 teachers in our sample. Figure 2 presents a scatter plot of value added for each outcome (math, reading, behavior, and attendance) when students are grouped by lagged outcomes. Each point represents one teacher-year observation where value added on students with below-average lagged outcomes is plotted on the y-axis against value added on students with above-average lagged outcomes on the x-axis (both measured in student standard deviations). Each plot also presents the correlation coefficient between the value added for the

[26]Interestingly, Petek and Pope (2023) report similarly weak (and often opposite-signed) results in the relationship between their analogous measure of noncognitive value added ("learning skills") and long term outcomes in an analogous quasi-experimental exercise (the cleanest comparison between our behavioral GPA results and the "learning skills" results in their Appendix Table A12).

[27]These patterns are also consistent with evidence that teachers adapt teaching practices little across classrooms with different levels of prior achievement (Aucejo et al., 2022).

two subgroups, as well as a slope coefficient for the line of best fit between the two. Appendix Table A.1 reports all of the standard deviations and the full correlation matrix for these estimates.

Figure 2. Value Added Varies Both Within and Across Teachers



(a) Math Scores

(b) Reading Scores

(c) Behavior GPA

(d) Attendance

Note: This figure shows our heterogeneous estimates of teacher value added on reading scores, math scores, behavior GPA, and attendance (each measured in student standard deviations). Each dot represents one teacher-year estimate of value added on students with above- or below-median lagged outcomes. Each correlation coefficient, ρ, is for the entire population of teacher-year observations. The dashed line shows the line of best fit with the slope β reported. For reference, a gray line with slope one is plotted in the background.

Figure 2 depicts meaningful differences in value added within *and* across teachers. Absolute advantage can be seen in the dispersion of teachers along the 45-degree line. Teachers

above and to the right generate larger gains than teachers below and to the left. Note that value added to math is much more dispersed (0.20 student standard deviations, σ) than value added to reading (0.13σ), behavior GPA (0.17σ), and attendance (0.10σ). Comparative advantage can be seen in the dispersion away from the 45-degree line. Teachers above and to the left have a comparative advantage in teaching lower-scoring students and teachers below and to the right have a comparative advantage teaching higher-scoring students. After shrinking across the two subgroups, correlations are high.[28] This means the average comparative advantage is modest, but it is not insignificant—Appendix Table A.1 shows that the mean absolute comparative advantage is 26–44% of the standard deviation in absolute advantage.[29] Appendix Table A.3 shows that we reject homogeneity across 1,000 class-stratified bootstraps using a standard difference-in-means t-test for 26% of teachers for math, 22% for reading, 31% for behavior, and 6% for attendance. Not only is there meaningful comparative advantage, but Appendix Table A.3 shows that comparative advantage is persistent across years: teachers who are at least a standard deviation better at teaching one subgroup display that same comparative advantage in 63–79% of subsequent observed years, depending on the outcome.

These patterns underpin the tensions in the teacher-assignment problem. On the one hand, there is significant dispersion in absolute and comparative advantage, so there will be gains to assigning teachers to classes based on class size and composition. At the same time, however, the high cross-subgroup correlations suggest that there may be swiftly diminishing marginal returns to considering richer and richer heterogeneity—especially given misallocation risk. We measure these returns and characterize the optimal second-best assignment in Section 4. Furthermore, because cross-outcome correlations tend to be much lower (see Appendix Table A.1), considering comparative advantage has the potential to generate even larger gains when the objective function includes multidimensionality—as will be explored in Section 5.

[28]The correlations are similar to those by socioeconomic status in North Carolina (0.9 for math in Bates et al., 2025) and lower than those by race in Chicago (0.97 for math in Delgado, 2025). Interestingly, when we estimate effects with additional subgroups, effects are most highly correlated for adjacent groups. For example, math value added on top-quartile students is much more correlated with value added on third-quartile students than with value added on first-quartile students (see Appendix Table A.2).

[29]This is about three times larger than the difference attributable to matches on observable student and teacher characteristics (Laverde et al., 2026) and slightly less than in a rich, saturated model with heterogeneity by 12 student characteristics (Ahn et al., 2025). As in Bates et al. (2025), we reject the null of perfect cross-subgroup correlation at the $[p < 0.001]$ level.

4. Efficiently Assigning Teachers to Classes

We now consider the assignment of teachers to classes by incorporating our value-added estimates into our theory. This section defines the empirical assignment problem, identifies the optimal assignment for maximizing average math scores, and assesses equity by trading off gains to higher- and lower-achieving students. To build intuition about the trade-offs related to heterogeneity, this section focuses only on math scores, but we consider multidimensionality in Section 5.

4.1 Optimization Problem and Solution

Consider a social objective $\tilde{\mathcal{W}}$ that places welfare weights ω_k on K groups of students. Recall that $\mathcal{J} : (i,t) \to j$ is an assignment function, telling us which teachers teach each student in each year. We consider a set of candidate assignments, $\mathscr{J}$, and define the following optimization problem (with subject subscripts suppressed):

$$\max_{\mathcal{J}\in\mathscr{J}} \tilde{\mathcal{W}}(\mathcal{J};\boldsymbol{\omega},\boldsymbol{\mu}_m) = \max_{\mathcal{J}\in\mathscr{J}} \frac{1}{N} \sum_{t=2003}^{2013} \sum_{k=1}^{K} \sum_{i:\, k_{i,t}=k} \omega_k \, \hat{\mu}^{\text{math}}_{m,\mathcal{J}(i,t),k} \tag{6}$$

where each $\omega_k \in [0.0, 1.0]$ represents the welfare weight on students in a given subgroup (such that $\sum_k \omega_k = 1$), and $\hat{\mu}_{j,k}$ are estimates of teacher j's value added on type k students. Recall that the reassignment sample is limited to students in grades 3–5 in 2003–2013 and that we restrict $\mathscr{J}$ to assignments that hold classes fixed to avoid introducing peer-effect biases into our welfare estimates. Our main results study district-wide reassignments in which a teacher from a given class can teach any same-grade class in the district, but we also analyze within-school reassignments where no teachers change schools or grades.

Four implications follow from Equation 6's formulation of welfare. First, in this static assignment problem, the social planner takes estimates as given and tries to maximize scores. Dynamics introduce an explore-versus-exploit trade-off in which the social planner may try to make early assignments to maximize information gain rather than only maximizing achievement gains. This economically interesting policy is likely to be more susceptible to manipulation than assignment rules based on pre-policy data. Second, this formulation assumes assignment $\mathcal{J}$ is respected. This is analogous to a partial-equilibrium interpretation in which the policy does not lead students to re-sort across classes (via requests), schools (via school choice), or districts (via in- or out-mobility). Third, because we do not change class composition, gains could be larger in districts with more class-level tracking (which increases the variance in class composition). Finally, district-wide reassignments might be practically

infeasible. For example, some assignments could be incentive incompatible given teacher preferences for locations and schools (Boyd et al., 2005; Johnston, 2025), and others could be in tension with state or union policies. In this case, the gains from reassignments highlight the shadow value of relaxing such constraints—something we study directly in Section 5.2.

We solve Equation 6 by linear programming. The district-wide reassignment problem has approximately 10^{690} assignments to search over each year and cannot be solved by heuristics such as assigning the best teachers to the largest classes or highly specialized teachers to classes with well-matched demographics (see Section 2). Instead, we characterize and solve Equation 6 as a (mixed-integer) linear programming problem (Bertsimas and Tsitsiklis, 1997)—see Appendix D.1 for details.

Two complications affect how we solve and analyze solutions to Equation 6 in practice. First, some teachers may never teach some types of students. While not an issue for standard (homogeneous) value added, this concern may become particularly pronounced in higher-dimensional models with more finely partitioned subgroups. We solve this by imputing Empirical Bayes predictions of unidentified effects. Appendix Table A.5 reports results separately by whether the social planner reassigns teachers with imputed value-added scores. Second, unlike standard linear programming problems, the teacher-assignment problem features uncertainty in the model parameters. To address this, we use robust optimization (see Gabrel et al., 2014, for an overview). This approach is a maximin optimizer that chooses an assignment with the highest gains under left-tail realizations of the parameters.[30] While we prefer the robust estimates, Appendix Table A.5 also reports results from standard optimization.

Once we have solved Equation 6, accurately reporting gains requires two more steps. First, to make "honest" out-of-sample predictions, we deflate the gains from each model by that model's out-of-sample quasi-experimental forecast bias (similar to the cross-fitting approach to honesty in Athey and Wager (2021)). This correction is important because the quality of any assignment depends on the quality of the value-added estimates, and Section 3 revealed that—despite being strong for all math models—forecast accuracy falls as models become richer. Empirically we estimate the quasi-experimental forecast parameter for each model, $\hat{\beta}_m$, and use it to deflate the predicted welfare for policies $\mathcal{J}$ evaluated with model m. We use the quasi-experimental forecast bias because it most closely reflects the policy counterfactual of interest: switching teachers across schools. Cross-sectional forecast bias is both more distant theoretically from the policy counterfactuals and is inadmissible for

[30]Because the social objective is monotonically increasing in each teacher's value added, we choose left-tail realizations in practice by using our 1,000 bootstraps to characterize the 5th empirical percentile of each value-added estimate for each teacher. We then find the best assignment using those (pessimistic) estimates.

deflation because it is directly targeted in the regularization of high-dimensional models.

Second, instead of reporting naive predicted gains, we report optimism-corrected welfare estimates (similar to the correction in Mbakop and Tabord-Meehan (2021)). As mentioned in Section 2, a winner's curse makes predicted gains of the selected assignment overoptimistic. We quantify optimism by measuring the average distance between the predicted gains from assignment $\hat{\mathcal{J}}_m$ and the gains under each of 1,000 bootstrapped value-added estimates. This process is computationally expensive as it requires estimating 1,000 bootstrapped estimates for each of our 38 models,[31] finding the optimal assignments for every welfare weight, and correcting the predicted gains with each set of bootstrapped estimates, but it is necessary to make accurate comparisons between the models. Because this approach is equivalent to integrating gains over the joint distribution of teacher effects, we refer to these gains as "expected gains," and report $\widehat{\mathcal{W}}^{\mathcal{J}} = \mathbb{E}_{\hat{\boldsymbol{\mu}}}\left[\hat{\beta}_m \mathcal{W}_m^{\mathcal{J}} | \omega\right]$.[32]

The comparison between naive predicted gains and forecast-deflated expected gains also allows us to quantify optimism empirically. Appendix Figure A.4 plots both series, illustrating how predicted gains create an illusion of convex returns to model complexity that is entirely driven by optimism. These two steps control the winner's curse and allow us to accurately and honestly evaluate welfare and compare models.

4.2 Gains from (Second-Best) Optimal Assignment Policies

We first consider the second-best problem for a social planner trying to maximize average math scores. In Equation 6, this objective implies equal (utilitarian) welfare weights on all students. Finding the second-best assignment requires two steps. First, the econometrician estimates multiple value-added models so the social planner can identify the best assignments using each model. Second, the social planner compares the gains and risk of using each model to determine the optimal policy.

4.2.1 Maximizing Math Scores

We begin by considering possible assignments made by using different models of value added on math. We consider heterogeneity across up to 10 lagged achievement quantiles, reported gender (female versus male), reported race and ethnicity (Black or Hispanic versus other race/ethnicity), and interactions. Recall that in the value-added estimation step, we did not

[31] 14 models for math, 14 for reading, 2 for behavior, 2 for attendance, 5 for math + reading, and 1 combining math + reading + behavior + attendance.

[32] Note, we estimate optimism and expected gains using the bootstrapped estimates $\hat{\mathbb{E}}_{\hat{\boldsymbol{\mu}}}\left[\mathcal{W}^{\mathcal{J}} | \omega\right] \equiv \frac{1}{B}\sum_b \sum_j \sum_k n_{j,k}(\mathcal{J})\ \bar{\omega}_{j,k}(\mathcal{J})\ \hat{\mu}_{m_b,j,k})$. We prefer the bootstraps to direct integration of Bayesian posteriors because a Shapiro-Wilk normality test rejects normality of the bootstrapped distributions at the 0.05 level for 55% of the teacher-year-subgroup estimates.

include students with missing test scores (lagged or actual), grade repeaters, and students who were absent for over 50% of the year. We do not drop these students from the reassignment step, however. Instead, we impute lagged achievement quantiles using analogous quantiles from other years where available and school-level averages where necessary (see Appendix D.1). This imputation is important for two reasons. First, the distribution of missing scores is not uniform across classrooms. As such, dropping students with missing scores before solving the assignment problem disproportionately reduces class sizes in lower-achieving schools, artificially inflating the gains from reassigning better teachers to larger classes and redistributing gains from lower-achieving schools toward higher-achieving schools. Second, in the real world, teachers will be assigned to teach all students—whether or not their prior achievement is known. Our approach provides a practical way to keep those students in the assignment problem.

Figure 3 presents the expected gains from robust reassignments based on value-added estimates from nine models with varying complexity.[33] Bars depict the expected achievement gains from an optimal assignment relative to the status quo. We report the gains as the average cumulative effect over grades 3–5 (see Appendix D.2 for details). Each 95% confidence interval comes from the 1,000 reevaluations of the robust assignments using forecast-deflated bootstrapped estimates. We also use these bootstrapped estimates to test the (one-sided) null hypothesis that the gains from each model are no larger than the gains from every other model and report p-values corresponding to the share of bootstraps in which the richer model produces smaller expected gains than the next most simple model.[34] All results are for the district-wide reassignment, but Appendix Table A.5 contains analogous within-school results. Furthermore, Appendix Table A.6 shows analogous results for reassignments maximizing reading scores and lifetime earnings.

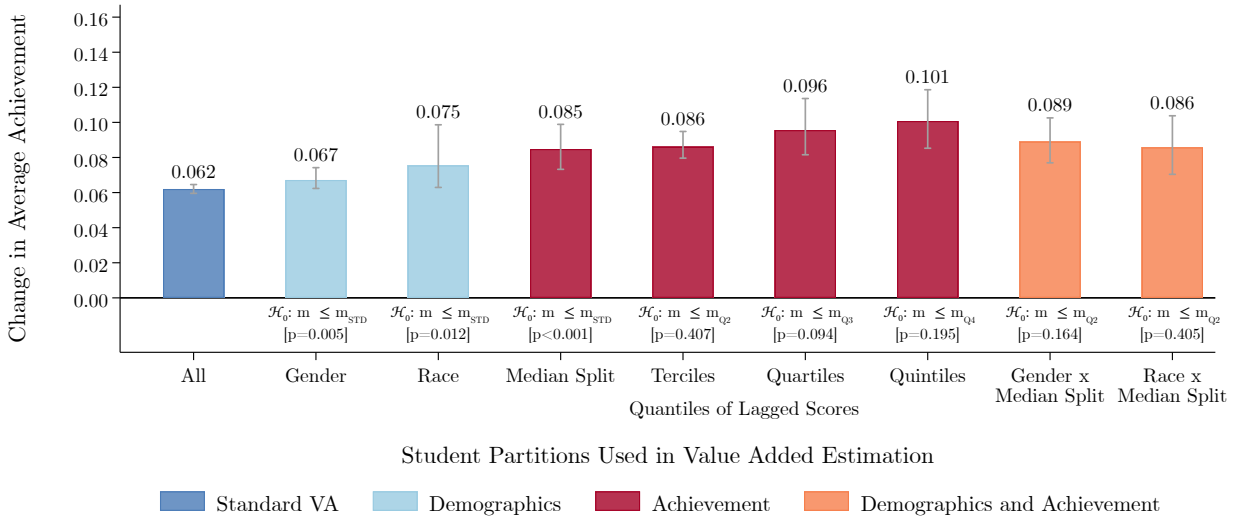
Two main findings emerge from Figure 3. First, there are substantial gains from using information about both absolute and comparative advantage; in each case, reassignments attain large gains. Simply using standard value added to put the best teachers into the biggest classes would produce expected gains of 0.06 standard deviations per student, 30% larger than a benchmark teacher "deselection" program.[35] In other words, policymakers who prefer not to fire dozens of teachers—or who cannot do so given collective bargaining agreements—could instead improve outcomes by reassigning these teachers to classes

[33]Appendix Figure A.4 contains the expected gains from the higher-dimensional models not included in Figure 3.

[34]This corresponds to models with one fewer quantile for achievement-based partitions and models without race and gender for demographic-based partitions. The p-values of all pair-wise model comparisons are reported in Appendix Table A.7.

[35]i.e., firing the worst 5% of teachers (Hanushek, 2009, 2011; Chetty et al., 2014b).

Figure 3. Reassignments Yield Large Gains—Especially when Using Lagged Achievement



Note: This figure shows the effects of new teacher assignments using different estimates of teacher value added on math scores. Each assignment uses the estimates of teacher effects to solve the problem in Equation 6 by robust optimization with the empirical distribution of teacher effects in 1,000 bootstrapped samples. We report the expected gains to student math scores under each model over grades 3–5 from the robust assignment by averaging gains from the joint distribution of 1,000 bootstrapped estimates, accompanied by empirical two-sided 95% confidence intervals. The figure also reports tests of the (one-tailed) null hypotheses that each assignment produces gains no larger than the assignment from the next most simple model. The reassignment sample includes students in the third-grade cohorts of 2003–2011.

where they do less harm and by assigning better teachers to larger classes (ideally with commensurate compensation). Using richer value-added estimates generates 8%–72% larger expected gains, indicating that information about comparative advantage can be almost equally important. But, as predicted in Section 2, there also seem to be diminishing returns to considering additional dimensions of heterogeneity, suggesting that the optimal policy will likely be implemented using a relatively simple model.

Second, estimating value added using achievement-based partitions captures the most welfare-relevant information. Given the frequent attention paid to heterogeneous teacher effects by gender or race (e.g., Dee, 2005; Delhommer, 2022; Delgado, 2025), the second series of bars in Figure 3 presents the gains from making assignments on these dimensions. Using these estimates does improve assignments by 10–20%, but value added based on quantiles of lagged achievement increases gains by 35–75%. Not only are the gains from considering lagged achievement larger, but conditional on achievement, the marginal impact of consider-

ing additional demographics is even smaller. For example, using the interaction of gender or race and above/below median prior-achievement generates 1–5% improvements over just the above- or below-median split, compared with 13% improvements from using four achievement quartiles. This concentration of welfare-relevant information in achievement-based partitions is especially relevant in practice since explicitly using race or gender in assignment decisions could raise substantial legal concerns. These results suggest that the core information for the optimal second-best assignment comes from using lagged test scores.

The fact that policy-relevant heterogeneity loads on lagged achievement is consistent with empirical research on differentiated instruction and accountability. For example, because teaching higher- and lower-achieving students requires different skills (see Betts (2011), Duflo et al. (2011), Small (2012), and Tomlinson (2017)), pre-service and in-service training programs intentionally emphasize differentiated instruction. In our theoretical framework, the best partitions capture the most variance in match effects; to the extent that differentiated instruction increases such variance, these training programs could increase gains from reassignment. Similarly, many education and accountability policies focus on the proficiency of lower-achieving students—even while expressing nondiscriminatory, identical preferences for students of different genders, races, and socioeconomic statuses. These policies can even cause teachers to allocate effort in ways that amplify their heterogeneous effects (Neal and Schanzenbach, 2010; Macartney et al., 2021) and also serve as a natural contributing source of the empirical patterns we document here.

While informative in many ways, the information in Figures 3 and A.4 are insufficient for us to compare models to choose the second best. The figure depicts expected gains from reassignments, but while many models produce similar gains, they may have very different amounts of misallocation risk. To identify the second-best policy, we need to define model selection criteria that trade off matching gains against the costs of estimation error.

4.2.2 Identifying the Second-Best Assignment Policy

To identify the optimal feasible policy, we need to compare assignments made using different value-added models. The theory in Section 2 states that the optimal policy balances expected marginal benefits from model specification (more matching gains) against marginal costs of estimation regret (more misallocation risk). In theory, the model with the highest forecast-deflated expected gains should satisfy this condition, but in practice there may be multiple models with comparable expected gains but different misallocation risk. This subsection outlines our preferred approach to model selection in such cases. Because our preferred criterion identifies a relatively sparse m^*, we show robustness to alternative approaches to model choice and discuss the sensitivity of our assignment results to other models.

Our main selection criterion focuses on expected gains and breaks ties using hyperparameter quality as a measure of misallocation risk. We proceed in four steps: (1) Calculate the forecast-deflated expected gains for every candidate model, (2) identify the model with the highest expected gains, $m_{\max}$, (3) assemble a set, $\mathcal{M}$, of all models with statistically indistinguishable gains from $m_{\max}$, and (4) select m^* as the model with the "best-behaved" hyperparameters within $\mathcal{M}$. We break ties with hyperparameter quality because Appendix B.4 identifies hyperparameter estimation as the key source of optimism and misallocation risk. We consider two measures of hyperparameter quality: the share of the estimated (auto)-covariance that is monotonic (assuming covariances decrease in subgroup distance and time) and the empirical distance from subgroup stationary (assuming equidistant covariances tend to be similar). See formal definitions in Appendix E.1.

Applying this approach identifies the lagged-math quartiles as the optimal model to use. In our estimates, heterogeneity across septiles and heterogeneity across deciles produce the highest forecast-deflated expected gains (tied at 0.107 standard deviations per student). Using a standard $\alpha = 0.05$,[36] the set $\mathcal{M}$ includes lagged math quartiles through deciles but not standard value added, above/below median, terciles, or any model with race or gender (see Appendix Table A.7). Appendix Table E.1 reveals that value added based on quartiles of lagged achievement has the best hyperparameter quality. Among subgroups and time periods, only 5.3% of the total covariance is non-monotonic (i.e., stronger between more distant time periods or subgroup pairs), compared to 20+% for the richest models. Although not necessary for a covariance matrix, quartiles also produces more stationary correlations with about half as much deviation from that benchmark.

Other model selection criteria also point to quartiles. For example, although it is an implicit piece of our model selection approach recall that quasi-experimental forecast bias quickly deteriorates with model complexity—with deciles generating 3.7× more bias than quartiles in our preferred specification (Appendix Table A.4). We also consider three other selection criteria. First, given the fact that optimism and misallocation risk are increasing in model complexity (Proposition 4 in Appendix B.4), a social planner with any risk aversion will select the simplest model from any equivalence class—generally lagged quartiles. Second, quantifying the welfare MSE derived in Appendix B.3 shows that either quartiles or sextiles make welfare predictions with the lowest MSE among that same equivalence class (see Appendix E.2). Finally, in Appendix E.3, we consider a more general loss function that penalizes models based on their distance from the first best and their empirical variance.

[36] Appendix Figure A.5 shows that model selection is broadly stable over a wide range of α for both math and earnings value added and across alternative inference approaches, including using bootstrapped $\hat{\beta}_m$ to account for additional uncertainty in the forecast deflation step.

This (implicitly loss-averse) criterion also identifies quartiles or quintiles as optimal. Independent of the model selection mechanism, in Appendix Figure A.6 we show that average gains are not sensitive to model selection in the neighborhood of m^*—and neither are our results about economic incidence discussed in the following subsection.

4.3 Restructuring Achievement Equity

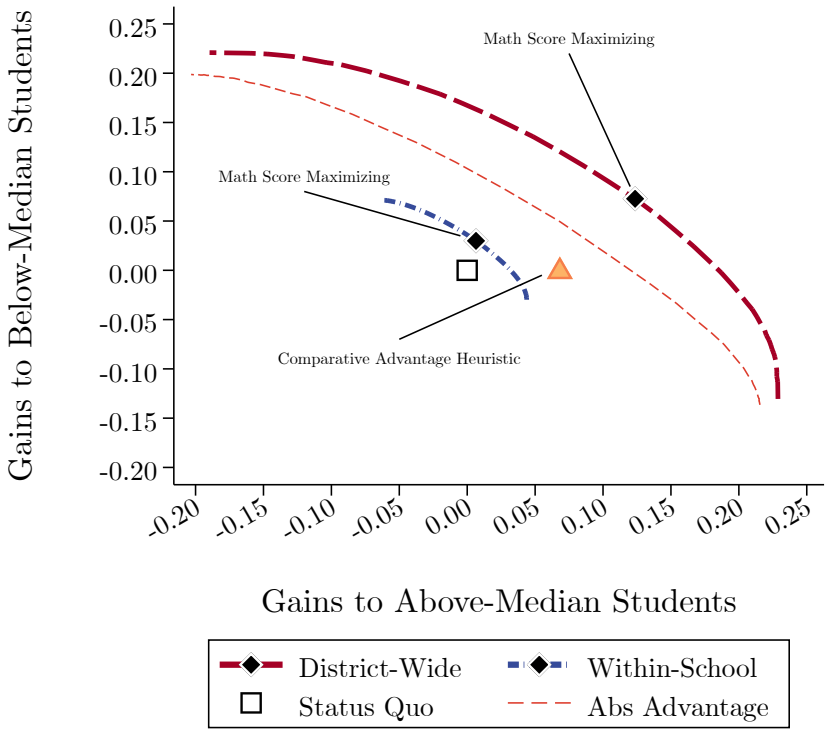
Whereas the previous results all focused on maximizing average scores, revealed preference suggests that policymakers also have distributional concerns. To operationalize this idea, we vary the welfare weights on students with below- versus above-median prior-year math scores, ω in Equation 6, from 0.0 (only care about above-median students) to 1.0 (only care about below-median students).[37] The changes in average subgroup scores characterize the trade-offs between helping students in each group—as in the stylized PPF in Figure 1. This is the frontier of second-best gains described by the theory in Section 2.

We depict the expected math score gains from four sets of policies in Figure 4. Changes to lower-achieving students' scores on the y-axis are plotted against changes for higher-achieving students on the x-axis. Gains above and to the right of the status quo (square marker) are always preferred by the social planner. The largest (red) and smallest (blue) PPFs depict the expected gains from district-wide assignments and within-school assignments using value-added estimates that vary by quartiles of lagged math achievement. Black diamond markers denote the math-score-maximizing assignments. We compare these PPFs with policies that make district-wide assignments based on absolute advantage using only standard value added, the central (orange) PPF, or only comparative advantage, the (yellow) triangle marker. For comparability, the gains from all assignments are quantified under the second-best model based on achievement quantiles.

Consider three main insights from Figure 4. First, the PPFs illustrate the potential for accomplishing distributional objectives. In this visualization, the curvature of the PPF indicates the scope for a policy to improve one subgroup's scores while not harming the other group on average. For example, the district-wide assignments using value added by math quartiles (red long-dashed PPF) curve outward much more than those using standard value added (orange dashed PPF). Considering comparative advantage significantly lowers the "price" of addressing achievement gaps because the social planner can leverage match effects to raise achievement for both groups rather than simply putting better teachers in classes with more lower-achieving students. In fact, a policymaker using value added by achievement

[37] Of course, the social planner could choose welfare weights that vary along different dimensions. We begin with achievement because of revealed preference seen in accountability policy. Furthermore, there is no requirement that the dimension of heterogeneity align with the dimension of welfare weights, although they often may in practice when choosing the second-best.

Figure 4. Optimal Assignments Can Achieve Substantial Redistribution



Note: This figure shows the expected test score gains from optimal (robust) assignments relative to the status quo. Three production possibility frontiers are presented: two optimally reallocating teachers across the district or within schools (both within grade) using heterogeneous value added by quartiles of lagged math scores, and one using standard value added. Each PPF is constructed by solving Equation 6 with various welfare weights on lower- and higher-scoring students, $\omega_k \in [0.0, 1.0]$. Diamond markers show the solutions from placing equal weight on all students. The triangle marker shows the gains from a heuristic policy that reassigns teachers across the district using comparative advantage on above- and below-median students ignoring class size.

quantiles could raise lower-achieving students' scores by 0.17 standard deviations without harming higher-achieving students on average, compared with 0.11 standard deviations using standard estimates.

Interestingly, although one rationale for using comparative advantage is to close achievement gaps, our results show that ignoring information about absolute advantage completely undermines this objective in SDUSD. The triangle marker in Figure 4 shows the results from a heuristic assignment policy that places teachers with the largest comparative advantage in teaching lower-achieving students in classes with larger shares of lower-achieving students (as proposed in Section 2.1 and in Delgado, 2025). While this policy does produce meaningful average benefits, in our setting they accrue only to higher-achieving students. On the other

hand, the second-best produces three times larger gains, and (relative to the heuristic) the majority of gains come from lower-achieving students.[38]

Second, in addition to helping lower-achieving students, achievement-based assignments could substantially reduce racial achievement gaps. For example, a race-blind district-wide assignment with 74% weight on lower-achieving students would shrink the racial achievement gap by 0.11 standard deviations (17%) by fifth grade without harming either underrepresented minorities (Black and Hispanic students) or other students. Because racial groups are fairly segregated across schools, however, there is little scope for reducing racial gaps using within-school assignments.[39] Interestingly, in addition to being legally feasible, policies in the frontier of second-best gains dominate many race-focused assignments (as in Delgado, 2025). Appendix Figure A.7 depicts the PPFs of race-conscious assignments, showing that for each assignment that improves both groups' scores, there is a strictly dominant version of the race-blind second-best policy based on achievement quartiles.[40]

The relevance of achievement heterogeneity to educational incidence on other dimensions has broader implications for our understanding of teacher effects and matching. For example, our results add nuance to prior research on "match effects" between students and teachers sharing observable characteristics like gender or race (e.g., Dee, 2005; Delhommer, 2022; Laverde et al., 2026). While these role-model effects are certainly important, this assortative match effect is only one component of comparative advantage in general. In our context, differentiation along the test-score distribution explains much more than demographic match, and, as shown in Figure E.3, including demographics can double expected MSE. More broadly, this pattern illustrates the importance of examining the key determinants of outcomes when estimating match effects—as in Dahlstrand (2022) with medical risk and Arnold et al. (2022) with criminal proclivity.

Third, Figure 4 also reveals that maximizing average scores does not have uniform incidence across student groups. This relationship can be seen visually by comparing the diamond markers to a 45-degree expansion path from the status quo. For example, the optimal within-school assignment benefits lower-achieving students a good deal more than higher-achieving students (0.028 versus under 0.006); while the optimal district-wide assignment

[38]The relative performance of the heuristic and the second best will be context dependent. In our setting, the reason the heuristic does not help lower-achieving students is because teachers with a comparative advantage at teaching lower-achieving students tend to have slightly lower absolute advantage (see Figure 2) and because classes with high shares of lower-achieving students tend to be smaller and so do not necessarily have more lower-achieving students overall.

[39]The relative value of estimating value added by demographics will depend on contextual factors like the distribution of achievement, demographics, and value added within and across schools.

[40]Unconstrained, the race-conscious policy can reduce the racial achievement gap most because it can more effectively reduce non-minority students' scores.

is much better for both groups, it reverses the incidence, generating relatively larger gains for higher-achieving students (0.121 versus 0.071). The redistribution of gains from lower-achieving students occurs because putting the best teachers in the biggest classes moves teachers out of lower-achieving schools, which typically have somewhat smaller classes.[41] This reversal illustrates how the distribution of students across classes interacts with the distribution of teacher effects to shape the scope and incidence of gains at the second best.

As we conclude this section, consider three notes on the interpretation of the results in the preceding sections. First, despite the net gains from reassignment, some students will be assigned less effective or worse-matched teachers than in the status quo.[42] Appendix Figure A.8 depicts the distribution of test-score changes under various policies, illustrating both harms and gains. Second, one might worry that the gains from proposed reassignments may not be accurately calibrated. While the purpose of identifying robust optima and reporting honest expected gains was to reduce this concern, assignment policies we consider may still differ too much from the teacher switches observed in practice for the estimated gains to be attainable. Appendix Figure A.9 helps address this concern, revealing that the proposed reassignments are similar to observed teacher switches in terms of changes in class size and achievement composition.[43] Finally, because teachers have distinct value added on different outcomes, the assignments that optimize math scores, reading scores, and behavioral outcomes are each distinct. While it may make sense to focus on math for assignments based on one subject (as in Bates et al., 2025) because the variance in teacher value added is relatively large, our theory suggests that ignoring multidimensionality may be suboptimal. This limitation motivates our need to aggregate gains over multidimensional outcomes and identify an assignment that maximizes welfare, not just math scores.

5. Making Welfare-Improving Assignments

Having illustrated the gains from the second-best policy for math scores, this section considers how teacher assignment might affect welfare more broadly. To that end, it considers multidimensional value added to maximize earnings gains for students, then considers policies that could induce teachers to participate in reassignment. Throughout this section we focus on maximizing students' welfare-weighted present-value lifetime earnings. We choose

[41] As a result, relatively more gains accrue to lower-achieving students at higher-achieving schools.

[42] The fact that some students will be harmed by a given assignment is also a feature of the status quo, so we can benchmark these harms against the fact that students' average "loss" from having a poorly matched teacher is about 0.10 standard deviations (only about 16% of the within-school-grade difference between the best and worst teacher).

[43] This congruence also strengthens our confidence that the quasi-experimental forecast deflation improves the accuracy of reassignment gains because the forecast parameters come from a set of switches with very similar identifying variation.

earnings because they are objective, policy neutral, and related to welfare weights; however, the score function could be extended to include other objectives, such as proficiency standards, high school graduation, criminal justice involvement, or subjective well being.

5.1 Reassignments Raise Lifetime Earnings

To find an assignment that maximizes present-value lifetime earnings, we define a score function that connects teachers' value added on math and reading. We focus on math and reading because of the poor quasi-experimental forecast performance of noncognitive value added. We generate the score function using the causal effects of math and reading value added on earnings in Chetty et al. (2014b):[44]

$$\hat{\mu}^{\mathcal{S}}_{i,j} = \$2234\,\hat{\mu}^{reading}_{i,j} + \$953\,\hat{\mu}^{math}_{i,j}$$

where $\hat{\mu}^{y}_{i,j}$ are the jointly estimated and shrunk effects of teacher j on outcome y for student i, and where gains are measured in nominal 2025 dollars. Following Chetty et al. (2014b), the causal change in predicted present-value lifetime earnings is calculated under three assumptions: (1) individuals may choose to work between the ages of 20 and 65; (2) gains from higher test scores apply to earnings at all ages; and (3) earnings gains are discounted at 3%.[45] These gains are discounted back to age 10, the average age of students in our sample. Later, we also present auxiliary results depicting earnings gains from score functions that also include returns to noncognitive gains.

We define the welfare maximization problem as

$$\max_{\mathcal{J}\in\mathscr{J}} \tilde{\mathcal{W}}(\mathcal{J};\boldsymbol{\omega},\boldsymbol{\mu}) = \max_{\mathcal{J}\in\mathscr{J}} \frac{1}{N} \sum_{t=2003}^{2013} \sum_{k=1}^{K} \sum_{i:\, k_{i,t}=k} \omega_k\, \hat{\mu}^{\mathcal{S}}_{i,t,\mathcal{J}(i,t)} \tag{7}$$

which is identical to Equation 6, but optimizes over the predicted change in score $\hat{\mu}^{\mathcal{S}}_{i,j}$ as a function of our jointly shrunk value-added scores on math and reading by lagged-achievement. Panel (b) of Appendix Figure A.4 reveals that the second-best assignment combines information about value added over lagged terciles of math and reading achievement—with more clearly distinguished losses from more complicated models than we saw in math. Panel (b) of Appendix Figures A.5 and A.6 present sensitivity analyses and Appendix E

[44]The earnings effect of increasing reading scores by one student standard deviation is $\frac{\$189}{0.124\sigma} = \$1,524$ in nominal 2010 dollars. The earnings effect of increasing math scores by one student standard deviation is $\frac{\$106}{0.163\sigma} = \$650$ in nominal 2010 dollars. We then inflate these gains into nominal 2025 dollars (a 46.6% increase). These earnings effects are estimated simultaneously; see Table 6 and the discussion on page 2268 in their paper.

[45]We assume a 5 percent discount rate partially offset by 2 percent real wage growth.

discusses other model selection criteria.

Note that this welfare formulation has two possible empirical limitations. Both arise from the unique empirical and econometric context of the Chetty et al. (2014b) earnings estimates. First, the estimates reflect the effects on students in New York between 1989 and 2009, but our analyses study students in California from 2003–2011. However, given that the data reflect overlapping years in similar large urban school districts, it seems plausible that teachers' pedagogy and students' opportunities are broadly similar across contexts. Second, Chetty et al. (2014b) do not estimate subgroup-specific earnings effects for value added on both math and reading, so our score function relies on population effects. To the extent that earnings effects vary across subgroups, the objective in Equation 7 will produce lower gains than a policy that assigns effective teachers to subgroups with stronger links between test scores and earnings. We assess the sensitivity of our results to subgroup-varying earnings returns below.

Figure 5 depicts the expected gains from reassignment. It shows expected present-value increases in lifetime earnings from experiencing optimal assignments in grades 3–5. Gains to students with below-median lagged math scores are on the y-axis, and gains to students with above-median lagged math scores on the x-axis (see aggregation details in Appendix D.2). The status quo is marked with a square, and PPFs trace out the frontier of second-best optima under different welfare weights, with diamond markers denoting the utilitarian assignments that produce the highest average earnings. We also plot the average expected gains from 50 different fully random (district-wide) assignments as a gray circle.

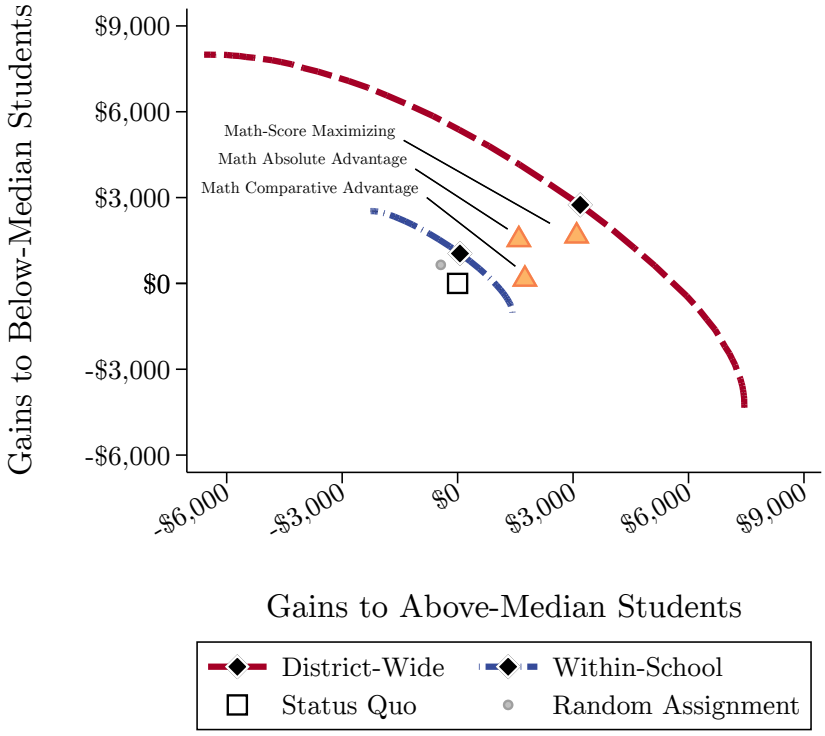
Figure 5 reveals substantial benefits from optimal teacher assignment. We estimate that the second-best optimum under district-wide assignment would generate \$2,700 in present-value earnings for below-median students and \$3,200 for above-median students (\$3,000 on average). These large gains are 2.3 times larger than the gains attainable through a benchmark "deselection" of 5% of teachers.

Policymakers concerned about inequality can also create large redistributive gains. For example, in the district-wide assignment, a social planner could increase the present value of lower-scoring students' earnings by over \$5,400 without hurting high-scoring students on average. A similar comparison reveals gains of about \$1,000 from within-school assignments.[46] Repeated year over year, these assignments could powerfully reduce both achievement and earnings inequality among students coming out of the district.

Furthermore, these gains are also larger than the gains from policies ignoring multidimensionality. Because math value added is more variable across teachers, math-focused assignments seem more desirable due to the greater scope for achievement gains. This in-

[46]This happens to be extremely close to the average-income-maximizing within-school assignment.

Figure 5. Combining Math and Reading Value Added Increases Earnings Gains



Note: This figure depicts the expected present-value lifetime earnings gains from optimally assigning teachers in grades 3–5. The PPFs are plotted by varying the welfare weights for students with above- and below-median math scores in third grade. The outer frontier reports the gains from a district-wide assignment, and the inner frontier from a within-school assignment. The diamond markers represent the (utilitarian) assignments that maximize average earnings. The small gray diamond represents the average gains to random district-wide assignment (50 iterations). Each assignment solves the problem in Equation 7 by robust optimization using jointly shrunk estimates of value added on math and reading by lagged achievement in each relevant subject. These estimates are mapped into earnings following the procedure in Chetty et al. (2014b). The sample includes students in the third-grade cohorts of 2003–2011. The triangle markers represent gains from heuristic assignment policies: "Math-Score Maximizing" is the optimal assignment when considering only math scores; "Math Absolute Advantage" is a policy that makes assignments based on standard math value added alone; and "Math Comparative Advantage" is a policy that assigns the teachers with the highest comparative advantage (based on above- and below-median prior math score value added) to the classes with the highest shares of below-median students.

tuition plays out in the policy experiments considered in Bates et al. (2025) and Delgado (2025), which both focus only on math. By evaluating the assignments proposed in the previous sections with the $\Delta\hat{S}$ estimates of earnings gains, however, our results suggest materially larger gains from considering both math and reading due to the larger earnings impact of reading value added. Specifically, we find that the earnings-maximizing assignment produces

3.2× larger gains (+$2,000) than the comparative advantage heuristic proposed by Delgado (2025), 1.9× larger gains (+$1,400) than the optimal assignment using standard value added mentioned in Bates et al. (2025), and even 1.3× larger gains (+$600) than the district-wide assignment maximizing average math scores from Section 4.

Not only does the second-best assignment increase average gains, but the incidence of considering multidimensionality loads significantly on lower-achieving students. This relationship is illustrated by the vertical distance between the triangle markers and the PPF in Figure 5. The earnings-maximizing assignment raises lower-achieving students' earnings by 142% (about $1,600 more than teacher deselection), 80% (about $1,200) more than an optimal assignment based on standard math value added, and 66% (about $1,000) more than the optimal assignment maximizing math scores. These large differences highlight the distributional implications of including optimal model choice in the social planner problem.

In terms of policy, these gains hint at enormous benefits from optimally reallocating teachers: our estimates suggest that implementing this policy with San Diego teachers in grades 3–5 over our 1998–2019 sample period could have generated present-value gains of roughly $700 million. Assuming comparable gains outside of the San Diego Unified School District, implementing this policy in all U.S. public schools[47] over ten years would generate gains on the order of $106 billion.[48]

Finally we use the information in Figure 5 to consider the implicit welfare considerations underpinning the status quo assignment. School and district administrators make intentional decisions, but the fact that this process is leaving millions of dollars of gains "on the table" does not account for the effect of assignment policies currently in place. Because the model cannot back out welfare weights justifying the status quo, we instead assess the status-quo in comparison to the random-assignment benchmark. The results reveal significant non-random sorting in the status quo that benefits above-median students but reduces gains to below-median students. This pattern is consistent with teachers' well-documented preference to teach in higher-achieving classrooms (e.g., Johnston, 2025) and with SDUSD policies giving hiring priority to more experienced teachers (who tend to have somewhat higher value added). Appendix Figure A.10 confirms this intuition showing that absolute advantage in math and reading value added has small, meaningful correlations with class composition. Although the absolute differences are relatively small, these results raise the question of whether educational administrators face a binding trade off between targeting talent and maintaining assignment incentive compatibility.

[47]There are roughly 3.6 million public school students in each cohort in the United States (NCES, 2024).

[48]Because reassignments will tend to produce smaller gains in smaller districts and larger gains in larger districts, these gains should be considered a rough benchmark rather than a precise calculation.

Robustness and Interpretation. Consider three notes on interpreting these results. First, up to this point our results have implicitly assumed zero-returns to noncognitive value added. Because this restriction may understate the gains from optimal assignment given evidence that noncognitive value added is more predictive of medium-run outcomes (Jackson, 2018; Petek and Pope, 2023), we assess the sensitivity of our results to the restriction. With no causal estimates of the long-term earnings gains from behavioral or attendance value added, we create alternative score functions by combining jointly shrunk value added on above- and below-median students in all four outcomes—math, reading, behavior, and absences—with hypothetical causal effects from \$0 to \$3,000. This is larger than the estimated range of cognitive effects, reflecting evidence of the relative importance of noncognitive skills in earnings gains (Chetty et al., 2011).[49] Appendix Figure A.12 compares these assignments to the earnings-maximizing assignment in Figure 5, showing up to 60% larger gains when using noncognitive outcomes to make assignments. Although this suggests a second-best frontier that is 1.5–4.6x larger than other policy proposals, our enthusiasm is tempered by the poor out-of-sample performance of noncognitive value added in the quasi-experimental forecasts (Appendix Table A.4). If we deflate the gains using the quasi-experimental forecast bias as we do with other results in our paper, the additional gains completely evaporate.

Second, we assess the sensitivity of our analysis to the assumption that earnings gains are uniform across subgroups. Motivated by the finding from Chetty et al. (2014b) that estimated earnings gains from teacher value added are twice as large for higher-income students, we rerun the analysis under the assumption that earnings gains are twice as large for either above- or below-median students, while preserving mean earnings gains overall. We consider each combination of subgroup-specific returns separately for math and reading, and then reassign teachers to optimize the resulting objective. Unsurprisingly, each new allocation favors groups with larger earnings gains. Appendix Figure A.11 displays full results with PPFs from each comparison.

Third, the focus on district-wide reassignments may detract from the significant promise of within-school assignments as a low-cost alternative. In practice, the multi-billion-dollar gains from optimal district-wide assignment policies may be infeasible, but the (relatively costless) within-school reassignments still generate nearly 20% of the gains. The remaining difference suggests a return to relaxing institutional constraints that prevent these multibillion-dollar gains. The following subsection explores one such policy.

[49] Our alternative score functions simulate results slightly larger than the range of cognitive effects that include estimates based on the effects of principal cognitive/noncognitive value added on employment in Texas (Hanushek et al., 2024) and the effects of teacher cognitive/noncognitive value added on postsecondary education outcomes in Greece (Lavy and Megalokonomou, 2024).

5.2 Funding Welfare with a Teacher Bonus Program

In this section we consider a hypothetical bonus pay program for teachers. As noted in Laverde et al. (2026), some assignments may not be incentive compatible without increasing compensation. The hypothetical bonuses allow us to consider welfare and incentive compatibility without explicitly modeling teacher preferences. We imagine a policy providing teachers with additional compensation for participating in reassignment—whether or not their school or class assignment is changed. As long as these bonuses are large enough to ensure incentive compatibility, the welfare under the resulting assignment is reflected in the marginal value of public funds (MVPF; Hendren and Sprung-Keyser, 2020) expended on the bonuses. Studying these bonuses allows us to benchmark the feasibility of the second-best optima and is a step toward implementing welfare-improving policies.

This MVPF is a "bang-for-the-buck" measure of each bonus program, calculated as the present value of total program benefits divided by the net cost of implementing it. Specifically, for a bonus of size b, the MVPF of assignment $\mathcal{J}$ is

$$MVPF^{\mathcal{J}}(b) = \frac{\sum_i (1-\tau)\hat{\mu}^{\mathcal{S}}_{i,\mathcal{J}(i)}}{Jb - \sum_i \tau \hat{\mu}^{\mathcal{S}}_{i,\mathcal{J}(i)}} \tag{8}$$

where $(1-\tau)\hat{\mu}^{\mathcal{S}}_{i,\mathcal{J}(i)}$ are the after-tax present-value monetary gains to each student from assignment $\mathcal{J}$ (given marginal tax rate τ), J is the number of teachers, and $\tau\hat{\mu}^{\mathcal{S}}_{i,\mathcal{J}(i)}$ is the present-value of gains recouped as tax revenue. We focus on earnings gains, $\hat{\mu}^{\mathcal{S}}_{i,\mathcal{J}(i)}$, from the second-best assignment with equal weights on all students and calibrate $\tau = 0.28$ with data from the Opportunity Atlas for San Diego (Chetty et al., 2018).[50]

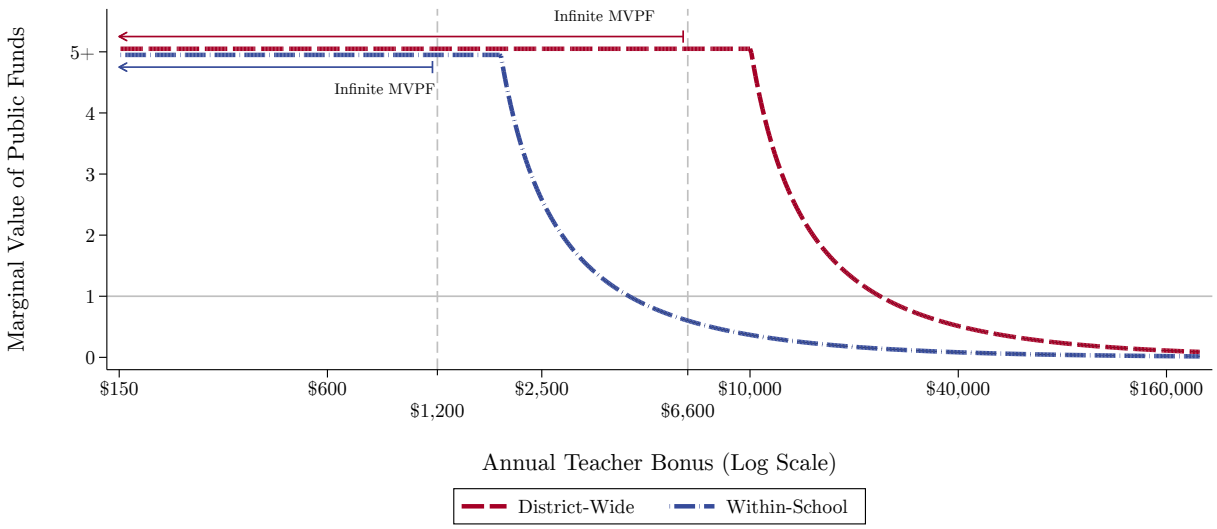
Two key assumptions are needed to make this statistic meaningful in practice. First, the MVPF in Equation 8 reflects the national value of the optimal assignment policy, so there must be a way to internalize the fiscal externality of the local district's policy (see Agrawal et al. (2023) for more on comparing local and national MVPFs). This assumption would be met if state and federal governments funded the teacher bonuses (or transferred the marginal tax revenue back to the district). Otherwise, the MVPF must be interpreted as the total value of implementing this policy nationwide. Second, this MVPF assumes that all teachers must be paid. As such, it is a lower bound on the social gains that could be further improved by targeting bonuses, for example by offering bigger bonuses to teachers who have to make bigger changes or by reassigning only a subset of teachers.

Figure 6 plots the MVPF of increasingly large bonus programs. The two series represent

[50]For children growing up in San Diego County, the median income at age 35 is $43,000. Because the majority of these individuals are unmarried (56%) and still living in the same commuting zone (68%), we apply the marginal tax rates from the United States and California for single filers, 0.22 and 0.06.

the MVPF of a bonus program of a given size for district-wide or within-school reassignments. The MVPF of any bonus program can be interpreted as dollars of social benefit produced for each dollar of net costs spent on bonuses. Values of the MVPF above 5 are reported at the same height on the y-axis. Policies with negative net costs, or $Jb < \sum_i \tau \hat{\mu}^{\mathcal{S}}_{i,\mathcal{J}(i)}$, have an "infinite" MVPF, as indicated in the figure.

Figure 6. Compensating Teachers for Reassignment Could Have Enormous Welfare Impacts



Note: This figure shows the marginal value of public funds (MVPF) of teacher bonus programs of different sizes for either within-school or district-wide assignments. Values are capped at 5 on the figure, the range for which the MVPF is infinite is indicated with arrows, and the x-axis is shown on a log scale.

The main takeaway from Figure 6 is that the MVPF of reassignment bonuses is very large for a broad range of bonus sizes. In fact, many of these bonuses would be *generating* (present-value) revenue while still increasing student earnings. For the district-wide assignment, the MVPF is infinite for bonuses of up to \$6,600 per year. For context, the starting salary for a new teacher is just under \$59,500, so this would imply an 11% annual bonus. For the within-school assignment, the MVPF is infinite for bonuses up to \$1,200 per year. This second value is particularly striking because the intervention is so noninvasive—especially given that teachers care a great deal about commuting distances (Boyd et al., 2005; Bates et al., 2025)—yet it generates gains large enough to justify reasonably large payments to teachers.

Even when the MVPFs are finite, they can remain large even for costly bonus programs that are likely to be incentive compatible. For example, for the district-wide assignment, a bonus program paying *every teacher* in the district \$15,000 per year to participate in the reassignment, regardless of whether they were actually reassigned, would still have an

MVPF of 2.0. In other words, it would generate $2 of present-value earnings gains for every dollar spent on bonuses. With respect to incentive compatibility, in a large randomized controlled trial, a similar payment was more than enough to induce some teachers to change schools (Glazerman et al., 2013).[51] For the within-school exercise, the Measures of Effective Teaching reassigned teachers within school for less than $1,000 (Kane et al., 2013), which would have an infinite MVPF. Although there are no survey data on teacher willingness to accept, conversations with teachers in other similar districts suggest that many would accept $15,000 to switch schools and nearly all would accept $1,000 to switch classes without hesitation. The MVPF of these policies is greater than one for bonuses up to $24,000 and $4,500 per year.

While this specific policy may never be adopted, the exercise highlights the value of similar teacher pay policies. For example, our results suggest that policymakers may consider paying effective teachers to teach slightly larger classes or paying specialized teachers to teach in new schools. Although these assignments could make some teachers worse off as uncompensated switches, the policies we explore tend to generate student earnings gains large enough to justify substantial teacher bonuses. These results suggest that, in a long-run equilibrium, there are many teacher pay programs that could pay for themselves, as long as districts have the capacity to make assignments flexible and receive some of the resulting tax revenue to cover the costs of implementation.

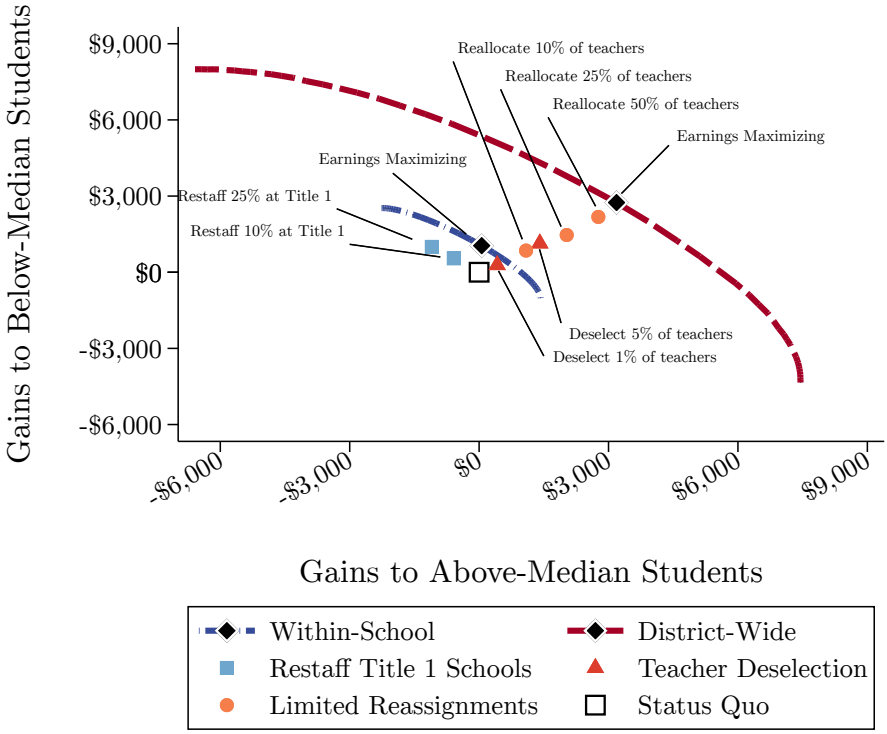
5.3 Welfare Comparisons with Other Policies

Having considered the optimal assignment policies motivated by our theory and quantified the expected gains from each, we conclude by comparing the performance of our preferred policies with three alternatives: limited reassignments, teacher deselection, and restaffing targeted schools. We discuss our approach to analyzing each policy and compare their effectiveness. Figure 7 summarizes the results, plotting the gains from selected counterfactuals compared with the optimal second-best policies.

Limited Reassignments. Because it may be more valuable to reassign certain teachers than others and because there may be diminishing marginal returns to additional reassignments, we consider a set of policies that restrict district-wide assignment to a limited share of teachers. We implement this counterfactual by assuming uniform welfare weights and constraining the linear program to change the assignments of no more than X% of teachers

[51] This experiment offered high value-added teachers $20,000 over two years in 2009 and 2010 or 2010 and 2011, equivalent to $29,300 in 2025 dollars, or roughly $15,000 per year. For context, the average baseline salary of teachers who transferred was $46,604 in 2010. Note that while this did induce changes, take-up was still fairly low.

Figure 7. Second-Best Reassignments Outperform other Proposed Policies



Note: This figure depicts the expected present-value lifetime earnings gains attained by various policies. It reports the gains from limited district-wide assignments (circle markers), restaffing low-resourced schools (square markers), and "deselecting" low-performing teachers (triangle markers) alongside the PPFs from fully optimal district-wide (dashed line) and within-school (dot-dashed line) assignments. The diamond markers represent the (utilitarian) assignments that maximize average earnings. Gains are presented comparing effects on students who had below-average math scores with those who had above-average math scores and are evaluated using the full distribution of jointly estimated value added on math and reading by above- or below-median lagged achievement in each subject.

in 1% increments. We find that reallocating (and paying bonuses to) as few as 25% of teachers achieves nearly 60% of the gains from the optimal district-wide assignment. As such, bonuses of up to $16,000 per year for those teachers would have an infinite MVPF.

Teacher "Deselection." A hallmark policy that uses value added is to remove or "deselect" the least effective teachers (Hanushek, 2009). We implement this counterfactual by identifying the lowest X% of teachers in standard math value added and replacing them with hypothetical teachers who have mean value added on all subgroups and outcomes. As expected, there are sizable gains. Removing the lowest 5% (1%) of teachers would achieve gains equal to about 43% (12%) of those attained by the second-best assignment. This policy is extremely costly to the removed teachers, but it affects far fewer teachers. In fact, because

these teachers tend to be so ineffective, policies to ensure incentive compatibility could have large returns. For example, if suitable replacements could be found, even a severance-pay policy of $59,000 to each removed teacher would have an infinite MVPF in either case.[52]

Restaff Targeted Schools. Another increasingly popular policy approach is to offer large bonuses to high-value-added teachers to teach in low-resourced schools. After pilot results in seven states (Glazerman et al., 2013), many districts have implemented similar policies (e.g., see the setting of Colas and Fu (2025)). We implement this counterfactual by identifying X% of teachers at Title I schools who have standard math value added below the top quartile and then swapping these teachers with randomly selected teachers from the top quartile who teach in the same grade at non-Title I schools. Consistent with the redistributive motive, there are no average gains, but the policy does produce meaningful gains for lower-achieving students. For example, restaffing 10% of positions increases their present-value earnings by $550 ($1,000 for 25%), with comparable losses to higher-achieving students relative to the status quo.

Policy Comparison. Figure 7 depicts three main patterns emerging from this policy comparison. First, the gains from the optimal district-wide assignments dominate all alternatives. Second, the low-cost within-school assignments produce comparable gains (or comparable redistribution) to much more invasive policies such as restaffing 25% of positions at Title I schools or completely removing 1% of teachers. Finally, note the nonuniform incidence of these policies. Unsurprisingly, restaffing Title I schools benefits only lower-achieving students on average, but deselection actually benefits higher-achieving students about 20% more than lower-achieving students. As a result, an equity-minded policymaker would strongly prefer a welfare-weighted second-best policy. Taken together, these results highlight the untapped potential for policies that allow for more flexible and informed teacher assignment and compensation based on value added.

6. Conclusion

A recent poll found that 53% of U.S. adults had a teacher who "changed their life for the better" (Dumitru, 2022). Given the massive potential for good that school teachers have for their students, it is of primary importance to find ways to more effectively utilize teachers' capacity and skills. In this paper, we outlined an approach to use value-added measures more effectively to assign teachers to classes, using tools from public finance to create, compare, and utilize value-added measures that are heterogeneous, multidimensional, and estimated with noise. We then used that approach to describe optimal second-best teacher assignment

[52]Of course, it is not guaranteed that such replacements can be found (Rothstein, 2015). In this case, our counterfactuals overstate the actual gains.

policies for the San Diego Unified School District, while taking multiple steps to reduce misallocation errors due to noise.

The core message of our paper is that there are substantial gains from using value added in the teacher assignment process, particularly when we model the multifaceted nature of teacher effectiveness. We documented large expected gains to both achievement and earnings from assignments based on value added. Back-of-the-envelope calculations suggest large welfare gains from these policies, even if they are quite expensive to implement. We also found that other approaches in the literature (e.g., Graham et al., 2023; Bates et al., 2025; Delgado, 2025; Ahn et al., 2025), while effectively highlighting the importance of value-added match effects, forgo substantial welfare gains. For example, papers that do not allow for gains from absolute advantage made possible by variation in class size (as in Delgado, 2025; Ahn et al., 2025) forgo roughly half of welfare gains. Additionally, these papers focus on one outcome, but considering multidimensional effects generates 20+% larger welfare gains. Likewise, using uniform welfare weights misses out on large gains whenever the social planner has distributional objectives.

We also explored a number of relevant considerations for practitioners studying value added. For example, we adapt the workhorse value-added estimation procedure to accommodate extremely high-dimensional value added and demonstrate the quantitative importance of the "winner's curse" in teacher assignment problems with uncertainty (Andrews et al., 2024). We show that, while innocuous for simple models, measures of this winner's curse, such as volatility and *ex post* regret, explode as models become increasingly complex. Finally, by using robust optimization and focusing on expected rather than predicted gains, we find that relatively simple models of heterogeneity tend to serve the social planner best.

In a broader context, these principles likely apply beyond teacher assignment and evaluation. In public policy, there are many settings where the best assignments likely depend on match effects, such as health (e.g., Dahlstrand, 2022), immigration (Norris, 2019), and criminal justice (Landon, 2024). Applying this approach in such domains could help policymakers improve both the efficiency of service delivery and the equity of outcomes for underserved groups—in short, allowing them to get more value out of value added.

References

Abdulkadiroğlu, Atila, Parag A Pathak, Jonathan Schellenberg, and Christopher R Walters (2020) "Do Parents Value School Effectiveness?," *American Economic Review*, Vol. 110, No. 5, pp. 1502–39.

Abrams, David S and Albert H Yoon (2007) "The Luck of the Draw: Using Random Case Assignment to Investigate Attorney Ability," *University of Chicago Law Review*, Vol. 74, p. 1145.

Agrawal, David, William Hoyt, and Tidiane Ly (2023) "A New Approach to Evaluating the Welfare Effects of Decentralized Policies," Working Paper.

Ahn, Tom, Esteban M Aucejo, and Jonathan James (2025) "The Importance of Student-Teacher Matching: A Multidimensional Value-Added Approach," *Review of Economics and Statistics*, pp. 1–45.

Aizer, Anna and Joseph J Doyle Jr (2015) "Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges," *The Quarterly Journal of Economics*, Vol. 130, No. 2, pp. 759–803.

Alatas, Vivi, Ririn Purnamasari, Matthew Wai-Poi, Abhijit Banerjee, Benjamin A Olken, and Rema Hanna (2016) "Self-targeting: Evidence from a Field Experiment in Indonesia," *Journal of Political Economy*, Vol. 124, No. 2, pp. 371–427.

Andrews, Isaiah, Toru Kitagawa, and Adam McCloskey (2024) "Inference on Winners," *The Quarterly Journal of Economics*, Vol. 139, No. 1, pp. 305–358.

Angrist, Joshua, Peter Hull, and Christopher Walters (2023) "Methods for Measuring School Effectiveness," *Handbook of the Economics of Education*, Vol. 7, pp. 1–60.

Arnold, David, Will Dobbie, and Peter Hull (2022) "Measuring Racial Discrimination in Bail Decisions," *American Economic Review*, Vol. 112, No. 9, pp. 2992–3038.

Athey, Susan, Raj Chetty, Guido W Imbens, and Hyunseung Kang (2025) "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely," *Review of Economic Studies*, p. rdaf087.

Athey, Susan and Stefan Wager (2021) "Policy Learning with Observational Data," *Econometrica*, Vol. 89, No. 1, pp. 133–161.

Aucejo, Esteban, Patrick Coate, Jane Cooley Fruehwirth, Sean Kelly, and Zachary Mozenter (2022) "Teacher Effectiveness and Classroom Composition: Understanding Match Effects in the Classroom," *The Economic Journal*, Vol. 132, No. 648, pp. 3047–3064.

Baron, E Jason, Richard Lombardo, Joseph P Ryan, Jeongsoo Suh, and Quitze Valenzuela-Stookey (2024) "Mechanism Reform: An Application to Child Welfare," NBER Working Paper.

Bates, Michael, Michael Dinerstein, Andrew C Johnston, and Isaac Sorkin (2025) "Teacher Labor Market Policy and the Theory of the Second Best," *The Quarterly Journal of Economics*, Vol. 140, No. 2, pp. 1417–1469.

Bau, Natalie (2022) "Estimating an Equilibrium model of Horizontal Competition in Education," *Journal of Political Economy*, Vol. 130, No. 7, pp. 1717–1764.

Bertsimas, Dimitris and John N Tsitsiklis (1997) *Introduction to Linear Optimization*, Vol. 6: Athena scientific Belmont, MA.

Betts, Julian R (2011) "The Economics of Tracking in Education," in *Handbook of the Economics of Education*, Vol. 3: Elsevier, pp. 341–381.

Beuermann, Diether W, C Kirabo Jackson, Laia Navarro-Sola, and Francisco Pardo (2023) "What is a Good School, and Can Parents Tell? Evidence on the Multidimensionality of School Output," *The Review of Economic Studies*, Vol. 90, No. 1, pp. 65–101.

Bhatt, Monica P, Sara B Heller, Max Kapustin, Marianne Bertrand, and Christopher Blattman (2024) "Predicting and Preventing Gun Violence: An Experimental Evaluation of READI Chicago," *The Quarterly Journal of Economics*, Vol. 139, No. 1, pp. 1–56.

Bhuller, Manudeep, Gordon B Dahl, Katrine V Løken, and Magne Mogstad (2020) "Incarceration, Recidivism, and Employment," *Journal of Political Economy*, Vol. 128, No. 4, pp. 1269–1324.

Biasi, Barbara, Chao Fu, and John Stromme (2021) "Equilibrium in the Market for Public School Teachers: District Wage Strategies and Teacher Comparative Advantage," NBER Working Paper.

Bobba, Matteo, Tim Ederer, Gianmarco Leon-Ciliotta, Christopher Neilson, and Marco G Nieddu (2026) "Teacher Compensation and Structural Inequality: Evidence from Centralized Teacher School Choice in Perú," NBER Working Paper.

Boyd, D., H. Lankford, S. Loeb, and J. Wyckoff (2005) "The Draw of Home: How Teachers' Preferences for Proximity Disadvantage Urban Schools," *Journal of Policy Analysis And Management*, Vol. 24, No. 1, pp. 113–132.

Branch, Gregory F, Eric A Hanushek, and Steven G Rivkin (2009) *Estimating Principal Effectiveness*: Urban Institute Washington, DC.

Chan, David C, Matthew Gentzkow, and Chuan Yu (2022) "Selection with Variation in Diagnostic Skill: Evidence from Radiologists," *The Quarterly Journal of Economics*, Vol. 137, No. 2, pp. 729–783.

Chandra, Amitabh, Amy Finkelstein, Adam Sacarny, and Chad Syverson (2016) "Health Care Exceptionalism? Performance and Allocation in the US Health Care Sector," *American Economic Review*, Vol. 106, No. 8, pp. 2110–2144.

Chernozhukov, Victor, Jerry A Hausman, and Whitney K Newey (2019) "Demand Analysis with Many Prices," NBER Working Paper.

Chernozhukov, Victor, Sokbae Lee, Adam M Rosen, and Liyang Sun (2025) "Policy Learning with Confidence," *arXiv preprint arXiv:2502.10653*.

Chetty, Raj, John N Friedman, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter (2018) "The Opportunity Atlas," *Opportunity Insights*.

Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan (2011) "How does your Kindergarten Classroom Affect your Earnings? Evidence from Project STAR," *The Quarterly Journal of Economics*, Vol. 126, No. 4, pp. 1593–1660.

Chetty, Raj, John N Friedman, and Jonah E Rockoff (2014a) "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, Vol. 104, No. 9, pp. 2593–2632.

——— (2014b) "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American*

Economic Review, Vol. 104, No. 9, pp. 2633–79.
Colas, Mark and Chao Fu (2025) "Information Friction and Teachers' Labor Market," NBER Working Paper.
Condie, Scott, Lars Lefgren, and David Sims (2014) "Teacher Heterogeneity, Value-Added and Education Policy," *Economics of Education Review*, Vol. 40, pp. 76–92.
Dahlstrand, Amanda (2022) "Defying Distance? The Provision of Services in the Digital Age," Working Paper.
Dee, Thomas S (2005) "A Teacher like Me: Does Race, Ethnicity, or Gender Matter?," *American Economic Review*, Vol. 95, No. 2, pp. 158–165.
Delgado, William (2025) "Disparate Teacher Effects, Comparative Advantage, and Match Quality," *Economics of Education Review*, Vol. 106, p. 102648.
Delhommer, Scott (2022) "High School Role Models and Minority College Achievement," *Economics of Education Review*, Vol. 87, p. 102222.
Dobbie, Will, Jacob Goldin, and Crystal S Yang (2018) "The Effects of Pre-Trial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, Vol. 108, No. 2, pp. 201–240.
Doyle, Joseph, John Graves, and Jonathan Gruber (2019) "Evaluating Measures of Hospital Quality: Evidence from Ambulance Referral Patterns," *Review of Economics and Statistics*, Vol. 101, No. 5, pp. 841–852.
Duflo, Esther, Pascaline Dupas, and Michael Kremer (2011) "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, Vol. 101, No. 5, pp. 1739–1774.
Dumitru, Oana (2022) "How are Teachers Changing their Students' Lives?," YouGov, August, URL: `https://today.yougov.com/society/articles/43501-how-teachers-changing-their-students-lives-poll`, Accessed *via* YouGov Today website.
Einav, Liran, Amy Finkelstein, and Neale Mahoney (2025a) "Producing Health: Measuring Value Added of Nursing Homes," *Econometrica*, Vol. 93, No. 4, pp. 1225–1264.
Einav, Liran, Amy Finkelstein, Neale Mahoney, and James C Okun (2025b) "Racial Differences in Nursing Home Value Added," NBER Working Paper, National Bureau of Economic Research.
Finkelstein, Amy and Matthew J Notowidigdo (2019) "Take-up and Targeting: Experimental Evidence from SNAP," *The Quarterly Journal of Economics*, Vol. 134, No. 3, pp. 1505–1556.
Gabrel, Virginie, Cécile Murat, and Aurélie Thiele (2014) "Recent Advances in Robust Optimization: An Overview," *European Journal of Operational Research*, Vol. 235, No. 3, pp. 471–483.
Gilraine, Michael and Nolan G Pope (2021) "Making Teaching Last: Long-Run Value-Added," NBER Working Paper.
Gilraine, Mike, Jiaying Gu, and Robert McMillan (2022) *A Nonparametric Approach for Studying Teacher Impacts*.
Glazerman, Steven, Ali Protik, Bing-ru Teh, Julie Bruch, and Jeffrey Max (2013) "Transfer Incentives for High-Performing Teachers: Final Results from a Multi-site Randomized Experiment,"Technical report, U.S. Department of Education.
Graham, Bryan S, Geert Ridder, Petra Thiemann, and Gema Zamarro (2023) "Teacher-to-Classroom Assignment and Student Achievement," *Journal of Business & Economic Statistics*, Vol. 41, No. 4, pp. 1328–1340.
Hanushek, Eric A (2009) "Teacher Deselection," *Creating a New Teaching Profession*, Vol. 168, pp. 172–173.
——— (2011) "The Economic Value of Higher Teacher Quality," *Economics of Education Review*, Vol. 30, No. 3, pp. 466–479.
Hanushek, Eric A, Andrew J Morgan, Steven G Rivkin, Jeffrey C Schiman, Ayman Shakeel, and Lauren Sartain (2024) "The Lasting Impacts of Middle School Principals," NBER Working Paper.
Harrington, Emma and Hannah Shaffer (2023) "Estimating Prosecutor Effects on Incarceration and Reoffense," Working Paper.
Harris, Lauren, Cecilia Moreira, and Aastha Rajan (2026) "Incentivizing Success: Assessing the Teacher Incentive Allotment Program's Impact on Teacher Labor Markets and Student Achievement," Working Paper.
Hendren, Nathaniel and Ben Sprung-Keyser (2020) "A Unified Welfare Analysis of Government Policies," *The Quarterly Journal of Economics*, Vol. 135, No. 3, pp. 1209–1318.
Hoerl, Arthur E and Robert W Kennard (1970) "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, Vol. 12, No. 1, pp. 55–67.
Hull, Peter (2020) "Estimating Hospital Quality with Quasi-Experimental Data," Working Paper.
Hussam, Reshmaan, Natalia Rigol, and Benjamin N Roth (2022) "Targeting High Ability Entrepreneurs Using Community Information: Mechanism Design in the Field," *American Economic Review*, Vol. 112, No. 3, pp. 861–98.
Ida, Takanori, Takunori Ishihara, Koichiro Ito, Daido Kido, Toru Kitagawa, Shosei Sakaguchi, and Shusaku Sasaki (2022) "Choosing Who Chooses: Selection-Driven Targeting in Energy Rebate Programs," NBER Working Paper.
Imberman, Scott A and Michael F Lovenheim (2016) "Does the Market Value Value-Added? Evidence from Housing Prices after a Public Release of School and Teacher Value-Added," *Journal of Urban Economics*, Vol. 91, pp. 104–121.
Ito, Koichiro, Takanori Ida, and Makoto Tanaka (2023) "Selection on Welfare Gains: Experimental Evidence from Electricity Plan Choice," *American Economic Review*, Vol. 113, No. 11, pp. 2937–2973.
Jackson, C Kirabo (2018) "What do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes," *Journal of Political Economy*, Vol. 126, No. 5, pp. 2072–2107.
Jackson, C Kirabo, Jonah E Rockoff, and Douglas O Staiger (2014) "Teacher Effects and Teacher-Related Policies," *Annu. Rev. Econ.*, Vol. 6, No. 1, pp. 801–825.
Jacob, Brian A and Lars Lefgren (2007) "What do Parents Value in Education? An Empirical Investigation of Parents' Revealed Preferences for Teachers," *The Quarterly Journal of Economics*, Vol. 122, No. 4, pp. 1603–1637.
Johnston, Andrew C (2025) "Preferences, Selection, and the Structure of Teacher Pay," *American Economic Journal: Applied Economics*, Vol. 17, No. 3, pp. 310–346.
Kane, Thomas J, Daniel F McCaffrey, Trey Miller, and Douglas O Staiger (2013) "Have we Identified Effective Teachers? Validating Measures of Effective Teaching using Random Assignment," in *Research Paper. MET Project. Bill & Melinda Gates Foundation*, Citeseer.
Kitagawa, Toru and Aleksey Tetenov (2018) "Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, Vol. 86, No. 2, pp. 591–616.
Kling, Jeffrey R (2006) "Incarceration Length, Employment, and Earnings," *American Economic Review*, Vol. 96, No. 3, pp. 863–876.

Landon, Taylor J. (2024) "Publicly Contracted Private Counsel: Defense Attorney Quality, Pay, and Defendant Outcomes," Working Paper.

Laverde, Mariana, Elton Mykerezi, Aaron Sojourner, and Aradhya Sood (2026) "Gains from Alternative Assignment? Evidence from a Two-Sided Teacher Market," Working Paper.

Lavy, Victor and Rigissa Megalokonomou (2024) "Alternative Measures of Teachers' Value Added and Impact on Short and Long-Term Outcomes: Evidence from Random Assignment," NBER Working Paper.

Macartney, Hugh, Robert McMillan, and Uros Petronijevic (2021) "A Quantitative Framework for Analyzing the Distributional Effects of Incentive Schemes," NBER Working Paper.

Martin, Ian WR and Stefan Nagel (2022) "Market Efficiency in the Age of Big Data," *Journal of Financial Economics*, Vol. 145, No. 1, pp. 154–177.

Mbakop, Eric and Max Tabord-Meehan (2021) "Model Selection for Treatment Choice: Penalized Welfare Maximization," *Econometrica*, Vol. 89, No. 2, pp. 825–848.

Mulhern, Christine (2023) "Beyond Teachers: Estimating Individual School Counselors' Effects on Educational Attainment," *American Economic Review*, Vol. 113, No. 11, pp. 2846–2893.

Mulhern, Christine and Isaac Opper (2021) "Measuring and summarizing the multiple dimensions of teacher effectiveness."

NCES (2024) "Enrollment in Public Elementary and Secondary Schools, by Level, Grade, and Race/Ethnicity: Selected Years, fall 2013 through Fall 2023,"Technical report, National Institute of Education Statistics.

Neal, Derek and Diane Whitmore Schanzenbach (2010) "Left Behind by Design: Proficiency Counts and Test-Based Accountability," *The Review of Economics and Statistics*, Vol. 92, No. 2, pp. 263–283.

Norris, Samuel (2019) "Examiner Inconsistency: Evidence from Refugee Appeals," *University of Chicago, Becker Friedman Institute for Economics Working Paper*, No. 2018-75.

Osborne-Lampkin, La'Tara and Lora Cohen-Vogel (2014) ""Spreading the Wealth": How Principals use Performance Data to Populate Classrooms," *Leadership and Policy in Schools*, Vol. 13, No. 2, pp. 188–208.

Petek, Nathan and Nolan G Pope (2023) "The Multidimensional Impact of Teachers on Students," *Journal of Political Economy*, Vol. 131, No. 4, pp. 1057–1107.

Rothstein, Jesse (2015) "Teacher Quality Policy when Supply Matters," *American Economic Review*, Vol. 105, No. 1, pp. 100–130.

San Diego Unified School District (2019) "Student–Teacher Linked Administrative Records, 1998–1999 through 2018–2019," Restricted-use administrative data.

Small, Marian (2012) *Good Questions: Great Ways to Differentiate Mathematics Instruction*: Teachers College Press.

Tikhonov, Andrei Nikolaevich, Alexander V Goncharsky, Vaceslav Vasilevic Stepanov, and Anatoly Grigorevich Yagola (1995) "Numerical Methods for the Approximate Solution of Ill-Posed Problems on Compact Sets," in *Numerical Methods for the Solution of Ill-posed Problems*: Springer, pp. 65–79.

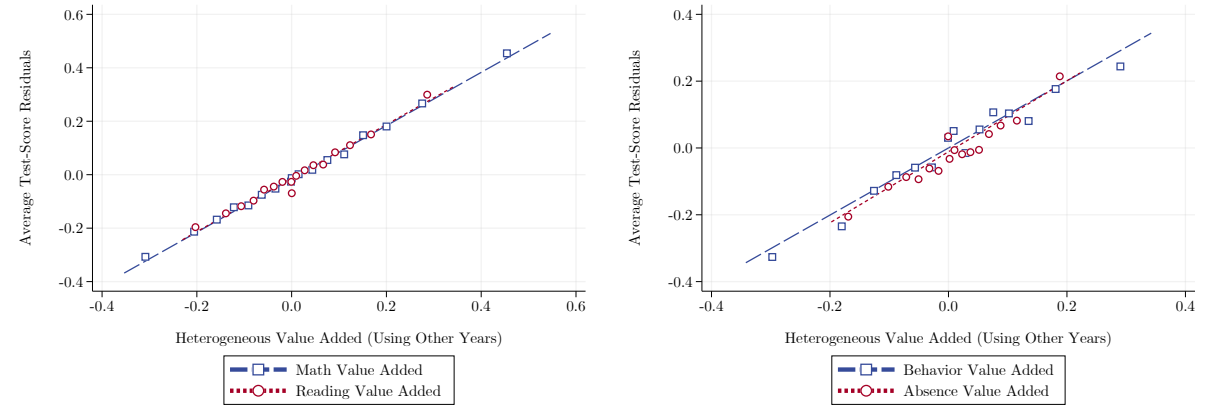
Tomlinson, Carol Ann (2017) *How to Differentiate Instruction in Academically Diverse Classrooms*: ASCD, 3rd edition.

Walters, Christopher (2024) "Empirical Bayes methods in labor economics," in *Handbook of Labor Economics*, Vol. 5: Elsevier, pp. 183–260.

Supplemental Appendix

A. Additional Tables and Figures

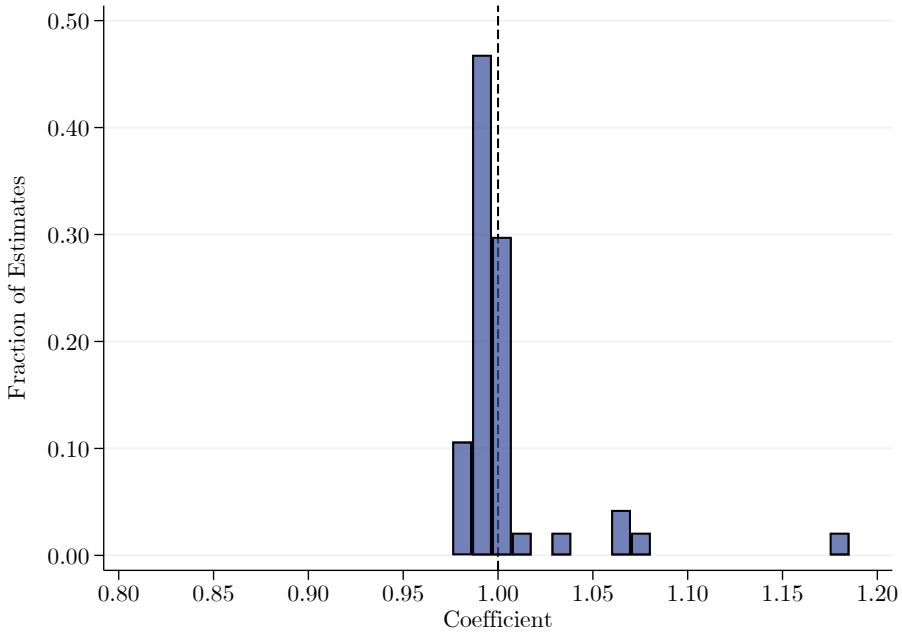
Figure A.1. Heterogeneous Value-added Measures Are Forecast Unbiased



(a) Math and Reading **(b)** Behavior and Absences

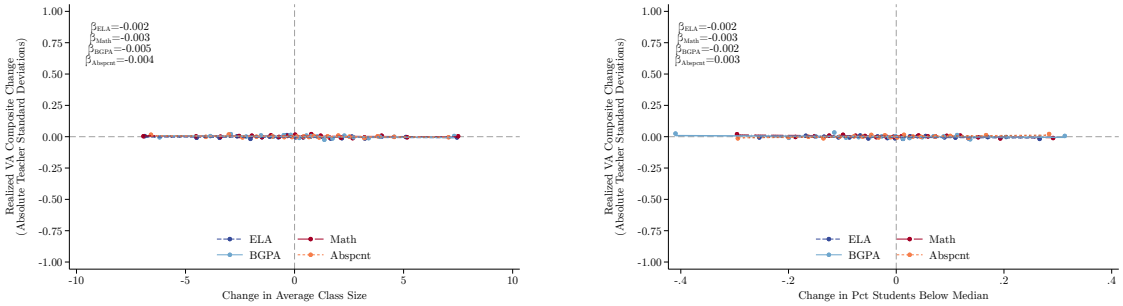
Note: This figure shows the relationship between average residuals for each teacher's class and their predicted value added (average estimated match effect on students in the class) based on prior years of data.

Figure A.2. All Models Exhibit Minimal Forecast Bias



Note: This figure shows the forecast-unbiasedness coefficients for different models. This figure shows the relationship between average residuals for each teacher's class and their predicted value added (average estimated match effect on students in the class) based on prior years of data. The four bars to the right of 1.05 are value added estimates on single noncognitive outcomes, i.e., standard or median split behavioral GPA value added and standard or median split attendance value added.

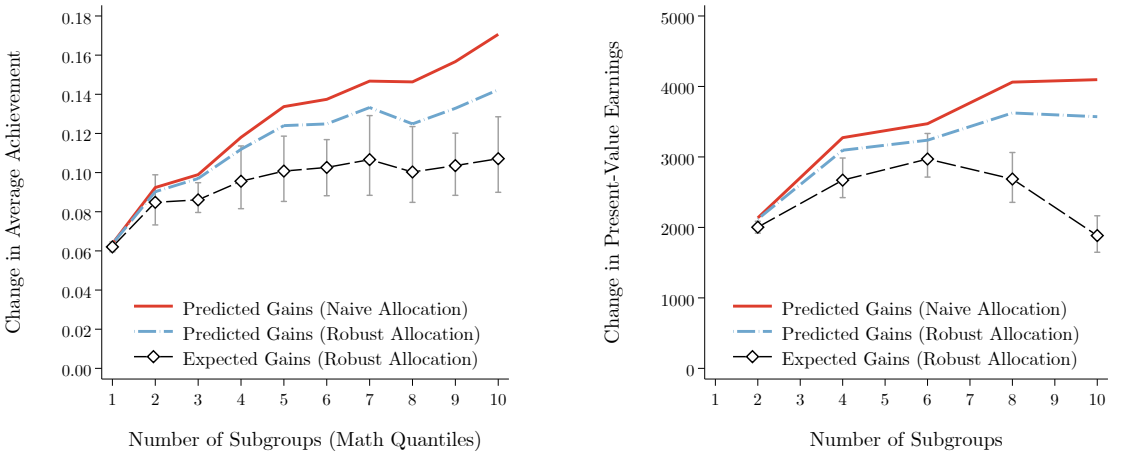
Figure A.3. Quasi-Experimental Changes in Teachers' Value Added are Unrelated to Class Size or Composition



(a) ΔVA by ΔClass Size **(b)** ΔVA by ΔPct Below Median

Note: This figure shows how our heterogeneous quasi-experimental estimates of teacher value added at the school-grade-year cell relate to experienced changes in class size and class composition. Panel (a) plots value-added changes over the number of students in class (β reports the within-teacher change associated with a five-student change in class size). Panel (b) plots value-added changes over the share of lower-achieving students (β reports the within-teacher change associated with a 10 percent change).

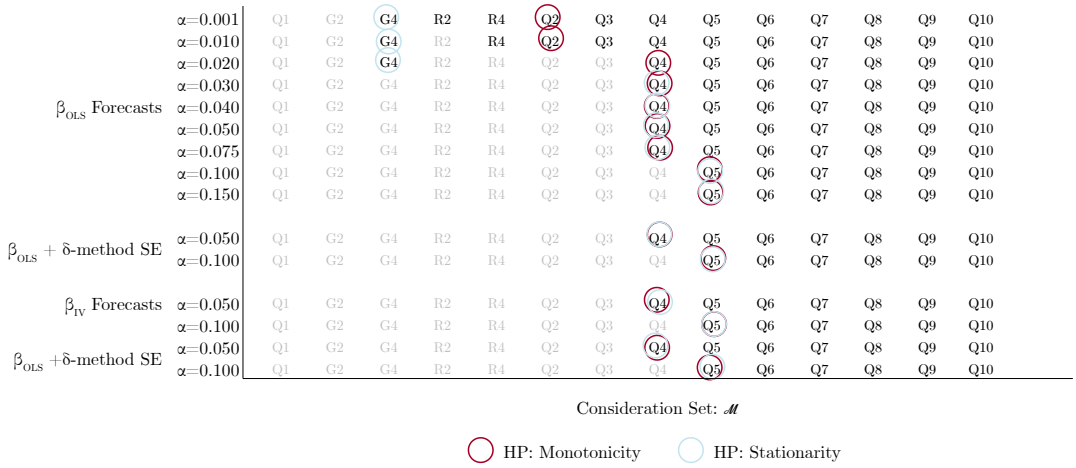
Figure A.4. Predicted Gains Are Far too Optimistic



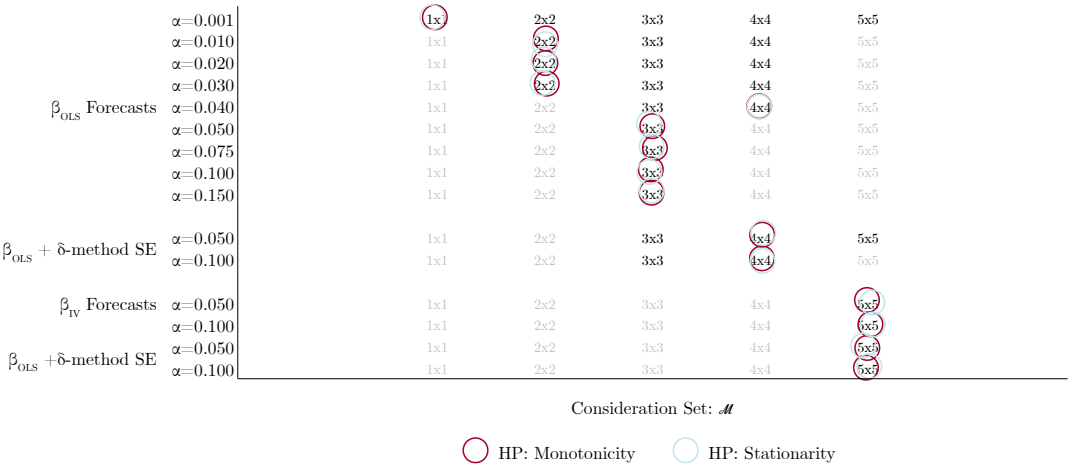
(a) Math Scores **(b)** Lifetime Earnings

Note: This figure shows the predicted gains from naive assignments and the predicted and expected gains from robust assignments. Predicted gains will be much larger than expected gains if over-optimism is a problem. The predicted gains tend to be (far) above even the 97.5th percentile of the empirical distribution of gains from the bootstrapped estimates. While not depicted in the figure, expected gains from the naive assignment tend to be slightly lower than expected gains from the robust assignment in most models, though they are not statistically distinguishable.

Figure A.5. Model-Selection Specification Curves



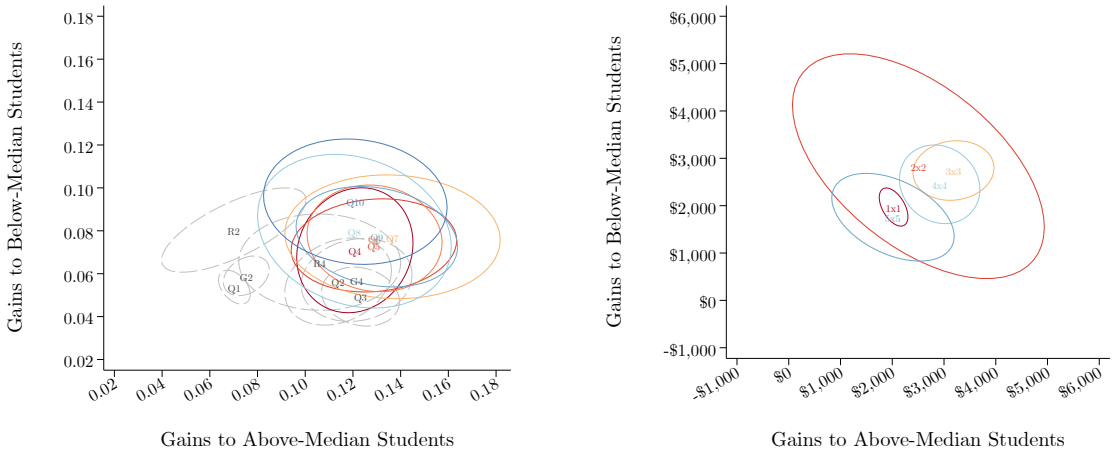
(a) Math Scores



(b) Lifetime Earnings

Note: This figure shows the results of the model selection exercise under a variety of different assumptions, approaches, and confidence levels. Each row reflects the model choice given (1) a forecast-bias parameter, (2) a standard error correction method, and (3) a confidence level for the one-tailed test. The columns indicate different models. For each row, we identify the model with the largest forecast-deflated expected gains, and identify a consideration set including all models with statistically indistinguishable gains given the inference method and confidence level. Models listed in black are included in the consideration set for each row while those in light gray are not. The model in each consideration set with the best hyperparameters under each criterion are circled in (dark) red or (light) blue. Delta method tests are done with a parametric bootstrap using the standard error of the forecast parameter.

Figure A.6. Models Close to m^* Produce Similar Incidence and Gains

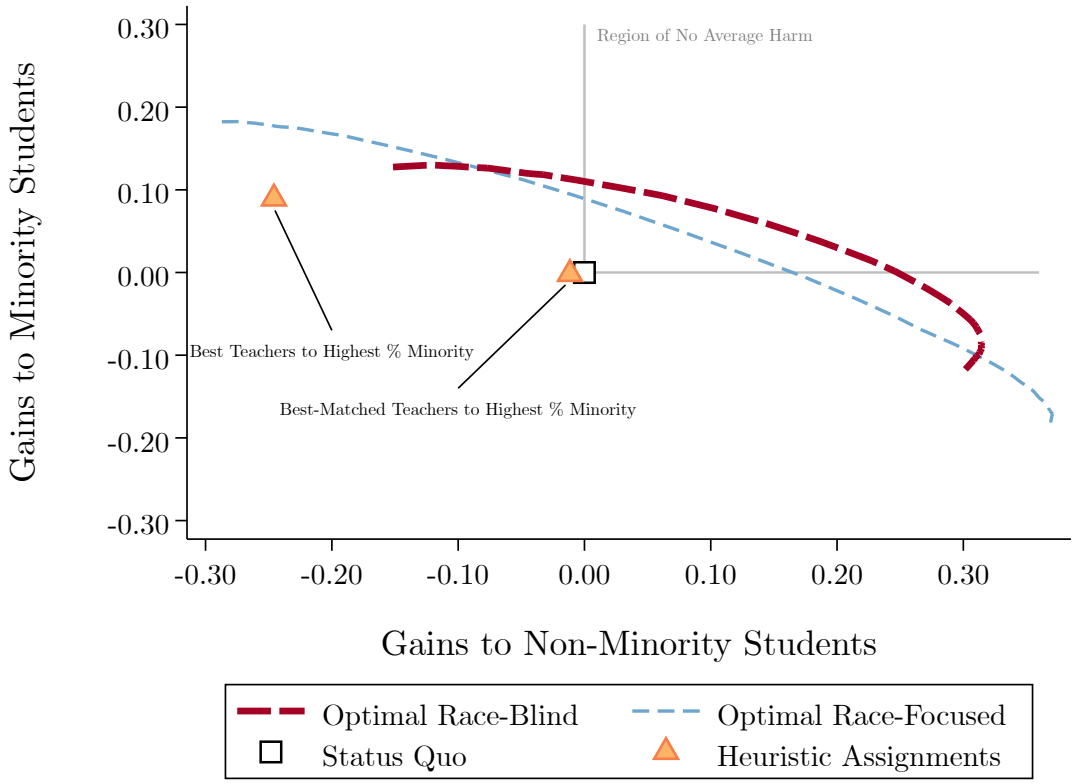


(a) Math Scores

(b) Lifetime Earnings

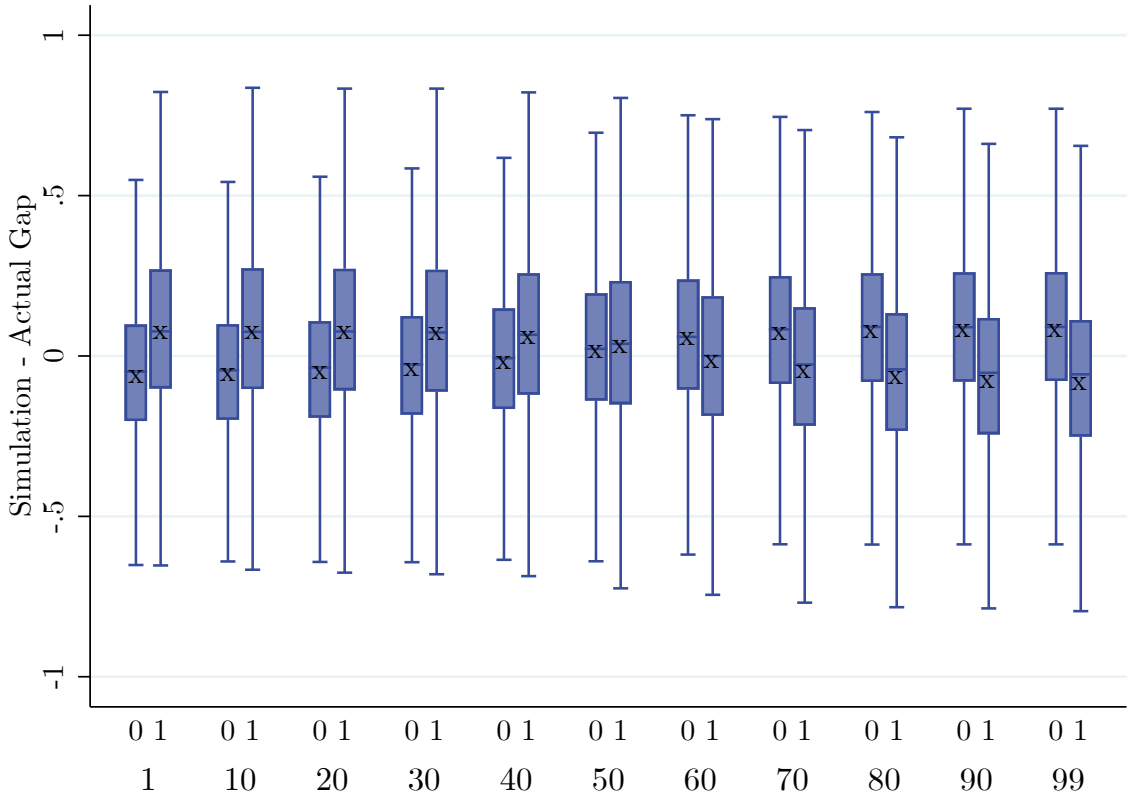
Note: This figure shows distribution of forecast-deflated expected gains across models. Gains to students with below-average prior math scores are depicted on the y axis with gains to students with above-average prior math scores on the x axis. Expected gains are indicated with a short label. Approximate elliptical 95% confidence intervals are calculated as based on the (co)variances across 1,000 bootstrapped evaluations: $\left(\bar{y}_{m,H} + \chi^{-1}(2,.95)\hat{\sigma}_H \cos(\theta), \bar{y}_{m,L} + \chi^{-1}(2,.95)\left[\frac{\hat{\sigma}_{L,H}}{\hat{\sigma}_H}\cos(\theta) + \sqrt{\hat{\sigma}_L^2 - (\frac{\hat{\sigma}_{L,H}}{\hat{\sigma}_H})^2}\sin(\theta)\right]\right)$. To avoid color repetition in panel (a), models outside of the consideration set are depicted in gray with dashed borders. Note that in both panels the chosen model produces economically indistinguishable gains from all comparable models.

Figure A.7. Race-Blind and Race-Conscious Reassignments



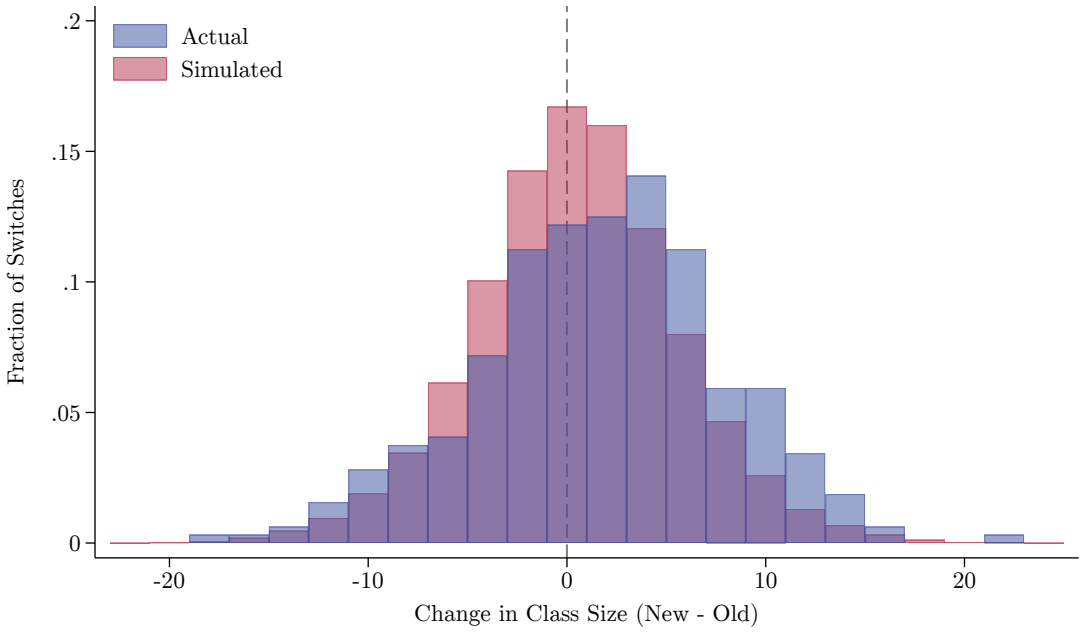
Note: This figure compares the gains from reassignments by student race using two racial groups: (1) Black and Hispanic students and (2) other students. We report two sets of results from optimal reassignments: one using value added by achievement and one using value added by race. We also report two heuristic policies: one placing the teachers with the largest comparative advantage at teaching minority students in the classes with the largest share of minority students, and the other placing teachers with the largest absolute advantage (average value added across race) in the classes with the largest share of minority students.

Figure A.8. Reassignments come with Gains to Many Students and Losses to Some Relative to the Status Quo

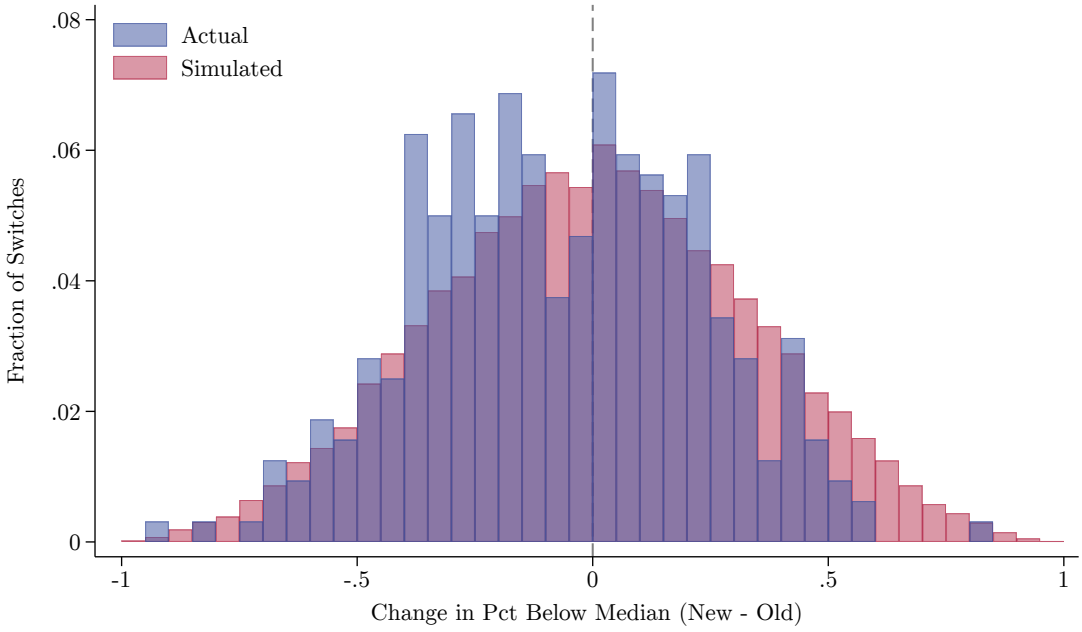


Note: This figure shows the distribution of differences in gains between the reassignment simulation and the status quo. The first row on the x-axis shows lower-achieving students (0) and higher-achieving students (1). The second row gives the percent of the weight placed on lower-achieving students. The "x" marks give the mean difference for the given group, and the boxes show the 5th, 25th, 50th, 75th, and 95th percentiles.

Figure A.9. Distribution of Observed and Proposed Switches



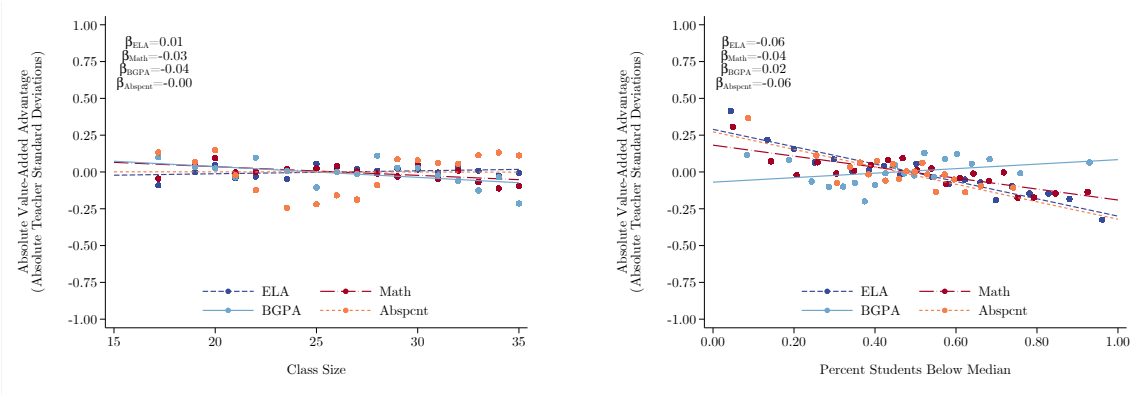
(a) Class Size



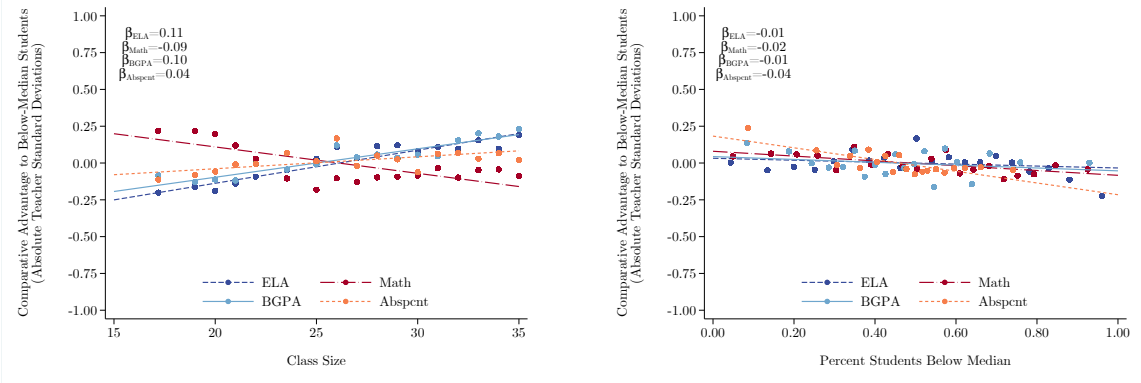
(b) Percent Below Median

Note: These figures show the distribution of changes in class size (panel a) and percent below median in the class (panel b) for teachers who actually switched schools in the data (new class - old class in each case) and for our proposed teacher reassignments.

Figure A.10. Status Quo Assigns Lower-Achieving Students Worse Teachers



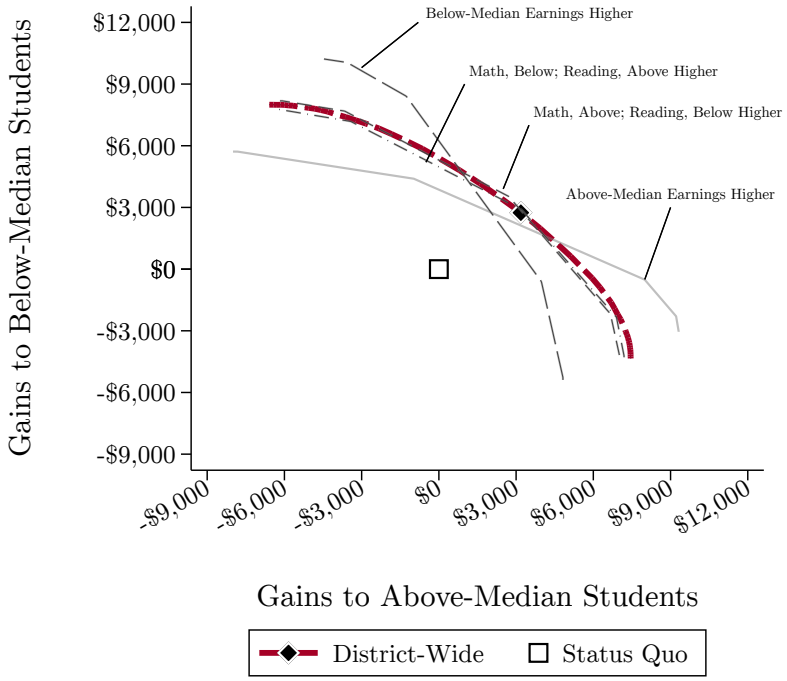
(a) Absolute Adv. by Class Size **(b)** Absolute Adv. by Pct Below Median



(c) Comparative Adv. by Class Size **(d)** Comparative Adv. by Pct Below Median

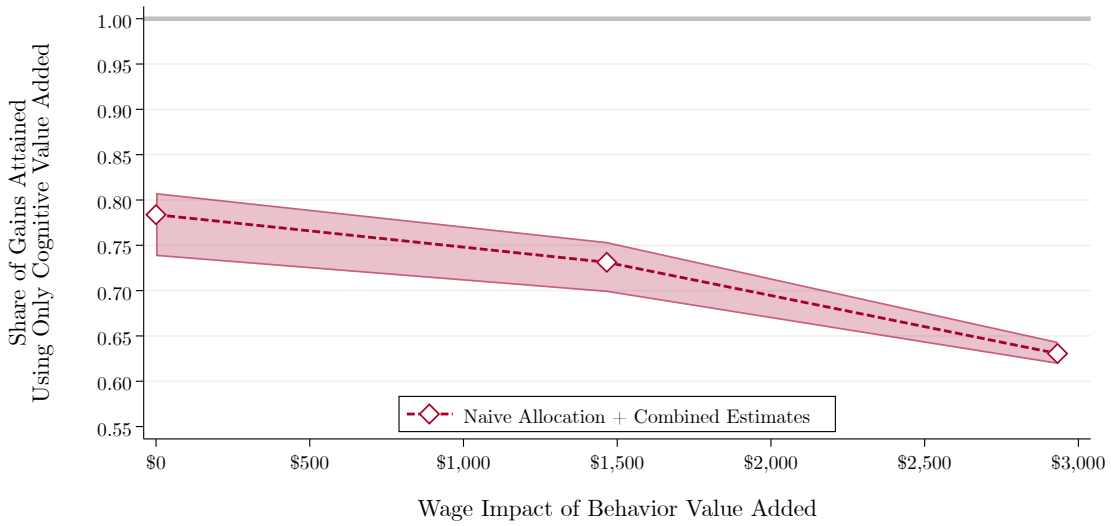
Note: This figure shows how our heterogeneous estimates of teacher value-added on reading, math, behavior, and absences relate to status quo class size and class composition. The top panels show teacher absolute advantage (average value added among higher- and lower-scoring students) and the bottom panels show the comparative advantage (difference in value added on higher- and lower-median scoring students). The panels on the left plot the cross-sectional variation of absolute advantage over the number of students in class where β reports the cross-sectional change associated with a five-student change in class size. The panels on the right plot the cross-sectional variation over the share of lower-achieving students where β reports the cross-sectional change associated with a 10 percent change in the fraction of below-median students in the class.

Figure A.11. Robustness to Assumptions About Earnings



Note: This figure depicts the expected present-value lifetime earnings gains from optimally assigning teachers in grades 3–5. The PPFs are plotted by varying the welfare weights for students with above- and below-median math scores in third grade. The red frontier reports the gains from a district-wide assignment and is the same as Figure 5. The white square is the status quo. Each assignment solves the problem in Equation 7 by robust optimization using jointly shrunk estimates of value added on math and reading by lagged achievement in each relevant subject. The gray PPFs show different assumptions about how earnings map to above- and below-median students. Following estimates from Chetty et al. (2014b), we preserve the mean earnings used and vary whether above- and below-median students have higher earnings separately for math and reading. In each case 'Higher' denotes twice the earnings to the indicated group relative to the other. All other splits (equal earnings on math or reading, a mix of equal on one subject and unequal on the other) fall in the middle of the PPFs displayed in the figure. The sample includes students in the third-grade cohorts of 2003–2011. As expected, when gains favor below-median students in both subjects the PPF is steeper; when gains favor above-median students, it becomes flatter. When gains favor different groups across different subjects, the gray PPFs track the original red PPF from Figure 5 closely. Ultimately, gains are similar in magnitude to the baseline results, though the slope of the PPF depends on the extent to which gains accrue disproportionately to one group or the other.

Figure A.12. Percent of Gains from Assignments Ignoring Noncognitive Skills



Note: This figure reports the overall share of gains attained by a naive assignment ignoring noncognitive outcomes. The plot reports the percentage of gains attained by assignments made using only math and reading relative to assignments using all four outcomes—evaluated using jointly shrunk value added across all four outcomes. The x-axis reports the gains for different values of the return to a student standard deviation increase in behavior value added, and the range plot covers the same range for attendance value added (with the connected scatter representing $1,500). Because all gains are evaluated using estimates jointly shrunk with all four outcomes, the (naive) assignment using only (jointly shrunk) math and reading value added is no longer optimal. Indeed, because the estimates are different, this naive evaluation produces slightly lower expected gains (about 19% less) than the optimal assignment using estimates jointly shrunk with all four outcomes, even when there is no causal effect of noncognitive scores on earnings.

Table A.1. Full Correlations Between Above- and Below-Median Value Added

	Math Below	Math Above	Reading Below	Reading Above	Behavior Below	Behavior Above	Absences Below	Absences Above
Math Below	-	0.94	0.67	0.67	-0.00	0.00	0.01	0.02
Math Above	0.94	-	0.66	0.67	0.00	0.01	-0.05	-0.07
Reading Below	0.67	0.66	-	0.94	0.01	0.00	0.02	0.00
Reading Above	0.67	0.67	0.94	-	0.01	0.01	0.03	0.02
Behavior Below	-0.00	0.00	0.01	0.01	-	0.93	0.12	0.11
Behavior Above	0.00	0.01	0.00	0.01	0.93	-	0.13	0.11
Absences Below	0.01	-0.05	0.02	0.03	0.12	0.13	-	0.85
Absences Above	0.02	-0.07	0.00	0.02	0.11	0.11	0.85	-
Standard Dev	0.18	0.20	0.13	0.12	0.16	0.14	0.09	0.10
$\mathbb{E}[\lvert CA\rvert]/\sigma_{AA}$	0.30		0.26		0.31		0.44	
CA Range / AA Range	0.36		0.36		0.42		0.57	

Note: The standard deviations are reported in the row labeled "Standard Dev," and each element in the matrix is the correlation between the row and the column. "Below" and "Above" refer to students below and above the median. The row labeled "$\mathbb{E}[|CA|]/\sigma_{AA}$" shows the mean absolute comparative advantage divided by the standard deviation of absolute advantage for the subject. The row labeled "CA Range / AA Range" shows the difference between a 99th-percentile teacher and a 1st-percentile teacher in terms of comparative advantage divided by the difference between a 99th-percentile teacher and a 1st-percentile teacher in terms of absolute advantage.

Table A.2. Correlations Between Quartile Estimates of Math and Gender Value Added

	Math 1st	Math 2nd	Math 3rd	Math 4th	Male Below	Male Above	Female Below	Female Above
Math 1st	-	0.95	0.88	0.83	0.95	0.82	0.95	0.88
Math 2nd	0.95	-	0.90	0.85	0.93	0.81	0.96	0.89
Math 3rd	0.88	0.90	-	0.99	0.96	0.96	0.86	0.98
Math 4th	0.83	0.85	0.99	-	0.93	0.97	0.82	0.98
Male Below	0.95	0.93	0.96	0.93	-	0.94	0.92	0.97
Male Above	0.82	0.81	0.96	0.97	0.94	-	0.76	0.98
Female Below	0.95	0.96	0.86	0.82	0.92	0.76	-	0.87
Female Above	0.88	0.89	0.98	0.98	0.97	0.98	0.87	-
Standard Dev	0.17	0.22	0.20	0.20	0.17	0.20	0.20	0.20

Note: The standard deviations are reported in the last row, and each element in the matrix is the correlation between the row and the column. "1st", "2nd", "3rd", and "4th" refer to students of the respective quartiles. "Below" and "Above" refer to students below and above the median in terms of prior math scores.

Table A.3. Estimates Show Meaningful Within-Teacher Comparative Advantage

	Math	Reading	Behavior	Absences
Frac 95% Same Sign	0.25	0.16	0.16	0.08
Frac Significantly Diff Above and Below	0.26	0.22	0.31	0.06
Frac Years w/ Same CA	0.79	0.65	0.76	0.63

Note: The first row shows the fraction of teacher-year estimates that have the same signed comparative advantage in 95% of the 1,000 bootstrapped estimates. The second row shows the fraction of teacher-year estimates for which we reject equality between the above- and below-median estimates using a standard difference-in-means t-test with standard errors calculated from the bootstrap distribution. The third row shows the fraction of years in which teachers in our data, whom we observe for at least three years, have a comparative advantage teaching the same subgroup of students that we estimated for them in their first year. In the third row, we also focus on teachers who should be consistently good at teaching one group by dropping teachers who have a comparative advantage within one standard deviation of zero in their first year.

Table A.4. Quasi-Experimental Forecast Bias

	Panel A: Math achievement				Panel B: ELA achievement			
	OLS		IV (switchers)		OLS		IV (switchers)	
Model	$\hat{b}$	(SE)	$\hat{b}$	(SE)	$\hat{b}$	(SE)	$\hat{b}$	(SE)
All students (STD)	1.020	(0.083)	0.980	(0.130)	0.992	(0.102)	1.106	(0.161)
Lagged-Achievement:								
Q2	1.049	(0.084)	0.988	(0.137)	1.001	(0.106)	1.073	(0.151)
Q3	0.989	(0.086)	0.917	(0.140)	0.963	(0.109)	1.094	(0.150)
Q4	0.977	(0.084)	0.903	(0.138)	1.024	(0.107)	1.124	(0.157)
Q5	0.964	(0.080)	0.898	(0.138)	0.933	(0.103)	1.105	(0.154)
Q6	0.961	(0.082)	0.878	(0.139)	0.875	(0.101)	1.044	(0.154)
Q7	0.946	(0.079)	0.863	(0.129)	0.807*	(0.104)	1.029	(0.156)
Q8	0.928	(0.078)	0.846	(0.126)	0.807**	(0.097)	1.038	(0.145)
Q9	0.939	(0.076)	0.900	(0.128)	0.761**	(0.112)	0.965	(0.195)
Q10	0.910	(0.077)	0.822	(0.128)	0.735**	(0.104)	1.003	(0.160)
Demographics:								
Gender	1.018	(0.082)	0.982	(0.128)	0.957	(0.100)	1.105	(0.160)
Race (White/Asian)	1.012	(0.084)	0.969	(0.130)	0.961	(0.105)	1.089	(0.157)
Gender × Q2	1.017	(0.083)	0.964	(0.135)	0.962	(0.105)	1.046	(0.156)
Race × Q2	0.964	(0.085)	0.928	(0.141)	0.917	(0.104)	1.029	(0.153)

	Panel C: Joint ELA+Math models			
	OLS		IV (switchers)	
Model	$\hat{b}$	(SE)	$\hat{b}$	(SE)
Math + Reading:				
$Q_1^M \times Q_1^R$	0.986	(0.090)	1.094	(0.139)
$Q_2^M \times Q_2^R$	0.960	(0.103)	1.015	(0.147)
$Q_3^M \times Q_3^R$	1.025	(0.094)	1.129	(0.146)
$Q_4^M \times Q_4^R$	0.918	(0.162)	1.128	(0.162)
$Q_5^M \times Q_5^R$	0.601	(0.309)	1.301	(0.186)
Demographics + Math + Reading:				
$G \times Q_1^M \times Q_1^R$	0.989	(0.089)	1.145	(0.141)
$G \times Q_2^M \times Q_2^R$	1.218**	(0.101)	1.268*	(0.153)
$R \times Q_1^M \times Q_1^R$	0.333***	(0.192)	1.088	(0.132)
$R \times Q_2^M \times Q_2^R$	0.508*	(0.291)	0.452*	(0.311)

	Panel D: Noncognitive outcomes and earnings			
	OLS		IV (switchers)	
Model	$\hat{b}$	(SE)	$\hat{b}$	(SE)
Behavioral GPA:				
All students (STD)	−0.885***	(0.346)	−1.273***	(0.592)
Q2	−0.633***	(0.331)	−1.260***	(0.591)
Absences:				
All students (STD)	0.146***	(0.256)	0.185**	(0.324)
Q2	0.304**	(0.275)	−0.017***	(0.294)
BGPA + Absences:				
Beh. GPA	−0.504***	(0.331)	−1.390***	(0.611)
Absences	0.527**	(0.233)	0.623	(0.264)
BGPA + Absences + Math + Reading:				
ELA	1.014	(0.139)	1.240	(0.196)
Math	1.097	(0.093)	1.108	(0.153)
BGPA	−0.123***	(0.339)	−0.429**	(0.687)
Absences	0.225***	(0.184)	0.243**	(0.331)

Note: This table presents the quasi-experimental estimates of forecast coefficients in our setting. We follow the methodology of Chetty et al. (2014a) with the OLS estimates, regressing the change in average student achievement in the school-grade-year cell on the change in average teacher value added in the school-grade-year cell and time fixed-effects. We implement the IV methodology of Gilraine and Pope (2021) with the IV estimates. Stars indicate that we reject that the estimates are forecast unbiased $\beta = 1$ at the indicated level, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.5. Naive and Robust Reassignment Results

	Math Standard	Gender	Race	Math Median Split	Math Terciles	Math Quartiles	Math Quintiles	Gender x Median Split	Race x Median Split
Across District:									
Naive (Not Imputed)	0.061	0.065	0.070	0.080	0.083	0.089	0.091	0.084	0.073
	[0.059,0.063]	[0.061,0.070]	[0.058,0.087]	[0.070,0.093]	[0.078,0.089]	[0.078,0.101]	[0.079,0.103]	[0.073,0.095]	[0.062,0.085]
Naive (All Teachers)	0.061	0.065	0.076	0.081	0.088	0.097	0.103	0.088	0.089
	[0.059,0.063]	[0.061,0.070]	[0.063,0.095]	[0.071,0.094]	[0.082,0.094]	[0.085,0.111]	[0.090,0.118]	[0.077,0.100]	[0.076,0.104]
Robust (All Teachers)	0.061	0.066	0.075	0.081	0.087	0.098	0.105	0.088	0.089
	[0.059,0.063]	[0.062,0.071]	[0.063,0.093]	[0.071,0.092]	[0.082,0.094]	[0.085,0.114]	[0.091,0.120]	[0.077,0.098]	[0.075,0.104]
Within School:									
Naive (Not Imputed)	0.012	0.013	0.012	0.015	0.014	0.015	0.014	0.014	0.010
	[0.012,0.013]	[0.012,0.014]	[0.010,0.015]	[0.012,0.018]	[0.013,0.016]	[0.012,0.018]	[0.011,0.018]	[0.012,0.017]	[0.009,0.012]
Naive (All Teachers)	0.012	0.013	0.014	0.015	0.016	0.018	0.019	0.016	0.015
	[0.012,0.013]	[0.012,0.014]	[0.012,0.016]	[0.013,0.019]	[0.014,0.018]	[0.015,0.022]	[0.015,0.024]	[0.013,0.019]	[0.013,0.017]
Robust (All Teachers)	0.012	0.013	0.014	0.015	0.016	0.018	0.020	0.016	0.015
	[0.012,0.013]	[0.012,0.014]	[0.012,0.016]	[0.013,0.018]	[0.014,0.018]	[0.015,0.023]	[0.016,0.024]	[0.014,0.019]	[0.013,0.018]

Note: This table shows the expected gains from three different assignments: a naive assignment ignoring parameter uncertainty and only reallocating teachers who have taught each subgroup; a naive assignment using imputed value added for teachers who have never taught certain subgroups; and a robust assignment using imputed value-added scores (our main specification). Expected gains and confidence intervals are computed using the 1,000 bootstrapped samples.

Table A.6. Reading and Dollar Reassignment Results

Panel A. Reading					
	Reading Standard	Reading Median Split	Reading Terciles	Reading Quartiles	Reading Quintiles
Robust (All Teachers)	0.034	0.048	0.056	0.062	0.066
	[0.032,0.035]	[0.044,0.052]	[0.049,0.066]	[0.054,0.072]	[0.056,0.076]
	Gender	Gender x Median Split	Race	Race x Median Split	
Robust (All Teachers)	0.038	0.054	0.044	0.061	
	[0.034,0.045]	[0.045,0.066]	[0.038,0.054]	[0.053,0.072]	
Panel B. Dollars					
	Standard	Median Split	Terciles	Quartiles	Quintiles
Robust (All Teachers)	2032	2782	2898	2926	3135
	[1946,2117]	[2523,3109]	[2648,3251]	[2566,3336]	[2741,3602]

Note: This table shows the gains analogous to Figure 3 (i.e., robust, outcome-maximizing reassignments across schools in the district) for reading and dollar outcomes, instead of focusing on math as the outcome. The first panel shows reading results split by prior achievement, the second shows reading results by demographics and interactions with prior achievement, and the last shows the results considering both reading and math and their impact on long-term earnings, with the indicated splits by prior achievement in both subjects using the gain estimates from Chetty et al. (2014b) of $1,524 for a one standard deviation increase in reading and $650 for a one standard deviation increase in math then inflating those gains to nominal 2025 dollars and reporting lifetime earnings. All gains are expected gains computed using the 1,000 bootstrapped samples. 95% confidence intervals are also reported from re-estimating value added, recomputing the optima, and evaluating gains in 1,000 bootstrapped samples.

Table A.7. *p*-Value on Pairwise Model Comparisons

Model Used	Share worse than ...												
in Assignment	STD	G2	R2	G4	R4	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
Q2 Math	0.000	0.003	0.169	0.836	0.595								
Q3 Math	0.000	0.000	0.143	0.671	0.467	0.407							
Q4 Math	0.000	0.000	0.044	0.179	0.129	0.056	0.094						
Q5 Math	0.000	0.000	0.021	0.060	0.029	0.021	0.027	0.195					
Q6 Math	0.000	0.000	0.020	0.047	0.025	0.017	0.012	0.150	0.398				
Q7 Math	0.000	0.000	0.016	0.042	0.025	0.014	0.012	0.121	0.238	0.341			
Q8 Math	0.000	0.000	0.032	0.128	0.085	0.067	0.040	0.253	0.515	0.661	0.756		
Q9 Math	0.000	0.000	0.017	0.049	0.032	0.025	0.025	0.168	0.350	0.434	0.601	0.303	
Q10 Math	0.000	0.000	0.007	0.029	0.017	0.014	0.015	0.093	0.197	0.275	0.481	0.185	0.283

Note: This table shows the *p*-values of the test of the one-sided null hypothesis that the model in each row produces *smaller* gains than the model in each column. *p*-values are created by comparing the forecast-deflated gains from the optimal assignment using each model across 1,000 bootstrapped estimation samples. For example [$p = 0.003$] indicates that the model in the column outperformed the model in the row for 3/1000 sets of bootstrapped estimates. STD indicates the standard (homogeneous value added model), G2 and R2 represent models with heterogeneity by gender or race, and G4 and R4 have heterogeneity by gender or race interacted with above- or below-median lagged outcomes. Models with richer quantiles of lagged achievement are indicated as Q#. Because of inverse symmetry only lower diagonal elements are reported.

B. Theory Appendix

B.1 Derivation of Social Planner Problem (Equation 1)

Definition 1 (Set Up). Let welfare be a weighted sum of lifetime utilities, $W^{\mathcal{J}} = \sum_{i=1}^{n} \phi_i U_i^{\mathcal{J}}$, where the utilities $U_i^{\mathcal{J}}$ may depend on the policy $\mathcal{J}$, but the *ex ante* welfare weights ϕ_i do not. Let $S_i^{\mathcal{J}} = s(\boldsymbol{Y}_i^{\mathcal{J}}, \boldsymbol{X}_i)$ be a score function summarizing outcomes, $\boldsymbol{Y}_i$, and characteristics, $\boldsymbol{X}_i$.

Assumption 1 (Policy Independence). Assume that for all i, the assignment policy $\mathcal{J}$ only affects lifetime utilities and score functions through $\mathcal{J}(i) = j$, the teacher assigned to teach student i.

Assumption 2 (Score Function). Assume $s(\cdot)$ (1) is additively separable into a student- and a match-specific component $S_i^{\mathcal{J}} = \bar{S}_i + \mu_{i,j}^{\mathcal{S}}$ for $j = \mathcal{J}(i)$ and (2) has assignment-invariant welfare content that is locally linear around $\boldsymbol{S}^0$ such that $\mathbb{E}[U_i^{\mathcal{J}} | S_i^{\mathcal{J}} = s] = \alpha_i + \beta_i s$

Proposition 1. Under Assumptions 1 and 2, the expected welfare of any assignment, conditional on the scores $(\boldsymbol{S}^{\mathcal{J}}, \boldsymbol{S}^0)$ and set of welfare weights $\boldsymbol{\phi}$ is $\mathcal{W}^{\mathcal{J}}$ up to a level normalization:

$$\mathcal{W}^{\mathcal{J}} \equiv \sum_j \sum_{i:\mathcal{J}(i)=j} \omega_i \mu_{i,j}^{\mathcal{S}}$$

where $\omega_i = \phi_i \beta_i$ are *ex ante* welfare weights adjusted for measurement in scores rather than lifetime utilities. Proof in Appendix B.5.

Remark (Surrogate Index). Consider the case that $S_i^{\mathcal{J}}$ is a surrogate index for lifetime utility, $S = \mathbb{E}[U|Y, X]$. In this case $\omega_i = \phi_i$. This is because under surrogacy and independence $\mathbb{E}[U^{\mathcal{J}}|S^{\mathcal{J}}] = S^{\mathcal{J}}$ (Athey et al., 2025). In other words by mapping the intermediate outcomes into the units of the long-term outcome of interest, surrogate index removes the necessity of adjusting welfare weights from changes in utilities into scores.

B.2 Derivation of Plug-In Welfare Error (Equation 2)

Definition 2 (Plug-In Welfare Estimate). Let $W^{\mathcal{J}} = \sum_j \sum_{i:\mathcal{J}(i)=j} \omega_i \mu_{i,j}^{\mathcal{S}}$. Then, given any set of subgroup value-added estimates $\{\mu_{j,k}\}_m$ where $k \in \mathcal{K}_m$ is a partition of students into subgroups used by model m. Suppressing the superscripts, $^{\mathcal{S}}$, let the model-m approximation of welfare be:

$$\widehat{\mathcal{W}}_m^{\mathcal{J}} = \sum_j \sum_{k \in \mathcal{K}_m} n_{j,k}(\mathcal{J}) \bar{\omega}_{j,k}(\mathcal{J}) \hat{\mu}_{j,k}$$

where $n_{j,k}(\mathcal{J})$ denotes the number of type k students assigned to teacher j under policy $\mathcal{J}$, $\bar{\omega}_{j,k}(\mathcal{J})$ is those students' average welfare weight, and $\hat{\mu}_{j,k}$ are estimates of teacher j's value added on type k students.

Proposition 2. Under Assumptions 1 and 2, the welfare estimation error of $\widehat{\mathcal{W}}^{\mathcal{J}}$ is

$$\begin{aligned}\mathcal{W}^{\mathcal{J}} - \widehat{\mathcal{W}}^{\mathcal{J}}_m &\equiv \sum_j \sum_{i:\,\mathcal{J}(i)=j} \omega_i \mu_{i,j} - \sum_j \sum_{k\in\mathcal{K}_m} n_{j,k}(\mathcal{J})\,\bar{\omega}_{j,k}(\mathcal{J})\,\hat{\mu}_{j,k} \\ &= \sum_j \sum_{k\in\mathcal{K}_m} n_{j,k}(\mathcal{J})\bar{\omega}_{j,k}(\mathcal{J}) \Bigg[\underbrace{\tilde{\mu}_{j,k}(\mathcal{J}) - \bar{\mu}_{j,k}}_{\text{Matching Gains}} + \underbrace{\frac{\text{Cov}(\omega_i, \mu_{i,j}; \mathcal{J})}{\bar{\omega}_{j,k}(\mathcal{J})}}_{\text{Distributional Gains}} + \underbrace{\bar{\mu}_{j,k} - \tilde{\mu}_{j,k}(\mathcal{J}_0)}_{\text{External Validity}} \\ &\qquad + \underbrace{\tilde{\mu}_{j,k}(\mathcal{J}_0) - \mu^{\star}_{j,k}}_{\text{Shrinkage Bias}} + \underbrace{\mu^{\star}_{j,k} - \hat{\mu}_{m,j,k}}_{\text{Estimation Error}} \Bigg]\end{aligned}$$

where $\bar{\mu}$ denotes the population average match effects, $\tilde{\mu}$ denote assigned-class match effects, and $\mu^{\star}$ and $\hat{\mu}$ represent oracle and feasible value-added estimates. Proof in Appendix B.5.

Remark (Expected Shrinkage Bias). Despite the fact that shrinkage bias is non-zero conditional on realized welfare, it may sum to zero in many cases. If noise is random and approximately normal, then $\sum_j (\mu^{\star}_{j,k} - \mu_{j,k}(\mathcal{J}_0)) \approx 0$ as the positive and negative shrinkage biases average out. This implies that we expect no effect of shrinkage bias in predicted welfare whenever $\big(n_{j,k}(\mathcal{J})\,\bar{\omega}_{j,k}(\mathcal{J})\big) \perp (\mu^{\star}_{j,k} - \mu_{j,k})$. While this is extremely likely for a randomly drawn $\mathcal{J}$, or across all $\mathcal{J}$ in expectation, Proposition 4 will show that this need not be the case once the social planner uses $\hat{\mu}$ to select an optimizing $\hat{\mathcal{J}}$.

B.3 Derivation of Mean Squared Error of Plug-In Welfare

Definition 3 (Model Misspecification). Let $M^{\mathcal{J}}_m$ represent the sum of three model misspecification terms from Proposition 2 such that $\mathcal{W}^{\mathcal{J}} = \boldsymbol{x}^{\mathcal{J}\prime}\boldsymbol{\mu}_m - M^{\mathcal{J}}_m$ where $\boldsymbol{x}^{\mathcal{J}}$ is a vector of $x^{\mathcal{J}}_{m,jk} = n_{j,k}(\mathcal{J})\,\bar{\omega}_{j,k}(\mathcal{J})$ terms that capture the cell-specific welfare distance of assignment $\mathcal{J}$ from the status quo and $\boldsymbol{\mu}_m$ denotes the stacked vector of target value-added parameters over $\mathcal{K}_m$.

Definition 4 (Feasible Shrinkage). Let $\hat{\boldsymbol{\mu}}_m$ the corresponding feasible estimates. Let $\hat{\boldsymbol{q}}_m = [\hat{\boldsymbol{A}}_m; \hat{\boldsymbol{\sigma}}_m]$ be the stacked subgroup residuals and hyperparameter estimates used to estimate $\hat{\boldsymbol{\mu}}_m$, with population analogues $\boldsymbol{A}_m$ and $\boldsymbol{\sigma}_m$. Let $D(\hat{\boldsymbol{q}}_m) = \hat{\boldsymbol{\mu}}_m$ be a function that maps all $j \times k$ residuals and $\mathcal{H}_m$ hyperparameters into the $j \times k$ value added estimates.

Assumption 3 (Hyperparameter Estimation). Assume that (1) hyperparameter estimation is unbiased, $\mathbb{E}[\hat{\boldsymbol{q}}_m] = \boldsymbol{q}_m$, (2) that $D(\cdot)$ is continuously differentiable and (3) that $D(\cdot)$ is locally linear around the true $\boldsymbol{q}$.

Proposition 3. Under Assumptions 1–3, the MSE of welfare from evaluating a fixed assignment $\mathcal{J}$ with $\hat{\boldsymbol{\mu}}_m$ is

$$\begin{aligned}\text{MSE}(\widehat{\mathcal{W}}_m^{\mathcal{J}}) &\equiv \mathbb{E}\left[\left(\widehat{\mathcal{W}}_m^{\mathcal{J}} - \mathcal{W}^{\mathcal{J}}\right)^2\right] \\ &= \left[\underbrace{M_m^{\mathcal{J}}}_{\text{Misspecification Bias}} + \underbrace{\boldsymbol{x}^{\mathcal{J}\prime}\left(D(\boldsymbol{q}_m) - \boldsymbol{\mu}_m\right)}_{\text{Shrinkage Bias}}\right]^2 + \underbrace{\boldsymbol{x}^{\mathcal{J}\prime} D_q \Omega_q D_q' \boldsymbol{x}^{\mathcal{J}}}_{\text{Welfare Variability}}\end{aligned}$$

where $\Omega_q = \mathbb{V}[\hat{\boldsymbol{q}}_m - \boldsymbol{q}_m]$ and $\mathbb{E}[\cdot]$ is taken over sampling uncertainty. Furthermore, if residuals are approximately independent across teacher-type cells, and $\hat{\boldsymbol{q}}_m - \boldsymbol{q}_m$ is approximately normal, then an optimistic characterization the elements of variance Ω_q is

$$\begin{aligned}s_k^2 \equiv \mathbb{V}(\hat{A}_{j,k} - \mu_{j,k}) &\geq \frac{\sigma_{\varepsilon k}^2 K}{n}\left[1 + \frac{\ell_{j,k}^{\hat{\beta}}}{J}\right] \\ \mathbb{V}(\hat{\sigma}_{k,k'} - \sigma_{k,k'}) &\geq \frac{1 + \mathbb{1}[k = k']}{J - 1}\left(\sigma_{k,k} s_{k'}^2 + \sigma_{k',k'} s_k^2 + s_{k'}^2 s_k^2\right)\end{aligned}$$

where n is the number of students per teacher, $\ell_{j,k}^{\hat{\beta}} = \bar{X}_{j,k}' \mathbb{E}[x_i x_i']^{-1} \bar{X}_{j,k}$ is a leverage term increasing in the distance between the student characteristics in (j,k) and in the population that captures the influence of first-stage estimation error, $\mathbb{V}(\hat{\beta} - \beta)$, on the variance. Proof in Appendix B.5.

Remark (Bayes Risk and Borrowing Strength). Note that by defining a true welfare $\mathcal{W}^{\mathcal{J}}$ Proposition 3 takes teacher effects as given. When Empirical Bayes estimates are integrated over the prior distribution (e.g., redrawing teachers from the super-population), then under the oracle hyperparameters, the sum of squared shrinkage bias and parameter variability are weakly decreasing as dimensions are added. This is not guaranteed for feasible Empirical Bayes or for shrunk estimates conditional on realized effects. In this light Proposition 3 can illustrate the power of borrowing strength. Consider two models with the same number of dimensions, but one is restricted so that all subgroups are treated independently. Mechanically this forces $D_q = 0$ for many elements. Without the smoothing that comes from averaging over correlated signals, the restricted model will be more variable than the unrestricted model.

Remark (Additional Estimation Error). Proposition 3 provides a lower bound on estimation errors under some relatively strict conditions. Common issues such as imbalances in classroom composition, unmodeled differences across schools or years, finite-sample issues with fixed-effect estimation, unmodeled heteroskedasticity or interdependence in ε, and

other design issues may increase variance relative to the balanced benchmark. Indeed, in Proposition 4 we show that the joint support of teachers who teach multiple subgroups is a crucial determinant of overfit welfare estimation error.

Remark (Hyperparameters and Welfare Variability). Finally, note that changes in welfare variability tend to be governed by the variability of hyperparameters. In general estimates of subgroup-class residuals become more noisy as the number of subgroups increase while the number of students in each affected cell decreases at exactly the same rate. If the rank of D_q stays proportional, welfare variability will not tend to change because of variability in these residuals. Hyperparameter variability, however, may continue to increase and will pass additional variance into the estimates and their weighted sum on predicted welfare.

B.4 Derivation of Expected Optimism

Definition 5 (Principal Components). Let Γ_K be the projection of the true/limiting-model-m covariance of teacher effects onto $\mathcal{K}_m$, then the R principal components of Γ_K are characterized by eigenvalues γ_r and $(K_m \times 1)$ eigenvectors $\boldsymbol{u}_r$.

Lemma 1 (Shrinkage Projection). Assume that for all k, $\sigma_{\epsilon,k}\left(1+\frac{\hat{\ell_k^\beta}}{J}\right)=\sigma$. Given this the oracle shrinkage matrix $B_K = \Gamma_K(\Gamma_K + s^2 I_K)^{-1}$ is diagonal in the principal-component basis, i.e., $U'B_KU$ is a diagonal matrix with elements $b_r = \frac{\gamma_r}{\gamma_r + s_r^2}$ representing reliability weights for the rth principal component and $s_r^2 = \frac{\sigma^2 \mathcal{K}_m}{n}$ $\forall r$. Proof in Appendix B.5

Definition 6 (Smoothed Assignment Rule). Let $\tilde{\mathcal{J}}(\hat{\boldsymbol{\mu}}_m;\kappa)$ be a smoothed assignment rule characterized by smoothing parameter κ. It indicates the "share" of each class c taught by teacher j, and let $\mathcal{X}_{j,r}(\hat{\boldsymbol{\mu}}_m) = \sum_c X_{rc}\tilde{\mathcal{J}}_{j,c}(\hat{\boldsymbol{\mu}};\kappa)$, be a function reporting the smoothed class-exposure for each teacher in each principal component as a function of the estimated effects, where $X_r = [\mathrm{I}_J \otimes u_r']\,\boldsymbol{x}^{\mathcal{J}}$.

Assumption 4 (Approximate Linear Program). Assume that (1) $\hat{\mathcal{J}}_m$ can be approximated arbitrarily well by $\tilde{\mathcal{J}}(\kappa)$, (2) $\mathcal{X}_r(\hat{\boldsymbol{\mu}}_m)$ is continuously differentiable and locally linear near $\boldsymbol{\mu}_m^\star$ (the oracle value-added estimates), and (3) that for $r > r_0$ the assignment responsiveness $\mathbb{E}\left[\nabla\mathcal{X}\right] = \frac{\mathbb{V}(X_r)}{J-1}Q_r$, where Q_r is an orthogonal projection matrix indicating how responsive assignments are to value-added in the rth principal component after accounting for all other principal components and $\mathbb{V}(X_r)$ describes how much variation there is across classes in this component. Finally (4) assume Q_r has rank $J-K+1-\tilde{R}$ where $\tilde{R}$ is the number of redundant dimensions conditional on the other components.

Assumption 5 (Principal Component Distributions). Assume that the information content of the principal components of Γ_K are contrasting, fall quickly, and become diffusely loaded such that for r' greater than some r_0: (1) $\sum_k u_{r',k} = 0$, (2) $\frac{\gamma_{r'}}{s^2_{r'}} < \epsilon$, and (3) $\max_k |u_{r',k}| \leq \frac{\bar{c}}{\sqrt{K}}$ for some $\bar{c}$.

Assumption 6 (Class Composition). Assume that class composition follows a multinomial distribution drawn from $\mathcal{K}_m$ equal-sized groups independently of class size in both the estimation and assignment samples and that there are sufficient students per class for X_r to be approximately normal in $r > r_0$.

Assumption 7 (Feasible Empirical Bayes). Assume that (1) feasible reliability weights $\hat{b}_r$ are unbiased, $\mathbb{E}[\hat{b}_r] = b_r$, and are constructed to be independent of each class-subgroup residuals (2) that feasible $\hat{\gamma}_r$ satisfies the first-order variance approximation:$\mathbb{V}(\hat{\gamma}_r) = \frac{2(\gamma_r + s_r^2)^2}{\tilde{J}_r}$, where the number of effective observations identifying each hyperparameter is $\tilde{J}_r$, and (3) $(\hat{\gamma}_r - \gamma_r) \perp \hat{\sigma}$

Proposition 4. Under Assumptions 1–2 and 4–7, the expected per-student optimism from the optimal assignment predicted by model m is increasing at least linearly in $\mathcal{K}_m$:

$$\begin{aligned}\text{Optimism}(m) &\equiv \mathbb{E}\left[\widehat{\mathcal{W}}_m^{\hat{\mathcal{J}}_m} - \mathcal{W}^{\hat{\mathcal{J}}_m}\right] \\ &= \underbrace{O^\star}_{\substack{\text{Oracle Optimism}}} + \underbrace{r_0 O_{r_0}}_{\substack{\text{Feasible EB Optimism} \\ \text{(High-Signal Components)}}} + \underbrace{(\mathcal{K}_m - r_0)\frac{2\sigma^2 \bar{L}}{J\chi(\mathcal{K}_m)}}_{\substack{\text{Feasible EB Optimism} \\ \text{(Low-Signal Components)}}}\end{aligned}$$

where $O^\star$ is the optimism under the oracle and is constant across $\mathcal{K}_m$ and zero in special cases, O_{r_0} is the feasible EB optimism from high signal-nodes, $\bar{L}$ the average degrees of freedom in the assignment problem, and $\chi(\mathcal{K}_m) \in (0, 1]$ is a decreasing function reflecting how additional subgroups reduce effective sample and increase variability in $\hat{b}_r$. This χ adjustment makes optimism slightly convex in $\mathcal{K}_m$.

Proof in Appendix B.5.

Remark (Oracle Optimism). As discussed with Proposition 3, there are commonly studied cases that imply no oracle optimism. Interestingly, even when these cases do not apply, the oracle optimism still handles low-signal nodes perfectly, such that any oracle optimism arises only from considerations surrounding the first r_0 high signal components.

Remark (Effective Teachers). Proposition 4 assumes that all classes in the estimation sample are drawn independently from an identical distribution of student types. Cross-class sorting beyond this independent baseline will decrease $\chi(\mathcal{K}_m)$ and will lead to greater

optimism. In our empirical setting, we find *much* lower effective teacher counts. For example, empirically we have $\chi(5) \approx 0.87$ and $\chi(10) \approx 0.70$ compared to 0.99 and 0.87 for the multinomial distribution. This implies empirical optimism multipliers in the range of 1.15–1.45× rather than 1.01–1.14×.

Remark (Hyperparameter Contamination). In practice, feasible Empirical Bayes estimates often do not construct leave-one-out estimates of hyperparameters. In this case, there may be additional optimism whenever the hyperparameters are nonlinear functions of the same residuals they shrink. If $\hat{b}_r = \hat{b}_r^{-j} + \frac{s_r^2(A_{j,r}^2 - \gamma_r - s_r^2)}{(\gamma_r + s_r^2)\tilde{J}_r}$, the optimism term in Proposition 4 should have an additional term proportional to $(\mathcal{K}_m - r_0)\frac{\sigma^2 \mathcal{K}_m}{\tilde{J}N}$.

Remark (Changing Leverage). Technically, the two feasible EB terms from Proposition 4 have a leverage component that slightly decreases in $\mathcal{K}_m$. Each new dimension of comparative advantage will absorb some of the residual variation available to other components. Because $\bar{L} = 1 - \frac{K-2}{J-1}$ the effective change with $\mathcal{K}_m$ is small, however, and empirically will be swamped out by $\chi(\mathcal{K}_m)$.

B.5 Proofs of Propositions

B.5.1 Proof of Proposition 1

Proof.

$$\begin{aligned}
\mathbb{W}^{\mathcal{J}} \equiv \mathbb{E}[W^{\mathcal{J}} | \boldsymbol{S}^{\mathcal{J}}, \boldsymbol{\phi}] &= \sum_{i=1}^{n} \mathbb{E}[\phi_i U_i^{\mathcal{J}} | S_i^{\mathcal{J}}, \phi_i] \\
&= \sum_{i=1}^{n} \phi_i \left(\alpha_i + \beta_i \left(\bar{S}_i + \mu_{i,j}^{\mathcal{S}}) \right) \right) \\
&\equiv \bar{W} + \sum_j \sum_{i:\mathcal{J}(i)=j} \omega_i \mu_{i,j}^{\mathcal{S}}
\end{aligned}$$

The first equality follows from the Policy Independence assumption which guarantees no spillovers and makes $S^{\mathcal{J}}$ sufficient for $U^{\mathcal{J}}$. The second equation applies the local linearity of $\mathbb{E}[U|S]$ and the additive separability of $s(\cdot)$. The last equation follows from the surjectivity of $\mathcal{J}$ (all students get taught) and defines $\omega_i = \phi_i \beta_i$ and $\bar{W} = \sum_i \phi_i \alpha_i + \phi_i \beta_i \bar{S}_i$.

An alternative normalization is to assume $\mu_{i,\mathcal{J}(i)}^{\mathcal{S}} = 0$ and to interpret $\mathcal{W}^{\mathcal{J}}$ as differences from the status quo:

$$
\begin{aligned}
\mathbb{W}^{\mathcal{J}} - \mathbb{W}' &\equiv \mathbb{E}[W^{\mathcal{J}}|\boldsymbol{S}^{\mathcal{J}}, \boldsymbol{\phi}] - \mathbb{E}[W^{0}|\boldsymbol{S}^{0}, \boldsymbol{\phi}] \\
&= \sum_{i=1}^{n} \mathbb{E}[\phi_i U_i^{\mathcal{J}}|S_i^{\mathcal{J}}, \phi_i] - \sum_{i=1}^{n} \mathbb{E}[\phi_i U_i^{0}|S_i^{0}, \phi_i] \\
&= \sum_{i=1}^{n} \phi_i \mathbb{E}[U_i^{\mathcal{J}} - U_i^{0}|S_i^{\mathcal{J}}, S_i^{0}] \\
&= \sum_{i=1}^{n} \phi_i \left(\alpha_i + \beta_i S_i^{\mathcal{J}} - \alpha_i - \beta_i S_i^{0}\right) \\
&= \sum_{j} \sum_{i:\mathcal{J}(i)=j} \phi_i \left(\beta_i(\bar{S}_i + \mu_{i,j}^{\mathcal{S}}) - \beta_i(\bar{S}_i + \mu_{i,\mathcal{J}_0(i)}^{\mathcal{S}})\right) \\
&\equiv \sum_{j} \sum_{i:\mathcal{J}(i)=j} \omega_i \mu_{i,j}^{\mathcal{S}}
\end{aligned}
$$

The first and second equalities follow from the Policy Independence assumption which guarantees no spillovers and makes $S^{\mathcal{J}}$ sufficient for $U^{\mathcal{J}}$. The third equation applies the local linearity of $\mathbb{E}[U|S]$. The fourth equation follows from the surjectivity of $\mathcal{J}$ (all students get taught) and the additive separability of $s(\cdot)$. The last equation defines ω_i.

B.5.2 Proof of Proposition 2

Proof.

$$
\begin{aligned}
\mathcal{W}^{\mathcal{J}} - \widehat{\mathcal{W}}_m^{\mathcal{J}} &\equiv \sum_{j} \sum_{i:\,\mathcal{J}(i)=j} \omega_i \mu_{i,j} - \sum_{j} \sum_{k \in \mathcal{K}_m} n_{j,k}(\mathcal{J}) \bar{\omega}_{j,k}(\mathcal{J}) \hat{\mu}_{j,k} \\
&= \sum_{j} \sum_{k \in \mathcal{K}_m} n_{j,k}(\mathcal{J}) \left[\left(\frac{1}{n_{j,k}(\mathcal{J})} \sum_{i} \omega_i \mu_{i,j} \right) - \bar{\omega}_{j,k}(\mathcal{J}) \hat{\mu}_{j,k} \right] \\
&= \sum_{j} \sum_{k \in \mathcal{K}_m} n_{j,k}(\mathcal{J}) \left[\frac{1}{n_{j,k}(\mathcal{J})} \sum_{i} \omega_i \mu_{i,j} + \bar{\omega}_{j,k}(\mathcal{J}) \Big(\tilde{\mu}_{j,k}(\mathcal{J}) - \tilde{\mu}_{j,k}(\mathcal{J}) + \mu_{j,k}^{\star} - \mu_{j,k}^{\star} \right. \\
&\qquad \left. + \bar{\mu}_{j,k} - \bar{\mu}_{j,k} - \hat{\mu}_{j,k} \Big) \right] \\
&= \sum_{j} \sum_{k \in \mathcal{K}_m} n_{j,k}(\mathcal{J}) \bar{\omega}_{j,k}(\mathcal{J}) \Bigg[\underbrace{\tilde{\mu}_{j,k}(\mathcal{J}) - \bar{\mu}_{j,k}}_{\text{Matching Gains}} + \underbrace{\frac{\text{Cov}(\omega_i, \mu_{i,j}; \mathcal{J})}{\bar{\omega}_{j,k}(\mathcal{J})}}_{\text{Distributional Gains}} + \underbrace{\bar{\mu}_{j,k} - \tilde{\mu}_{j,k}(\mathcal{J}_0)}_{\text{External Validity}} \\
&\qquad + \underbrace{\tilde{\mu}_{j,k}(\mathcal{J}_0) - \mu_{j,k}^{\star}}_{\text{Shrinkage Bias}} + \underbrace{\mu_{j,k}^{\star} - \hat{\mu}_{m,j,k}}_{\text{Estimation Error}} \Bigg]
\end{aligned}
$$

The first equality comes from factoring out $n_{j,k}(\mathcal{J})$. The second from adding and subtracting $\tilde{\mu_{j,k}}(\mathcal{J})$, $\bar{\mu}_{j,k}$, and $\mu_{j,k}^{\star}$, where $\bar{\omega}_{j,k}(\mathcal{J})$ and $\tilde{\mu}_{j,k}(\mathcal{J})$ represent the average welfare weights

of students in teacher j's assigned class and the average match effect of teacher j on those students, $\bar{\mu}_{j,k}$ is their population average match effect, and $\mu^{\star}_{j,k}$ is the oracle value-added estimate. The third comes from adding and subtracting $\bar{\omega}_{j,k}(\mathcal{J})\tilde{\mu}_{j,k}(\mathcal{J}_0)$, the average welfare weights times the true average match effect in the class teacher j actually taught in the estimation sample, and from the definition of covariance (we define the within-class covariance as $\text{Cov}(\omega_i, \mu_{i,j}; \mathcal{J}) \equiv \frac{1}{n_{j,k}(\mathcal{J})} \sum_{i:\, \mathcal{J}(i)=j} \omega_i \mu_{i,j} - \bar{\omega}_{j,k}(\mathcal{J})\tilde{\mu}_j(\mathcal{J})$).

B.5.3 Proof of Proposition 3

Proof. The MSE of welfare from evaluating a fixed assignment $\mathcal{J}$ with $\hat{\boldsymbol{\mu}}_m$ is

$$\begin{aligned}
\text{MSE}(\widehat{\mathcal{W}}^{\mathcal{J}}_m) &\equiv \mathbb{E}\left[\left(\widehat{\mathcal{W}}^{\mathcal{J}}_m - \mathcal{W}^{\mathcal{J}}\right)^2\right] \\
&= \mathbb{E}\left[\widehat{\mathcal{W}}^{\mathcal{J}}_m - \mathcal{W}^{\mathcal{J}}\right]^2 + \mathbb{V}\left(\widehat{\mathcal{W}}^{\mathcal{J}}_m - \mathcal{W}^{\mathcal{J}}\right) \\
&= \mathbb{E}\left[\boldsymbol{x}^{\mathcal{J}\prime}\left[D(\hat{\boldsymbol{q}}_m) - \boldsymbol{\mu}_m\right] + M^{\mathcal{J}}_m\right]^2 + \mathbb{V}\left(\boldsymbol{x}^{\mathcal{J}\prime} D(\hat{\boldsymbol{q}}_m)\right) \\
&= \left(\boldsymbol{x}^{\mathcal{J}\prime}\mathbb{E}\left[D(\boldsymbol{q}_m) + (\hat{\boldsymbol{q}}_m - \boldsymbol{q}_m) D_q - \boldsymbol{\mu}_m\right] + M^{\mathcal{J}}_m\right)^2 + \boldsymbol{x}^{\mathcal{J}\prime}\mathbb{V}\left(D(\hat{\boldsymbol{q}}_m)\right)\boldsymbol{x}^{\mathcal{J}} \\
&= \left[\underbrace{M^{\mathcal{J}}_m}_{\text{Misspecification Bias}} + \underbrace{\boldsymbol{x}^{\mathcal{J}\prime}\left(D(\boldsymbol{q}_m) - \boldsymbol{\mu}_m\right)}_{\text{Shrinkage Bias}}\right]^2 + \underbrace{\boldsymbol{x}^{\mathcal{J}\prime} D_q \Omega_q D_q' \boldsymbol{x}^{\mathcal{J}}}_{\text{Welfare Variability}}
\end{aligned}$$

The second equality applies the definition of variance, the third uses the definition of $M^{\mathcal{J}}_m$ and notes that there is no sampling variance in $\mathcal{W}^{\mathcal{J}}$. The fourth equality Taylor expands $D(\hat{\boldsymbol{q}}_m)$ under Assumption 3. The bias term in the fifth equality applies the unbiasedness of $\hat{\boldsymbol{q}}_m$ under Assumption 3 and the fact that D_q is fixed under the oracle. For the variance term, it applies the delta method to get the variance of $D(\cdot)$ from the variance of $\mathbb{V}(\hat{q}) \equiv \Omega_q$.

First consider the estimation error in subgroup-class residuals. We assume that because of sample size, $\hat{\beta}_k$ is essentially independent of the class-level error terms:

$$\begin{aligned}
\mathbb{V}\left(\hat{A}_{j,k} - \mu_{j,k}\right) &= \mathbb{V}\left(\left(\frac{1}{n_{j,k}}\sum_{i\in(j,k)} y_i - \hat{\beta}_k x_i\right) - \mu_{j,k}\right) \\
&= \frac{\mathbb{V}\left(\sum_{i\in(j,k)}\varepsilon_i\right) + \mathbb{V}\left(\sum_{i\in(j,k)}(\hat{\beta}_k - \beta_k)x_i\right)}{n^2_{j,k}} \\
&= \frac{\sigma^2_{\varepsilon k} + \bar{X}'_{j,k}\mathbb{V}\left(\hat{\beta}_k - \beta_k\right)\bar{X}'_{j,k}}{n_{j,k}}
\end{aligned}$$

$$
\begin{aligned}
&\geq \frac{\sigma^2_{\varepsilon k} + \frac{\sigma^2_{\varepsilon k}}{N_k} \bar{X}'_{j,k} \mathbb{E}\left[x_i x_i'\right]^{-1} \bar{X}'_{j,k}}{n/K} \\
&= \frac{\sigma^2_{\varepsilon k} K}{n} \left(1 + \frac{\ell^{\hat{\beta}}_{j,k}}{J}\right)
\end{aligned}
$$

The second equality applies $y_i = x_i'\beta_k + \mu_{j,k} + \varepsilon_i$ and the assumed independence. The third takes the variance of the sums and cancels $n_{j,k}$. The fourth line has an inequality implied by the lowest-case variance scenario that all classes have equal class sizes and class shares: $n_{j,k} = \frac{n}{K}$ where $n = \frac{N}{J}$ and notes that $\mathbb{V}(\hat{\beta}) = \frac{\sigma_{\varepsilon k} K}{N} \mathbb{E}[x_i x_i']^{-1}$. The final equality simplifies.

Next we create an optimistic benchmark on hyperparameter estimation error. We focus on the cross-subgroup correlation parameters $\sigma_{k,k'}$. We assume

$$
\begin{aligned}
\mathbb{V}\left(\hat{\sigma}_{k,k'} - \sigma_{k,k'}\right) &= \mathbb{V}\left(\frac{1}{J-1}\sum_{j=1}^{J}(\hat{A}_{j,k} - \bar{\hat{A}}_k)(\hat{A}_{j,k'} - \bar{\hat{A}}_{k'}) - \frac{1}{J-1}\sum_{j=1}^{J}(A_{j,k} - \bar{A}_k)(A_{j,k'} - \bar{A}_{k'})\right) \\
&= \frac{1}{(J-1)^2}\mathbb{V}\left(\sum_{j=1}^{J}(A_{j,k} - \bar{A}_k)\eta_{j,k'} + (A_{j,k'} - \bar{A}_{k'})\eta_{j,k} + \left(\eta_{j,k}\eta_{j,k'} - \mathbb{E}[\eta_{j,k}\eta_{j,k'}]\right)\right) \\
&\geq \frac{1}{J-1}\left(\sigma_{k,k} s^2_{k'} + \sigma_{k',k'} s^2_k + s^2_k s^2_{k'}\right) \qquad \text{if } k \neq k' \\
&\geq \frac{2}{J-1}\left(2\sigma_{k,k} s^2_k + s^4_k\right) \qquad \text{if } k = k'
\end{aligned}
$$

Here the second equality follows from defining $\eta_{j,k} = A_{j,k} - \hat{A}_{j,k}$ distributing terms so that the $(A - \bar{A})$ terms in $\hat{\sigma}$ cancel with σ. Then the third inequality follows from the definition of variance under independence of shocks between k and k', or the fourth for $k = k'$—where the inequality follows from s^2_k being a lower bound if all cross-group correlations are positive.

B.5.4 Proof of Lemma 1

Proof. Because $\mathbb{V}(\hat{\boldsymbol{A}}_k - \boldsymbol{\mu}_k) = s^2 I_K$, the signal covariance is $\mathbb{V}(\hat{\boldsymbol{A}}_j) = \Gamma_K + s^2 I_K$. and the oracle linear projection of $\boldsymbol{\mu}_j$ on $\hat{\boldsymbol{A}}_j$ is $\mathbb{E}[\boldsymbol{\mu}_j \mid \hat{\boldsymbol{A}}_j] = B_K \hat{\boldsymbol{A}}_j$ where $B_K = \Gamma_K(\Gamma_K + s^2 I_K)^{-1}$. Because s^2 is a scalar constant, $\Gamma_K(\Gamma_K + s^2 I_K)^{-1} u_r = \frac{\gamma_r}{\gamma_r + s^2} u_r$, and therefore $U' B_K U = \text{diag}(b_1, \ldots, b_K)$, with

$$
b_r = \frac{\gamma_r}{\gamma_r + s^2}.
$$

B.5.5 Proof of Proposition 4

Proof. $\text{Optimism}(m) \equiv \mathbb{E}\left[\widehat{\mathcal{W}}^{\hat{\mathcal{J}}_m}_m - \mathcal{W}^{\hat{\mathcal{J}}_m}\right]$

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\mathcal{W}}_m^{\hat{\mathcal{J}}_m} - \mathcal{W}^{\hat{\mathcal{J}}_m}\right] &= \mathbb{E}\left[\hat{\boldsymbol{x}}'\left(\hat{\boldsymbol{\mu}}_m - \boldsymbol{\mu}_m\right)\right] \\
&= \sum_r \sum_j \mathbb{E}\left[\mathcal{X}_{j,r}(\hat{\boldsymbol{\mu}})\left(\hat{\mu}_{j,r} - \mu_{j,r}\right)\right] \\
&= \sum_r \sum_j \mathbb{E}\left[\left[\mathcal{X}_{j,r}(\boldsymbol{\mu}^\star) + \nabla\mathcal{X}_{j,r}(\boldsymbol{\mu}^\star)\left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\star\right)\right]\left(\hat{\mu}_{j,r} - \mu_{j,r}\right)\right] \\
&= \sum_r \sum_j \Bigg[\mathbb{E}\left[\mathcal{X}_{j,r}(\boldsymbol{\mu}^\star)\left(\mu_{j,r}^\star - \mu_{j,r}\right)\right] + \mathbb{E}\left[\mathcal{X}_{j,r}(\boldsymbol{\mu}^\star)\left(\hat{\mu}_{j,r} - \mu_{j,r}^\star\right)\right] \\
&\qquad + \mathbb{E}\left[\nabla\mathcal{X}_{j,r}(\boldsymbol{\mu}^\star)\left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\star\right)\left(\mu_{j,r}^\star - \mu_{j,r}\right)\right] \\
&\qquad + \mathbb{E}\left[\nabla\mathcal{X}_{j,r}(\boldsymbol{\mu}^\star)\left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\star\right)\left(\hat{\mu}_{j,r} - \mu_{j,r}^\star\right)\right]\Bigg] \\
&= \sum_r \sum_j \mathbb{E}\left[\mathcal{X}_{j,r}(\boldsymbol{\mu}^\star)\left(\mu_{j,r}^\star - \mu_{j,r}\right)\right] \\
&\qquad + \mathbb{E}\left[\nabla\mathcal{X}_{j,r}(\boldsymbol{\mu}^\star)\mathrm{Cov}\left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\star, \hat{\mu}_{j,r} - \mu_{j,r}^\star\right)\right] \\
&= \underbrace{\sum_r \sum_j \mathbb{E}\left[\mathcal{X}_{j,r}(\boldsymbol{\mu}^\star)\left(\mu_{j,r}^\star - \mu_{j,r}\right)\right]}_{\text{Oracle Optimism (integrates to 0 under prior)}} \\
&\quad + \underbrace{\sum_r \sum_j \mathbb{V}\left(\hat{b}_r\right)\mathbb{E}\left[\frac{\partial\mathcal{X}_{j,r}}{\partial\mu_{j,r}^\star}\hat{A}_{j,r}^2\right] + \mathbb{V}\left(\hat{b}_r\right)\mathbb{E}\left[\sum_{j'\neq j}\frac{\partial\mathcal{X}_{j,r}}{\partial\mu_{j',r}^\star}\hat{A}_{j,r}\hat{A}_{j',r}\right]}_{\text{Feasible EB Optimism Penalty}}
\end{aligned}
\tag{B.1}
$$

The second equality comes from mapping welfare into the true principal components and the assumption that $\mathcal{X}_r(\boldsymbol{\mu})$ approximates $\hat{\boldsymbol{x}}_r$ arbitrarily well. The third uses a Taylor approximation of $\mathcal{X}$ around $\boldsymbol{\mu}^\star$ which is locally linear under assumption 4. The fourth separates $(\hat{\mu} - \mu)$ into $(\hat{\mu} - \mu^\star)$ and $(\mu^\star - \mu)$ and distributes it across terms. The fourth equality applies the fact that because $\mathbb{E}[\hat{b}_r] = b_r$ and is independent from $\hat{A}_{j,r}$, $\mathbb{E}[\hat{\boldsymbol{\mu}}_r] = \mathbb{E}[\hat{b}_r]\mathbb{E}[\hat{\boldsymbol{A}}_r] = b_r\boldsymbol{\mu}_r \equiv \boldsymbol{\mu}^\star$; therefore, the second and third terms both equal zero and the covariance enters the fourth term. The fifth equation separates the covariances into sums over (j, r) and (j', r') pairs, noting that because each principal component is orthogonal the r' terms are zero, but the j' terms remain because either residual may affect a given teacher's assignment in the rth dimension.

We consider each of these terms in turn.

Oracle Optimism. The oracle optimism reports how far the expected welfare would be from the realized welfare if the shrinkage hyperparameters were known. One benefit of

Empirical Bayes is that optimism will be zero when integrated over the prior. This property reflects the fact that while $\mathbb{E}_\epsilon[\mu^\star - \mu|\mu] \neq 0$, $\mathbb{E}_{\mu,\epsilon}[\mu^\star - \mu] = 0$, so over repeated draws of teachers, there will be no optimism on average under assignment $\hat{\mathcal{J}}$. In cases, however, where the social planner cares about welfare given the *realized* teacher effects, some optimism will persist.

This optimism has two components: oracle winner's curse and shrinkage bias. To see this, apply the definition of covariance and then Stein's Lemma to the Oracle Optimism term above.

$$\begin{aligned} O_r^\star &\equiv \sum_j \mathbb{E}\left[\mathcal{X}_{j,r}(\boldsymbol{\mu}^\star)\left(\mu_{j,r}^\star - \mu_{j,r}\right)\right] \\ &= \sum_j \text{Cov}\left(\mathcal{X}_{j,r}(\boldsymbol{\mu}^\star), \mu_{j,r}^\star - \mu_{j,r}\right) + \mathbb{E}\left[\mu_{j,r}^\star - \mu_{j,r}\right]\mathbb{E}\left[\mathcal{X}_{j,r}(\boldsymbol{\mu}^\star)\right] \\ &= \sum_j \underbrace{b_r^2 \frac{\partial \mathcal{X}_{j,r}}{\partial \mu_{j,r}^\star} s_r^2}_{\text{Oracle Winner's Curse}} \underbrace{- (1-b_r)\mu_{j,r}\mathbb{E}[\mathcal{X}_{j,r}]}_{\text{Shrinkage Bias}} \end{aligned} \tag{B.2}$$

This oracle winner's curse term is analogous to the feasible EB penalty but without estimation error in b_r. Also note that whereas winner's curse is strictly positive, shrinkage bias actually *reduces* optimism. For high-signal principal components, these two terms could sum to something positive or negative unless the social planner integrates over all possible draws of teacher effects—in which case they exactly cancel.

Whereas high-signal principal components follow Equation B.2, both terms go to zero for low-signal principal components. To see this, note that when $\frac{\gamma_r}{s_r^2} < \epsilon$, b_r^2 goes to zero quickly. This pattern indicates that the oracle correctly downplays the new information with small to no marginal increase in winner's curse. Furthermore, the small γ_r implies a small shrinkage bias since $\mu_{j,r}$ is close to zero: $\sum_j \mu_{i,j}\mathbb{E}[\mathcal{X}_{j,r}] \approx J\text{Cov}_J(\mu_{j,r}, \mathbb{E}[\mathcal{X}_{j,r}]) = J\frac{\gamma_r n}{K}\rho \leq J\sigma^2\epsilon \approx 0$.

Combining these results, we get the following formula for Oracle Optimism which is always constant in $\mathcal{K}_m$ and which has a closed-form approximation in special cases:

$$\begin{aligned} O^\star \equiv \sum_{r=1}^{K} O_r^\star &= \sum_{r=1}^{r_0}\sum_j b_r^2 s_r^2 \frac{\partial \mathcal{X}_{j,r}}{\partial \mu_{j,r}^\star} - (1-b_r)\mu_{j,r}\mathbb{E}[\mathcal{X}_{j,r}] \\ &\approx \frac{\sigma}{n} \end{aligned} \tag{B.3}$$

The final approximation holds in the special case where absolute advantage is the only principal component with high signal to noise and it is very precise, i.e., $r_0 = 1$ and $b_r = 1$.

Feasible EB Optimism Penalty. The optimism created by using feasible Empirical Bayes instead of the oracle depends on the variance of the estimated hyperparamters, $\hat{b}_r$, and on the assignment sensitivity $\frac{\partial \mathcal{X}}{\partial \mu}$. The own-effect assignment sensitivity, $\frac{\partial \mathcal{X}_j}{\partial \mu_j}$, will be positive (conditional on effects and assignments in other dimensions, r) while the cross-effects will generally be small or zero.

We begin with characterizing the variance of $\hat{b}_r$. Recall that the true $b_r = \frac{\gamma_r}{\gamma_r + s_r^2}$. If the variance of $\hat{\gamma}_r = \frac{2(\gamma_r + s_r^2)^2}{\tilde{J}_r}$, then by the delta method

$$\begin{aligned}\mathbb{V}\left(\hat{b}_r\right) = \frac{\partial \hat{b}_r}{\partial \gamma_r}^2 \mathbb{V}\left(\hat{\gamma}_r\right) &= \left(\frac{s_r^2}{(\gamma_r + s_r^2)^2}\right)^2 \frac{2(\gamma_r + s_r^2)^2}{\tilde{J}_r} \\ &= \frac{2 s_r^4}{\tilde{J}_r (\gamma_r + s_r^2)^2}\end{aligned} \tag{B.4}$$

So the variance decreases with the signal-to-noise ratio and with the effective number of teachers identifying each b_r.

Next we note that the optimism formula for the high-signal nodes does not simplify much. In high-signal nodes, the residuals $\hat{\boldsymbol{A}}_r$ are used to determine $\mathcal{X}(\boldsymbol{\mu}^\star) = \mathcal{X}(b_r \hat{\boldsymbol{A}})$, so even though A_j and $A_{j'}$ are independent, neither is independent of $\frac{\partial \mathcal{X}}{\partial \mu^\star}$. As such we define the average high-node optimism penalty from feasible EB:

$$O_{r_0} = \frac{\mathbb{V}\left(\hat{b}_r\right)}{r_0} \sum_{r=1}^{r_0} \sum_j \mathbb{E}\left[\frac{\partial \mathcal{X}_{j,r}}{\partial \mu_{j,r}^\star} \hat{A}_{j,r}^2\right] + \mathbb{E}\left[\sum_{j' \neq j} \frac{\partial \mathcal{X}_{j,r}}{\partial \mu_{j',r}^\star} \hat{A}_{j,r} \hat{A}_{j',r}\right] \tag{B.5}$$

which is somewhat unwieldy but is constant in $\mathcal{K}_m$.

Low signal components have a much more tractable formula. In low signal r, $b_r \to 0$, so the assignments are only based only on considerations from high-signal components, breaking the covariance between A_r and $\frac{\partial \mathcal{X}_r}{\partial \mu_r^\star}$. Thus for $r > r_0$

$$\begin{aligned}\sum_j \mathbb{E}\left[\frac{\partial \mathcal{X}_{j,r}}{\partial \mu_{j,r}^\star} \hat{A}_{j,r}^2\right] + \sum_j \mathbb{E}\left[\sum_{j' \neq j} \frac{\partial \mathcal{X}_{j,r}}{\partial \mu_{j',r}^\star} \hat{A}_{j,r} \hat{A}_{j',r}\right] &= \sum_j \mathbb{E}\left[\frac{\partial \mathcal{X}_{j,r}}{\partial \mu_{j,r}^\star}\right] \mathbb{E}\left[\hat{A}_{j,r}^2\right] + \sum_j \mathbb{E}\left[\sum_{j' \neq j} \frac{\partial \mathcal{X}_{j,r}}{\partial \mu_{j',r}^\star}\right] \mathbb{E}\left[\hat{A}_{j,r} \hat{A}_{j',r}\right] \\ &= \sum_j \frac{\mathbb{V}(X_r)}{J-1} Q_r \left(s_r^2 + \gamma_r\right) + 0 \\ &= \frac{n}{K} L_r \left(s_r^2 + \gamma_r\right)\end{aligned} \tag{B.6}$$

The second equality applies Assumption 4 to the first term and notes that A_j and $A_{j'}$ are

independent. The third uses the variance of a multinomial class composition $\left(\frac{n}{K}\right)$ and defines $L_r = \frac{\text{rank}(Q_r)}{J-1} = \frac{J-K+1}{J-1}$, L_r contains the residual leverage available to make assignments based on component r (out of $J-1$ unconstrained degrees of freedom). The fact that Q_r remains of full rank comes from the independence of u_r across subgroups.

Combining these terms back into Equation B.1, we derive the Proposition 4 result:

$$\begin{aligned}
\text{Optimism}(m) = & \underbrace{\sum_r \sum_j \mathbb{E}\left[\mathcal{X}_{j,r}(\boldsymbol{\mu}^\star)\left(\mu^\star_{j,r} - \mu_{j,r}\right)\right]}_{\text{Oracle Optimism (integrates to 0 under prior)}} \\
& + \underbrace{\sum_r \sum_j \mathbb{V}\left(\hat{b}_r\right)\mathbb{E}\left[\frac{\partial \mathcal{X}_{j,r}}{\partial \mu^\star_{j,r}}\hat{A}^2_{j,r}\right] + \mathbb{V}\left(\hat{b}_r\right)\mathbb{E}\left[\sum_{j'\neq j}\frac{\partial \mathcal{X}_{j,r}}{\partial \mu^\star_{j',r}}\hat{A}_{j,r}\hat{A}_{j',r}\right]}_{\text{Feasible EB Optimism Penalty}} \\
& = O^\star + r_0 O_{r_0} + \sum_{r>r_0} \frac{2s_r^4}{\tilde{J}_r(\gamma_r + s_r^2)^2}\frac{n}{K}L_r\left(s_r^2 + \gamma_r\right) \\
& = O^\star + r_0 O_{r_0} + \sum_{r>r_0} \frac{2\sigma^2}{\tilde{J}_r}\bar{L} \\
& = O^\star + r_0 O_{r_0} + (\mathcal{K}_m - r_0)\frac{2\sigma^2\bar{L}}{J\chi(\mathcal{K}_m)}
\end{aligned}$$

The second equality applies the results from Equations B.3–B.6. The third removes the small $\frac{\gamma_r}{s_r^2}$ terms and simplifies using the definition of $s_r^2 = \frac{\sigma^2 K}{n}$, and notes that $L_r = \bar{L}\forall r$. The fourth sums over all remaining $r > r_0$ noting that because the u_r are diffuse for higher r, $\mathbb{E}[\tilde{J}_r] = J\chi(\mathcal{K}_m)$, where $\chi(\mathcal{K}_m)$ is a decreasing function determining the joint overlap of subgroups under the multinomial distribution.

C. Data and Estimation Details

C.1 Data Description

Our data from the San Diego Unified School District (SDUSD) are administrative, linked data between students and teachers for the 1998–1999 through 2018–2019 school years. The outcomes we use from the data include test scores, attendance, and behavioral GPA. The data also include ethnicity and gender for both students and teachers, as well as information on teachers' credentials.

Test Scores. California had no statewide tests in the 2013–2014 school year because that year it transitioned from one test administered in spring 2013 to a new test in spring 2015. Throughout this time period, the district had three different testing regimes (summarized in Table C.1). In the early years of our data, students in grades 2–6 took the Stanford Achievement Test, Ninth Edition (SAT9). During the 2001–2002 school year, students transitioned to taking the California Standards Test (CST) each year in grades 2–6. This continued until the 2012–2013 school year. Thereafter, the district shifted to the California Assessments of Student Performance and Progress (CAASPP), administered as Smarter Balanced Summative Assessments to students in grades 3–8. For our purposes, we focus on students in grades 3–5.

Table C.1. Testing Regime by Grade and Year (With Tests in Spring)

	1998–2001	2002–2013	2015–2019
Grade 2	SAT9	CST	-
Grade 3	SAT9	CST	CAASPP/Smarter Balanced
Grade 4	SAT9	CST	CAASPP/Smarter Balanced
Grade 5	SAT9	CST	CAASPP/Smarter Balanced

Noncognitive Outcomes. In addition to using these test score outcomes for reading and math in each year (standardized yearly at the district level), we also include as outcomes the fraction of the year that a student attended school and behavioral GPA, standardized yearly at the district level.

The measurement of absences is straightforward, but the measurement of behavioral GPA is more nuanced. We transformed teachers' categorical ratings of various elements of behavior into linear scales, averaged them, and then standardized them by grade and year. There were two forms of the elementary school report card over time. Between the school years ending in 2002 through 2007, the four variables we used included measures on a five-

point scale for whether the student begins work promptly, follows directions, and measures of self-discipline and overall classroom behavior. Thus, the first three variables focus on a student's attentiveness to classwork, while the fourth is more about overall emotional behavior.

In 2008, some schools began a transition to a new standards-based report card that was then used in all later years. We use three variables from this newer report card. Two are similar to the first three variables above, in that they indicate how diligent and attentive the student is to classwork. These variables are whether the student shows interest in learning and completes assignments when due. The third new variable, whether the student respects people and property, is similar to the overall classroom behavior variable in the older system. Another difference is that in the new report card, instead of reporting on a five-point scale teachers reported on a three-point scale. We handle this issue by standardizing so that the overall behavioral GPA has a mean of zero and a standard deviation of one for each grade and year.

As a check on the consistency of the behavioral GPA in 2008 relative to earlier years, we calculated the correlation between behavioral GPA at the student level in year t and $t-1$ for t corresponding to spring 2003 through 2011, and checked for a large drop in the autocorrelation in spring 2008, which marked the transition to the new report card format. From spring 2003 through 2007, the correlation ranged from 0.632 to 0.664; in the key transition year of 2008, the correlation was very close to this range, at 0.610. From 2009 through 2011, the range was 0.554 to 0.568, perhaps reflecting slightly greater noise due to the use of three rather than four variables in the new index.

C.2 Sample Creation

We generate both an estimation sample and a policy counterfactual sample using the SDUSD data. We start with a sample of 582,579 student-year observations in third- through fifth-grade classrooms from the 1998–1999 through 2018–2019 school years and make five restrictions intended to target typical teachers in SDUSD teaching traditional third- through fifth-grade classes.

- We drop students who are absent from their assigned class more than 50 percent of the time. This restriction is nominal, dropping less than 0.1% of students.
- We drop classrooms teaching only special education students. This drops 6,293 student-year observations.
- We drop mixed-grade classrooms. Because SDUSD pushed to eliminate mixed-grade classrooms in 2002, this is our biggest sample restriction in the early years, dropping

70,052 student-year observations, 40% of which are prior to the 2002–2003 school year. This restriction is important to guarantee consistency when reassigning teachers across classes.

- We drop classes outside of the student-weighted 2.5th and 97.5th percentiles of class size, limiting our sample to classes between 13 and 35 students. This drops an additional 31,710 student-year observations, all reflecting non-typical teaching situations. Including these classes results in much larger gains by placing the best teachers in these few enormous classes.

- We drop grade repeaters from our estimation and reassignment sample, since identifying prior-year test scores to estimate teacher value added is fraught. This is a small limitation, dropping 8,047 student-year observations across the 20 school years in our sample.

These restrictions leave us with 466,150 student-year observations.

In the estimation sample, we include only students with at least two consecutive years of data on at least one relevant outcome. This leaves us with 378,399 student-year observations to estimate teacher value added for 2,306 unique teachers across 19 years. For the policy counterfactual sample, we include all students but limit to the third-grade cohorts of 2002–2003 through 2010–2011. This includes 73,235 students.

Table C.2 shows the composition of classes across grades in both the estimation and policy counterfactual samples. There are a few main takeaways from this table. First, the typical class (regardless of grade) is balanced on prior achievement in math and reading, prior behavioral GPA and attendance, and gender. Just under 60 percent of students in the district are Black or Hispanic, and this is reflected in the average racial/ethnic composition of classes in our two samples. Second, differences across grades and samples in the composition of classes along these dimensions are not meaningful, though classes in the policy counterfactual sample perhaps have a slightly higher fraction (about 6 percentage points) of students with worse prior behavior and attendance than those in the estimation sample. This amounts to an average of roughly one more student below the median in these categories relative to the estimation sample. Third, the dispersion of classes on these metrics can be seen clearly from the 25th and 75th percentiles of each. The widest dispersion occurs in the math, reading, and racial/ethnic composition of classes. This reflects the geographic sorting of students across schools in the district and is one reason why we observe the largest gains from reassigning teachers using these dimensions.

Table C.2. Summary Statistics by Grade and Sample

	Estimation Sample			Reallocation Sample		
Grades	3rd	4th	5th	3rd	4th	5th
Below median Math	0.51	0.51	0.51	0.51	0.51	0.51
	[0.31, 0.71]	[0.32, 0.70]	[0.32, 0.70]	[0.32, 0.70]	[0.33, 0.69]	[0.34, 0.70]
Below median Reading	0.52	0.52	0.52	0.51	0.52	0.52
	[0.30, 0.73]	[0.32, 0.73]	[0.32, 0.72]	[0.31, 0.71]	[0.33, 0.72]	[0.33, 0.71]
Below median Behavior	0.48	0.48	0.49	0.51	0.51	0.56
	[0.33, 0.60]	[0.33, 0.60]	[0.33, 0.62]	[0.38, 0.65]	[0.38, 0.63]	[0.40, 0.71]
Below median Attendance	0.46	0.45	0.46	0.52	0.50	0.51
	[0.37, 0.57]	[0.38, 0.55]	[0.38, 0.56]	[0.41, 0.63]	[0.42, 0.61]	[0.42, 0.62]
Female	0.49	0.49	0.49	0.50	0.50	0.50
	[0.43, 0.55]	[0.44, 0.55]	[0.44, 0.54]	[0.43, 0.56]	[0.44, 0.55]	[0.44, 0.55]
Black/Hisp	0.59	0.61	0.60	0.60	0.62	0.62
	[0.35, 0.86]	[0.36, 0.87]	[0.34, 0.86]	[0.35, 0.88]	[0.38, 0.89]	[0.36, 0.88]

Note: Each row represents the average fraction of the given category across classes in the given sample. The square brackets contain the 25th and 75th percentiles of the given variable.

C.3 Value Added Estimation

C.3.1 Model and Identification

We model student achievement following Delgado (2025) and Bates et al. (2025), as a function of observable student characteristics, teacher value added, teacher experience, school effects, time effects, and classroom shocks specific to student subtypes, as follows:

$$y_{i,t} = \mu^y_{\mathcal{J}(i,t),k,t} + \beta^y_k X_{i,t} + \phi^y_{\ell(i,t)} + \phi^y_t + \theta^y_{\mathcal{C}(i,t),k} + \varepsilon^y_{i,t}$$

where $y_{i,t}$ is the outcome from student i in year t, $\mu^y_{\mathcal{J}(i,t),k,t}$ is the value added from the assigned teacher j for students of type k, $X_{i,t}$ are the observed student characteristics, $\phi^y_{\ell(i,t)}$ are all school-specific factors impacting student achievement on outcome y from the student's assigned school $\ell(i,t)$, ϕ^y_t are outcome specific time effects, $\theta^y_{\mathcal{C}(i,t),k}$ are outcome-specific classroom shocks to students of type k for assigned classroom $\mathcal{C}(i,t)$, and $\varepsilon^y_{i,t}$ are

idiosyncratic student-level shocks.

Because we are studying primary school teachers, students and teachers are typically assigned to one and only one class each year. As in Delgado (2025), this means that we cannot separately identify teacher value added from classroom shocks.

We make two standard assumptions to identify our value-added model following Delgado (2025) and Bates et al. (2025).

Assumption 8. (Joint Stationarity) Assume that subgroup-specific teacher effects, classroom shocks, and student-level shocks follow a stationary process.

$$\begin{aligned}
\mathbb{E}[\mu^y_{\mathcal{J}(i,t),k,t}|y,k,t] = \mathbb{E}[\theta^y_{C(i,t),k,t}|y,k,t] = \mathbb{E}[\varepsilon^y_{i,t}|y,k,t] = 0 \\
\text{Cov}(\mu^y_{\mathcal{J}(i,t),k,t}, \mu^{y'}_{\mathcal{J}(i,t),m,t+h}) = \sigma_{\mu^y_k,\mu^{y'}_m,h} \\
\text{Cov}(\theta^y_{C(i,t),k,t}, \theta^{y'}_{C(i,t),m,t}) = \sigma_{\theta^y_k,\theta^{y'}_m} \\
\text{Cov}(\varepsilon^y_{i,t}, \varepsilon^y_{i,t+h}) = \sigma_{\varepsilon^y_k,h} \\
\text{Cov}(\varepsilon^y_{i,t}, \varepsilon^{y'}_{i,t}) = \sigma_{\varepsilon^y,\varepsilon^{y'}}
\end{aligned}$$

Assumption 9. (Fixed-Effect Independence) Let $\bar{\alpha}^y_{j,k}$ be teacher j's mean value added for student type k and let $\ell(j,t)$ return teacher j's assigned school in year t. Assume that teachers' drift is independent of the school effects on each outcome.

$$(\alpha^y_{j,k,t} - \bar{\alpha}^y_{j,k}) \perp \phi^y_{\ell(j,t)} \forall k, y$$

C.3.2 Estimation Details

Estimation follows Bates et al. (2025) and Delgado (2025), with adaptations for multidimensionality. We follow the first three steps of the four-step estimation procedure in Bates et al. (2025) exactly for each outcome, y.

First, we regress our outcome $y_{i,t}$ on a set of student characteristics, $X_{i,t}$ and teacher-subgroup-year fixed effects, $\alpha^y_{\mathcal{J}(i,t),k,t}$ for each outcome and student subgroup:

$$y_{i,t} = \beta^y_k X_{i,t} + \alpha^y_{\mathcal{J}(i,t),k,t} + v^y_{i,t}$$

where i indexes students, y indexes outcomes, t indexes years, k indexes student subgroups, and $\mathcal{J}(i,t)$ indexes the teacher assigned to student i in year t. In this regression, $X_{i,t}$ includes cubic polynomials in prior-year math, reading, attendance, and behavior, each interacted with student grade level. We also include ethnicity, gender, age, lagged suspensions and

absences, and indicators for special education and English language learner status.[53]

Second, we form residuals from this initial regression, $\hat{v}^y_{i,t}$, by subtracting the effects of student covariates (but not teacher fixed effects):

$$\hat{v}^y_{i,t} = y_{i,t} - \hat{\beta}^y_k X_{i,t}$$

and project these residuals separately for each outcome y onto teacher (α^y_j), school ($\phi^y_{\ell(i,t)}$), and year fixed effects (ϕ^y_t), as well as a teacher experience function $f(y,z)$. We specify this function to allow for different returns for each year of experience up to 6 years then one estimate for all years of experience thereafter (i.e., 7+ years). We estimate this regression separately for each outcome, allowing different estimates for each component across outcomes y.

$$\begin{aligned}\hat{v}^y_{i,t} &= \sum_{e=1}^{6} \delta^y_e \mathbb{1}\{z_{\mathcal{J}(i,t),t} = e\} + \delta^y_7 \mathbb{1}\{z_{\mathcal{J}(i,t),t} \geq 7\} + \alpha^y_{\mathcal{J}(i,t)} + \phi^y_{\ell(i,t)} + \phi^y_t + \eta^y_{i,t} \\ &= f(y, z_{\mathcal{J}(i,t),t}) + \alpha^y_{\mathcal{J}(i,t)} + \phi^y_{\ell(i,t)} + \phi^y_t + \eta^y_{i,t}\end{aligned}$$

Third, we form a second set of student-level residuals in which we remove the school effects and teacher-experience effects:

$$\begin{aligned}A_{y,i,k,t} &= \hat{v}^y_{i,t} - \left(\hat{\phi}^y_{\ell(i,t)} + \hat{f}(y, z_{\mathcal{J}(i,t),t})\right) \\ &= y_{i,t} - \hat{\beta}^y_k X_{i,t} - \hat{\phi}^y_{\ell(i,t)} - \hat{f}(y, z_{\mathcal{J}(i,t),t})\end{aligned}$$

which we can aggregate into average residuals for each teacher-year-subgroup-outcome:

$$\bar{A}_{y,j,k,t} = \frac{1}{n_{j,k,t}} \sum_{i:\mathcal{J}(i,t)=j, k_i=k} A_{y,i,k,t}$$

Finally, we then stack average residuals from all outcomes and student-subgroups from years prior to t into a single vector for each teacher, $\boldsymbol{A}^{-t}_j$ and estimate teacher j's vector of value added for each outcome y and student subgroup k in year t using only the prior data:

$$\boldsymbol{\mu}_{j,t} = \boldsymbol{\psi}'_j \boldsymbol{A}^{-t}_j \tag{C.1}$$

The matrix $\boldsymbol{\psi}'_j$ is a teacher-specific set of reliability weights with dimensions $M \times M(t-1)$,

[53] One small difference from Bates et al. (2025) is that they include student socioeconomic status, proxied by free and reduced-price lunch—which is not provided to researchers by SDUSD.

where M is the number of outcome-subgroup divisions.[54] The reliability weights capture the population relationship between average residuals in year t and prior years and are adjusted for teacher-specific signal strength (reflected by the number of students taught in each subgroup and intersection of subgroups).

Following the notation of Delgado (2025), for each m, $\boldsymbol{\psi}_j$ can be written as $\Gamma_j^{-1}\boldsymbol{\gamma}$. We denote $\Gamma_j = \Gamma - D_j + \lambda I$. The first component of Γ_j, Γ, captures the overall variance-covariance structure across time, subgroups, and outcomes. This is a shared block variance-covariance matrix of dimension $(M(T-1) \times M(T-1))$, where each block reports the t-autocorrelations of m and m'. The second component, D_j, is a teacher-specific matrix for information adjustment based on sample size. This block matrix is composed of $M \times M$ diagonal blocks, where the $T-1$ diagonal elements are $\frac{n_{m,m',t}\sigma_{\epsilon_{m,m'}}}{n_{m,t}n_{m',t}}$ where $n_{m,m',t}$ is the intersection of students in both subgroup-outcome cells m and m'. For example, in the diagonal blocks this corresponds to $\frac{\sigma_{\epsilon_k}}{n_k}$ as in Bates et al. (2025) and Delgado (2025), but in the presence of multidimensionality, it also adjusts for within-student correlations across outcomes $\sigma_{\epsilon^y,\epsilon^{y'}}$ in the off-diagonal blocks. Intuitively, this adjustment makes Γ_j larger and more spherical when we have noisier information about teacher j's effects, shrinking $\psi_{j,m,t}$ toward zero. The third component, λI, is a ridge-type adjustment that keeps Γ_j positive definite despite the D_j adjustment—which can be very noisy in high-dimensional cases. We discuss details of the cross-validated selection, properties, and robustness of λ across models in Appendix C.4. Finally, the $\boldsymbol{\gamma}$ matrix is an $(M(T-1) \times M)$ matrix of correlations between average residuals for each subgroup and outcome in periods $t-h$ and period t.

In addition to the regularization, the only other substantive deviation from the approaches of Delgado (2025) and Bates et al. (2025) is highlighted in Equation C.1. Whereas those papers studied only one outcome at a time, we use information from all outcomes and subgroups to predict each value-added estimate. As such, $\boldsymbol{\mu}_{j,t}$ is a $M \times 1$ vector capturing the teacher's impact on each student subgroup and outcome. Since we only observe one teacher per class, we combine the structural estimation of σ_{θ_k} and σ_{μ_k} as in Delgado (2025) rather than Bates et al. (2025), though both papers differ only slightly in their estimation of the relevant structural parameters. We also impose a drift limit of seven (as in Bates et al. (2025)), after which all elements of the autocorrelation vector are set to $\sigma_{\mu_k^y,\mu_m^{y'},8}$ for any k, m, y, y'. Our final estimate for teacher j's value added in year t on outcome y for students in subgroup k is as follows:

$$\hat{\mu}^y_{j,k,t} = \hat{\boldsymbol{\psi}}^y_{j,k}\boldsymbol{A}_j^{-t} + \hat{f}(y, z_{\mathcal{J}(i,t),t})$$

[54] Note that students need not be subdivided into the same number of subgroups for each outcome (e.g., one could estimate three subgroups for reading and two for math, meaning that $M = 5$)

where $\hat{\boldsymbol{\psi}}^{y}_{j,k}$ are the relevant elements of $\boldsymbol{\psi}'_j$.

C.4 Regularization of Γ_j

When estimating additional value-added parameters on the same sample, Γ_j can become ill-conditioned and generate low-quality estimates. This can occur with both multidimensionality and heterogeneity, as each additional parameter requires estimating $(T-1) \times M$ additional hyperparameters. Although the population Γ is guaranteed to be positive definite, sampling noise can lead to small (or even negative) eigenvalues in $\Gamma - D_j$ and a large condition number. This causes large swings in the reliability weights and produces value-added estimates that have literally no predictive power for the current-year test scores ($\beta = 0$ or 100% forecast bias). This occurs almost exclusively in the highest-dimensional models, i.e., those with $M > 6$.

We address this issue using a ridge-type regularization. Rather than treating $\Gamma_j = \Gamma - D_j$ as in Delgado (2025), we add a small positive number, λ, to the diagonal elements of Γ_j affected by D_j. This well-known solution to ill-conditioning (Tikhonov et al., 1995) works because adding λ to the diagonals of Γ_j increases all eigenvalues by λ. This ensures that the matrix is positive definite, reduces the condition number (by adding λ to both the numerator and the denominator), and guarantees that Γ_j is well behaved when inverted. The resulting reliability weights $\hat{\boldsymbol{\psi}}_j = (\Gamma - D_j + \lambda I)^{-1}\gamma$ share intuition with a ridge regression (Hoerl and Kennard, 1970), and similar corrections have been used in other empirical applications, such as the estimation of price elasticities (Chernozhukov et al., 2019) and financial returns (Martin and Nagel, 2022).

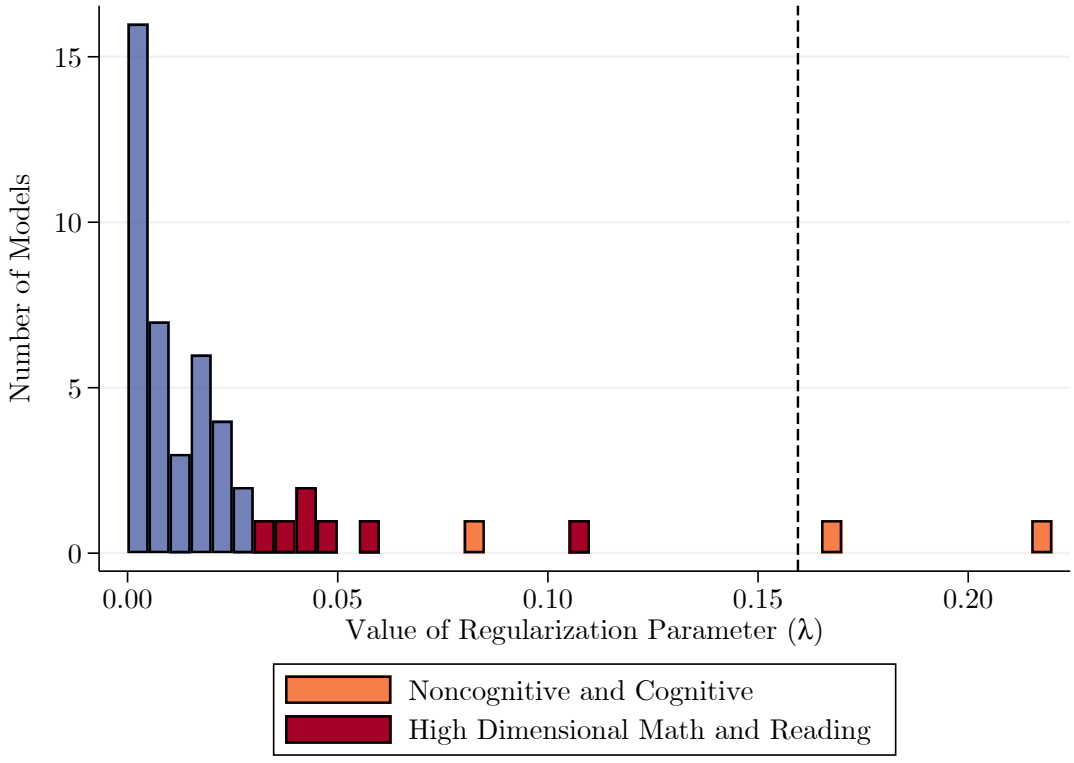
In practice, we choose a unique λ for each model by minimizing the forecast bias $(1-\beta)$ for the estimates (an approximate out-of-sample prediction[55]). We use a standard optimization routine, initializing at $\lambda = 0.005$, calculating the implied estimates, estimating the forecast bias, and updating λ by gradient descent. Since we are minimizing forecast *bias*—one minus the forecast coefficient—there may be local minima with a forecast bias of one in instances where Γ_j is very poorly conditioned (i.e., the model has so little predictive power that small changes in λ do not help). By construction, the only other minimum is at a forecast bias of zero (perfect forecasting), so we add a heuristic adjustment to λ of 0.2 when stuck in a local minimum. Note that while this process seeks to minimize forecast bias, it is not guaranteed to eliminate it. An oversaturated model may still function poorly even after regularization. The fact that this approach produces accurate out-of-sample predictions for models with

[55]Forecast bias is not precisely an out-of-sample prediction because data from all years are used to estimate the hyperparameters—including the year being forecast. However, given the large number of teachers, years, and students, contamination from using the same students or teachers to construct hyperparameters is small.

even 6–12 subgroups is strong evidence of its usefulness in practice.

Figure C.1 shows the optimal λ values for each model. Most models include a very small amount of regularization—55 percent of models have regularization smaller than 10 percent of the on-diagonal variance for that model, and 75 percent are smaller than 15 percent. Models with multiple outcomes typically require more regularization than models using only heterogeneity. For example, the models labeled "Noncognitive and Cognitive" (which shrink across math, reading, behavior, and attendance—with various degrees of heterogeneity) and "High Dimensional Math and Reading" (which shrink across math and reading with three or more degrees of heterogeneity in each) require all of the largest λ values.

Figure C.1. Most Models Require Little Regularization



Note: The dotted line in the figure gives the average value of the variance that we regularize along the diagonal of Γ_j across all models. "Noncognitive and Cognitive" models are those where we shrink not only across reading and math, but also across behavioral GPA and absences, with varying degrees of heterogeneity in math and reading. "High Dimensional Math and Reading" models shrink across $M \geq 5$ math and reading subgroups.

One concern with this approach is the potential need to calculate a new optimal λ for each set of bootstrapped estimates. We do this for a sample of bootstraps, and find that the standard deviation in the optimal λ calculated using the bootstrapped samples is very small,

ranging from 0.0001–0.006 across models. These differences are inconsequential, causing virtually no difference in the estimates (0.99+ correlation across estimates calculated with different λ values). Although the estimates are very similar, we also test whether differences in the value of λ matter in optimal assignments. The standard deviation of the gains from assignments using the estimates generated with these different λ values is 0.088%–0.93% of the predicted gain. We conclude that neither estimation nor reassignment is sensitive to the re-optimization λ for a given model.

Ultimately, most of our models are forecast unbiased with or without regularization. This breaks down for more complex models, where estimating the model without regularization suffers from an ill-conditioned Γ_j, leading to large swings in value-added estimates and high degrees of forecast bias. However, even these models are nearly all forecast unbiased with regularization—specifically, they go from having no predictive power on current-year test scores to showing nearly zero forecast bias once regularized. For the more parsimonious models used for our primary results in the paper, the correlation between the regularized estimates and the typical value-added estimates is 0.998. The correlation in forecast bias between these two sets of models is also very high, 0.95. Even though this regularization matters little for our primary results, we use it with every model for consistency.

C.5 Validation and Robustness

To show the robustness of our teacher value-added estimates, we demonstrate in this section that using an alternative estimation strategy to calculate the impact of teachers on students still yields very similar answers to our preferred methodology. In particular, we estimate teacher effectiveness following the strategy pioneered by Chetty et al. (2014a) by including rich controls, with cubic polynomials in each of our four lagged outcomes interacted with covariates (grade, ethnicity, gender, age, special education status, English learner status), grade and year indicators, cubic polynomials of leave-one-out class and school-grade means of prior-year outcomes interacted with grade, and leave-one-out class and school-year means of all other individual covariates. This strategy also assumes independence of teacher value-added measures across outcomes and student subgroups.

In Table C.3, we show the correlations of these estimates with our preferred estimates for value added on math and reading achievement. In both reading and math, we show the correlations between the two estimation strategies for two different student subgroups: above- and below-median students in prior math and reading achievement. The correlations between estimates for the same student subgroup and outcome are all high, ranging from 0.6 to 0.7. This suggests a strong relationship between our preferred estimates and those obtained using this alternative specification.

Table C.3. Full Correlations Between Our Preferred and Alternative Estimates

	Math Below	Math Above	Reading Below	Reading Above	Math' Below	Math' Above	Reading' Below	Reading' Above
Math Below	-	0.94	0.67	0.67	0.70	0.64	0.53	0.51
Math Above	0.94	-	0.66	0.67	0.68	0.71	0.51	0.54
Reading Below	0.67	0.66	-	0.94	0.54	0.47	0.68	0.58
Reading Above	0.67	0.67	0.94	-	0.53	0.48	0.65	0.61
Math' Below	0.70	0.68	0.54	0.53	-	0.80	0.70	0.60
Math' Above	0.64	0.71	0.47	0.48	0.80	-	0.57	0.68
Reading' Below	0.53	0.51	0.68	0.65	0.70	0.57	-	0.67
Reading' Above	0.51	0.54	0.58	0.61	0.60	0.68	0.67	-
Standard Dev	0.18	0.20	0.13	0.12	0.19	0.22	0.12	0.14

Note: The standard deviations are reported in the last row, and each element of the matrix is the correlation between the row and the column. Every row or column with a prime (') gives the alternative estimates without school fixed effects, whereas those without primes give our preferred estimates. "Below" and "Above" refer to the students below and above the median.

Furthermore, correlations in Table C.3 between above- and below-median math students are substantially higher for our preferred estimates (0.93) than for the alternative specification (0.80). This is also true for reading value added: 0.94 for our preferred estimates versus 0.67 for the alternative specification. The last row of the table shows that the amount of across-teacher variation in value added is similar. Taken together, this implies that the scope for gains from comparative advantage using these alternative estimates is larger than from our preferred estimates, suggesting that our estimates provide a more conservative picture on the gains from reallocating teachers than would be obtained using these alternative estimates.

D. Assignment Policy Details

D.1 Optimal Assignment with Linear Programming

To assess the gains from alternative assignments, we first assign each student in the policy counterfactual sample to a relevant subgroup. For students with observed lagged outcomes and demographics, this is trivial. For other students, we impute their subgroups and then hold them constant throughout all assignment counterfactuals as follows. (1) use the current-year score rather than the lagged score because it is the closest proxy for lagged scores. (2) use the student's modal subgroup for the relevant outcome across years we observe the student. (3) use the subgroup of a randomly drawn peer of the same gender, race, school, year, and grade. (4) use the subgroup of a random peer from students in their school in that same year (or (5) at any time). This procedure classifies all students and to account for them in reassignment. We include all teachers with value added estimates in the reassignment exercise; other teachers are assigned to the class they actually taught.

We formulate the following mixed-integer linear programming problem. We assign teachers and classes to reassignment strata defining allowed assignments. Each strata is effectively its own optimization problem. For within-school swaps, these strata are school-grade-year, and for across-school swaps they are grade-year. We then represent each possible assignment as a binary variable, $x_{j,c}$, indicating whether teacher j was assigned to class c.

Ultimately, this leaves us with $(N_C)^2$ assignment variables in each cell where N_c is the number of classes or teachers in the cell. Let $C : (i,t) \to c$ be an assignment function indicating which class a student is assigned to each year. Formulated in this way, the objective function we maximize in the optimization problem as following:

$$\hat{V}_{\mathscr{C}} = \sum_{c \in \mathscr{C}} \sum_{j} x_{j,c} \left(\sum_{i:C(i,t)=c} \omega_{k_{i,t}} \hat{\mu}_{j,k_{i,t},t} \right)$$

where $\mathscr{C}$ denotes the cell (fixing year, t, grade, and school where applicable), $\omega_{k_{i,t}}$ is the welfare weight assigned to students of type $k = k_{i,t}$ and $\hat{\mu}_{j,k_{i,t},t}$ is the estimated value added of teacher j for students of type $k = k_{i,t}$. Because each $x_{j,c}$ is an assignment indicator, this function sums the welfare-weighted value added from each teacher assignment made, and adds zero from assignments that are not made.

The linear constraints for the problem are as follows:

$$\sum_{c} x_{j,c} = 1, \forall j \qquad \sum_{j} x_{j,c} = 1, \forall c$$

meaning that each teacher must be assigned to one and only one class, and each class must have one and only one teacher. We add one additional constraint when restricting the number of swaps that can be made:

$$\sum_{c} \sum_{j} \mathbb{1}\{j \neq c\} x_{j,c} \leq N_c \times f$$

where f is the fraction of swaps we allow. Since this sums the binary assignment variables where $j \neq c$—that is, those where the teacher is assigned to a different class than the status quo—it represents the number of switches that are made. The solution to either the constrained or unconstrained problem yields an optimal mapping of teachers to classes.

D.2 Aggregating Gains from Reassignment

This section describes how we aggregate gains across students for each assignment. Once we solve for the optimal assignment of teachers to classes in each instance, we use the resulting mapping to assign each student a 'new' value-added score—that is, the heterogeneous estimate from their newly assigned teacher for the student's subgroup. The 'gain' for each student in a given year is the newly assigned value-added score (i.e., what the student would get if the alternative assignment were implemented) minus the original value-added score (i.e., what they would have had in the status quo).[56] This means that if we are unable to reallocate a teacher in a particular year, her students are all given a gain of zero since gains are relative—comparing the assignment with the status quo—and those students keep the same teacher in both.

Our policy counterfactual sample allows us to report the average expected three-year gain for a student who experienced the policy for three years. We do this first by summing the total gains for students within each grade level and subgroup and then dividing by the number of relevant student-years to obtain the average gains to students in each grade and subgroup.[57] We then sum these grade-specific averages into the total expected three-year gain for students of each type.

[56] This, of course, can be negative if the newly assigned teacher is worse for the particular student or can be zero if a student is assigned to the same teacher in the alternative assignment that they would have been assigned to in the status quo.

[57] We divide by student year counts to account for students who leave or enter the data between 3rd and 5th grade—just as a real policy maker would have to.

E. Second-Best Model Selection

This appendix presents details of the model selection criteria proposed in the body of the paper and their results for assignments targeting math scores and lifetime earnings. The first two criteria are tie-breaking rules based on other information for a social planner who wants to maximize expected gains. The second set of criteria are alternatives that directly penalize variance or misallocation risk.

E.1 Hyperparameter Quality

Given the theoretical result that optimism and estimation regret are caused by misestimating hyperparameters in the feasible Empirical Bayes approach, we use model hyperparameters to assess model quality. Recall that the hyperparameters used in shrinkage are the cross-subgroup and cross-time-period covariances of teacher residuals. Furthermore, while the number of hyper-parameters to estimate increases quadratically in $\mathcal{K}_m$, the effective number of teachers identifying each parameter actually shrinks because each cross-subgroup or cross-time-period covariance requires observing teachers who teach both types of students or in both time periods. This pattern suggests that hyperparameter quality is likely to deteriorate as value-added models become more complex. We focus on two intuitive metrics for quality: the monotonicity and stationarity of these covariances.

First, consider monotonicity. The monotonicity measure is grounded on the assumption that the correlation between more proximate subgroups and time periods should be higher than the correlation between more distant subgroups and time periods. For example, monotonicity would imply that the correlation between value added on the first and second decile of lagged student achievement should be higher than the correlation between value added on the third and ninth deciles. To order all subgroups, we also assume that correlation between students of similar achievement quantiles is higher than that between students of similar demographic characteristics and that same subject correlation across different quantiles is higher than cross subject correlation. Similarly, this assumption implies that the correlation between any group this year and another group last year should be larger than that between that other group five years ago.

We measure the empirical deviation from monotonicity and report the share of total covariance that is non-monotonic. Denoting the temporal distance between time periods as, h, we define the difference in covariance between adjacent pairs $\Delta^k_{k,k',h} = \sigma^2_{k,k'+1,h} - \sigma^2_{k,k',h}$[58] and time periods $\Delta^t_{k,k',h} = \sigma^2_{k,k',h+1} - \sigma^2_{k,k',h}$. Under monotonicity these differences should all

[58]Technically this is for $k' > k$ and it is $\Delta_{k,k',h} = \sigma^2_{k,k'-1,h} - \sigma^2_{k,k',h}$ for $k' < k$because monotonicity is only assumed going toward the $k = k$ diagonal from each side.

be negative. We report the share of that total covariance that is non-monotonic:

$$\text{HP}_{\text{Mono}} = \frac{\sum_t \sum_k \sum_{k'} \max(\Delta^k_{k,k',t}, 0) + \max(\Delta^t_{k,k',t}, 0)}{\sum_t \sum_k \sum_{k'} \text{abs}(\Delta^k_{k,k',t}) + \text{abs}(\Delta^t_{k,k',t})}$$

Second, consider stationarity. The stationarity measure is grounded on the assumption that the correlation between equidistant pairs of subgroups should be relatively similar. For example, stationarity would imply that the correlation between value added on the first and second decile of lagged student achievement should be relatively similar to the correlation between the fifth and sixth decile since both pairs are adjacent. This assumption is grounded in the empirical norm established by Chetty et al. (2014a) of assuming and enforcing stationarity across time periods in standard value-added estimation. Whereas our concept of monotonicity seems all but guaranteed for the true variance-covariance matrix, there are many reasonable cases in which modest deviations from stationarity are plausible. In our view, however, significant deviations from stationarity are likely still indicative of poor hyperparameter quality.

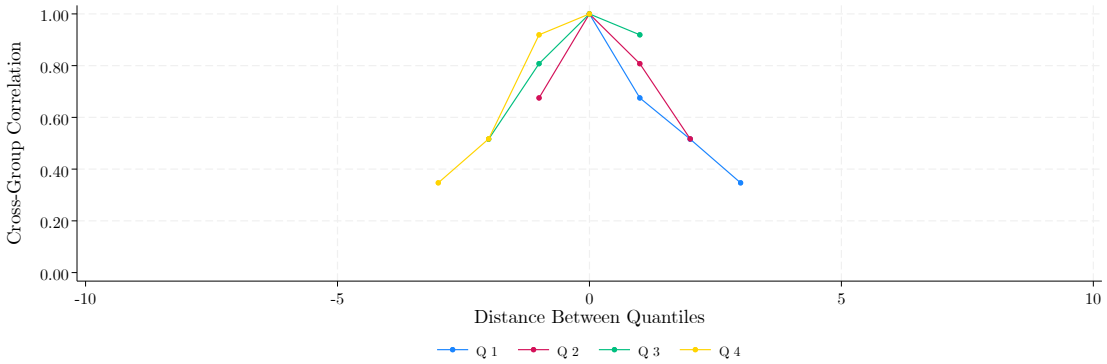
We measure the empirical deviation from stationarity and report the root mean squared error implied by imposing stationarity on the empirical hyperparameter estimates (i.e., the ratio of non-stationary covariance to total covariance). To do this we consider the same (auto)covariance matrix, and measure the variance of equidistant subgroup-pairs around the average covariance: $\sigma^2_{k,k'} - \bar{\sigma}^2_{|k-k'|}$, where the average is $\bar{\sigma}^2_{|k-k'|} = \frac{1}{k-d} \sum_{(k,k'):|k-k'|=d} \sigma^2_{k,k'}$. We report the mean squared error of these deviations relative to the total squared covariance:

$$\text{HP}_{\text{Stat}} = \sqrt{\frac{\sum_t \sum_k \sum_{k'} (\sigma^2_{k,k'} - \bar{\sigma}^2_{|k-k'|})^2}{\sum_t \sum_k \sum_{k'} (\sigma^2_{k,k'})^2}}$$

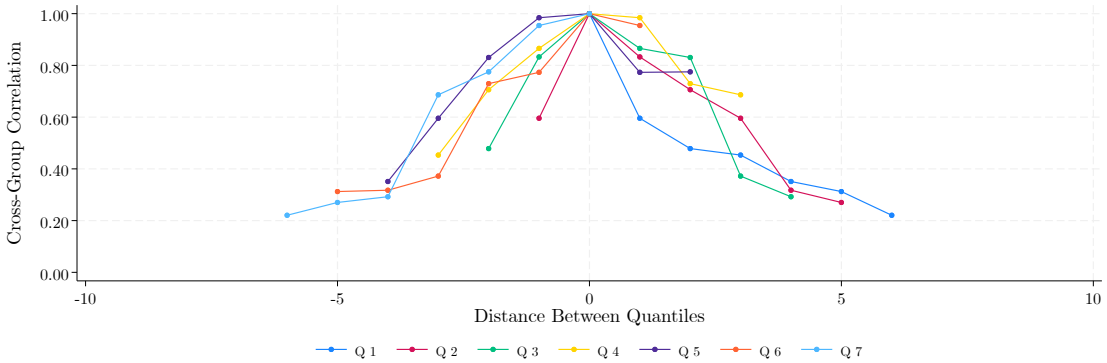
To build intuition, Figure E.1 plots three examples from math heterogeneity by achievement that illustrate how monotonicity and stationarity break down as models become too complex. Panel (a) shows the contemporaneous ($h = 0$) correlations between subgroup residuals over the distance between quantiles across quartiles of lagged math scores.[59] The cross-group correlations estimated in this model are monotonic in the distance between quartiles and appear stationary (e.g., the correlation between 1 and 3 and 2 and 4 are similar), with adjacent quartiles being correlated at around 0.8 and the furthest quartiles correlated at under 0.4. Increasing model complexity complicates these patterns. In Panel (b) of Figure E.1, we observe the cross-group correlations across septiles of prior math achievement.

[59] For example, for the Q_1 line, +1 denotes quartile 2 and +3 denotes quartile 4.

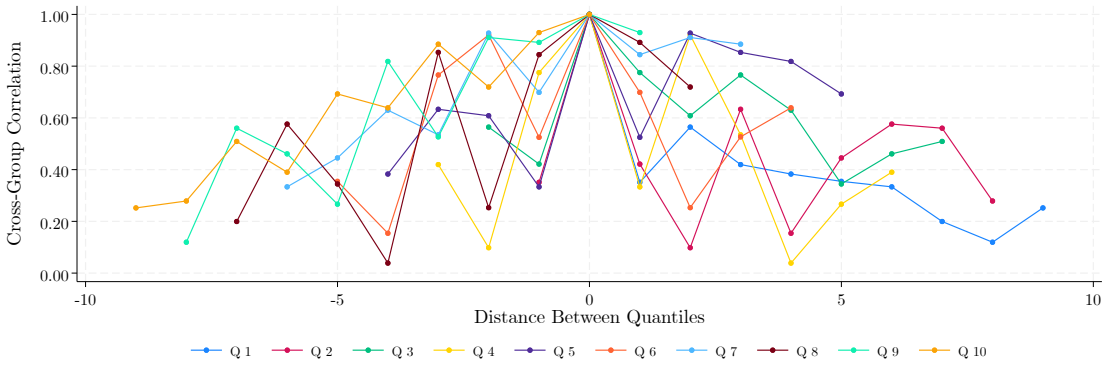
Figure E.1. Hyperparameter Stability Decreases with Model Complexity



(a) Quartiles



(b) Septiles



(c) Deciles

While typically monotonic in the distance between quantiles, these parameters are much less consistent and are also less stationary. These problems explode in the case with 10 subgroups (Panel (c)). The correlation structure varies wildly and is not monotonic (e.g., subgroups 8 and 3 are correlated at 0.35, 8 and 4 at 0.05, 8 and 5 at 0.85, and 8 and 6 at 0.25). It seems unlikely that these patterns reflect the true data-generating process, indicating that models with the highest degrees of heterogeneity are substantially less reliable. Table E.1 reports the full set of results across math-, reading-, and earnings-focused value added.

E.2 Predicted Welfare MSE

Because breaking ties with our metrics of hyperparameter quality could seem somewhat arbitrary, we also consider using the mean squared error of predicted welfare as a direct measure of misallocation risk conditional on model performance. Appendix B.3 showed that MSE jointly penalizes model misspecification (analogous to bias) and estimation error (analogous to variance), so we define

$$\widehat{\text{WMSE}}(m) \equiv \mathbb{E}\left[\widehat{\mathcal{W}}_m^{\hat{\mathcal{J}}_m} - \widehat{\mathcal{W}}_{\text{ens}}^{\hat{\mathcal{J}}_m}\right]^2 + \mathbb{V}\left(\widehat{\mathcal{W}}_m^{\hat{\mathcal{J}}_m}\right)$$

Note that the true welfare MSE would use $\mathcal{W}^{\hat{\mathcal{J}}_m}$ in the bias term, but because this object is unknowable, we approximate it with $\widehat{\mathcal{W}}_{\text{ens}}^{\hat{\mathcal{J}}_m}$, a leave-model-out ensemble measure of the gains at $\hat{\mathcal{J}}_m$ based on all other value-added models.

To calculate $\widehat{\mathcal{W}}_{\text{ens}}^{\hat{\mathcal{J}}_m}$ we revaluate the different optimal allocations proposed by each model, $\{\hat{\mathcal{J}}_m\}$, using the value-added estimates form the 1,000 bootstraps of all of the other models. We then predict the relationship between these different evaluations, $\mathbb{E}\left[\mathcal{W}_{m'}^{\hat{\mathcal{J}}_m}\right]$, and the model complexity in a quadratic regression of the form:

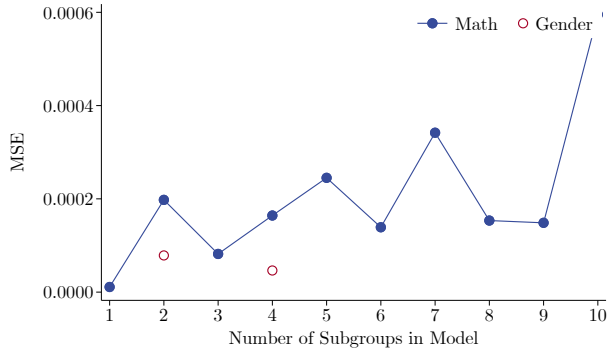
$$\mathbb{E}\left[\widehat{W}_{m'}^{\hat{\mathcal{J}}_m}\right] = \beta_{1,m}\mathcal{K}_{m'} + \beta_{2,m}\mathcal{K}_{m'}^2 + e_{m'}$$

where $\mathbb{E}\left[\widehat{W}_{m'}^{\hat{\mathcal{J}}_m}\right]$ are the expected gains from $\hat{\mathcal{J}}_m$ predicted by model m'. We then define $\widehat{\mathcal{W}}_{\text{ens}}^{\hat{\mathcal{J}}_m} = \hat{\beta}_{1,m}\mathcal{K}_m + \hat{\beta}_{2,m}\mathcal{K}_m^2$ to calculate the MSE of predicted welfare.

Figure E.2 shows that the MSE of predicted welfare is increasing with $\mathcal{K}_m$ but not monotonically. Standard value added and value added by gender have extremely low variability and bias (but they produce the smallest gains). The variability of deciles is enormous. In the consideration set defined in the text models with four, six, eight, and nine quantiles all have comparably low MSE.

Note that the creation of $\widehat{\mathcal{W}}_{\text{ens}}^{\hat{\mathcal{J}}_m}$ has the added benefit of correcting a small bias in expected gains that persists even after the forecast deflation and joint integration. The "self-evaluation

Figure E.2. Predicted Welfare MSE is Roughly Increasing with $\mathcal{K}_m$



Note: Each point in the figure represents the Predicted Welfare MSE using $\widehat{\mathcal{W}}^{\hat{\mathcal{J}}_m}_{\text{ens}} = \hat{\beta}_{1,m}\mathcal{K}_m + \hat{\beta}_{2,m}\mathcal{K}_m^2$ to calculate the MSE of predicted welfare. The dots labeled 'Gender' use math test scores with heterogeneous value added estimated along gender partitions for the two subgroups dot and along gender by two achievement group partitions for the 4 subgroup dot.

bias" this process corrects occurs from the fact that student subgroups are observed with error. Idiosyncratic errors may cause a student whose true academic ability is in the third quartile to be empirically classified in the fourth quartile. Because bootstrapping takes student subgroups as given, it doesn't account for this uncertainty and so the set of models based on the same partition will perpetuate slight over-certainty about the quality of specific types of matches. Thus all of the bootstraps will give a slight premium to assignments made based on the same partition of students. In practice, this self-evaluation bias turns out to be fairly small, but it does increase in $\mathcal{K}_m$, providing additional evidence that the predicted gain from the richest value-added models may not be entirely accurate.

E.3 Quadratic Loss from the First-Best

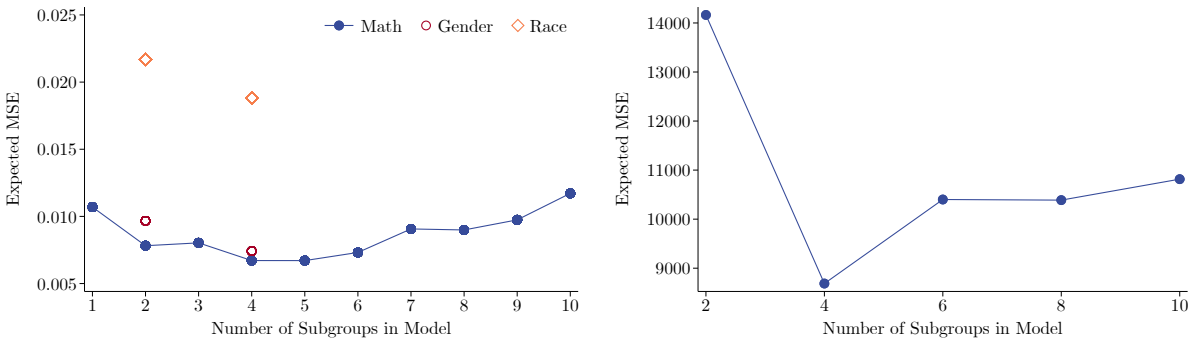
The previous subsection considered the MSE of predicted welfare as a diagnostic for model quality. But while that approach penalized inaccuracy, it did not penalize sub-optimality. Because models with greater matching gains will have less misspecification regret, and because estimation regret is conceptually related to variance, we propose an alternative MSE-like model selection criterion based on expected quadratic loss relative to the first best:

$$\begin{aligned}\mathcal{L}(m) &= \mathbb{E}_b\left[\left(\widehat{\mathcal{W}}^{\hat{\mathcal{J}}_m}_{m_b} - \mathcal{W}^*_{\text{First Best}}\right)^2\right] \\ &= \mathbb{V}\left(\widehat{\mathcal{W}}^{\hat{\mathcal{J}}_m}_{m_b}\right) + \left(\mathbb{E}\left[\widehat{\mathcal{W}}^{\hat{\mathcal{J}}_m}_{m_b}\right] - \mathcal{W}^*_{\text{First Best}}\right)^2\end{aligned}$$

Unfortunately, $\mathcal{L}(m)$ is not directly estimable because the gains from the first-best assignment are unknown. To overcome this limitation, we draw plausible values for $\mathcal{W}^*_{\text{First Best}}$ and calculate the expected loss for each model m:

$$\overline{\mathcal{L}}(m) = \int_{\mathcal{W}'} \mathbb{V}\left(\widehat{\mathcal{W}}^{\hat{\mathcal{J}}_m}_{m_b}\right) + \left(\mathbb{E}\left[\widehat{\mathcal{W}}^{\hat{\mathcal{J}}_m}_{m_b}\right] - \mathcal{W}'\right)^2 \mathrm{d}\mathcal{W}'$$

Figure E.3. Loss across Models with Differing Complexity



(a) Loss across Math Models

(b) Loss across Earnings Models

Note: Each point in the figure represents the average loss across different values of the true parameter. This true parameter is drawn uniformly from $[0.05, 0.25]$ for Panel (a) and from $[1000, 7000]$ for Panel (b). The dots labeled 'Math' in Panel (a) correspond to models including only math scores with differing numbers of subgroups. Those labeled "Gender" and "Race" include demographics or demographics interacted with above- and below-median math achievement in the four-subgroup case. Each earnings model in Panel (b) includes estimates of the specified number of subgroups for both math and reading, used jointly.

In practice, we estimate $\overline{\mathcal{L}}(m)$ using the sample analogues $\hat{\mathbb{E}}[\widehat{\mathcal{W}}^{\hat{\mathcal{J}}_m}_{m}] = \frac{1}{B}\sum_b \widehat{\mathcal{W}}^{\hat{\mathcal{J}}_m}_{m_b}$ and $\hat{\mathbb{V}}(\widehat{\mathcal{W}}^{\hat{\mathcal{J}}_m}) = \frac{1}{B-1}\sum_b \left(\hat{\mathbb{E}}[\widehat{\mathcal{W}}^{\hat{\mathcal{J}}}_{m}] - \widehat{\mathcal{W}}^{\mathcal{J}_m}_{m_b}\right)^2$. We draw $\mathcal{W}'$ uniformly from $[0.05, 0.25]$ for math and $[1000, 7000]$ for earnings (which each cover about one third of a teacher standard deviation of gains). The average loss across models is shown in Figure E.3. In Panel (a), when focusing on math scores, the model allowing for heterogeneity across four subgroups performs best, even when compared with models that also use demographics. In Panel (b), the earnings model that performs best is that allowing for different value added above and below median in both math and reading scores.

Table E.1. Diagnostics for Hyperparameters for Each Model

		Panel A: Math achievement				Panel B: ELA achievement			
		Inversion Share			Stationarity	Inversion Share			Stationarity
Model	$\mathcal{K}_m$	Group	Time	Combined	RMS	Group	Time	Combined	RMS
All students (STD)	1	–	0.035	0.035	0.000	–	0.149	0.149	0.000
Lagged-Achievement:									
Q2	2	0.044	0.045	0.045	0.086	0.131	0.124	0.127	0.175
Q3	3	0.038	0.067	0.053	0.055	0.076	0.138	0.103	0.200
Q4	4	0.041	0.070	0.053	0.084	0.065	0.143	0.099	0.166
Q5	5	0.073	0.090	0.080	0.112	0.115	0.157	0.133	0.193
Q6	6	0.244	0.106	0.203	0.168	0.222	0.181	0.207	0.232
Q7	7	0.200	0.126	0.176	0.174	0.113	0.182	0.142	0.213
Q8	8	0.156	0.114	0.140	0.158	0.291	0.223	0.269	0.261
Q9	9	0.267	0.126	0.221	0.181	0.239	0.211	0.228	0.262
Q10	10	0.245	0.134	0.208	0.201	0.345	0.227	0.310	0.304
Demographics:									
Gender	2	0.111	0.049	0.064	0.027	0.123	0.165	0.155	0.045
Race (White/Asian)	2	0.149	0.036	0.076	0.056	0.146	0.166	0.162	0.069
Gender × Q2	4	0.359	0.077	0.246	0.113	0.268	0.151	0.212	0.159
Race × Q2	4	0.080	0.075	0.078	0.076	0.193	0.146	0.169	0.154

		Panel C: Joint multi-subject models			
		Inversion Share			Stationarity
Model	$\mathcal{K}_m$	Group	Time	Combined	RMS
Math + Reading:					
$Q_1^M \times Q_1^R$	2	0.000	0.115	0.061	0.095
$Q_2^M \times Q_2^R$	4	0.205	0.098	0.166	0.149
$Q_3^M \times Q_3^R$	6	0.262	0.115	0.212	0.191
$Q_4^M \times Q_4^R$	8	0.248	0.127	0.206	0.180
$Q_5^M \times Q_5^R$	10	0.274	0.152	0.234	0.216
Demographics + Math + Reading:					
$G \times Q_1^M \times Q_1^R$	4	0.130	0.132	0.131	0.107
$G \times Q_2^M \times Q_2^R$	8	0.338	0.138	0.265	0.161
$R \times Q_1^M \times Q_1^R$	4	0.198	0.119	0.162	0.105
$R \times Q_2^M \times Q_2^R$	8	0.262	0.121	0.205	0.144

Note: Each row is one value-added model. $\mathcal{K}_m$ indicates the number of subgroups times outcomes in the given model. The group inversion share is the (magnitude-weighted) share of group covariance estimates that move the wrong way (where a more distant group is more highly correlated than a closer group). For models with demographics and achievement, we assume that students in the same achievement group are more closely related than those in the same demographic group. For models with multiple outcomes, we assume that different groups within the same outcome are more correlated than those across outcomes. The time inversion share is the (magnitude-weighted) fraction of covariance estimates where a further time period is more highly correlated than the closer time period. Combined inversion share is the average of the two. The Stationarity RMS is root-mean-squared fraction of the differences between correlations between two time periods and their sample average, normalized by the overall size of the correlation, meaning that 0 indicates full stationarity in the data.