Welfare Added?
Optimal Teacher Assignment with Value-Added Measures

Tanner S. Eastmond[*] Michael David Ricks[†]

Julian Betts[‡] Nathan Mather[§]

January 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 increasing misallocation risk. 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 $2,800 and increase lower-achieving students' earnings by 70–156% 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, 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, and the NBER Education 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. 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 among 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 begin by proposing 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., Petek and Pope, 2023), "better" assignments are conceptually ill-defined (Condie et al., 2014). Aggregating gains using welfare weights and an index for lifetime utility characterizes the (full-information) first-best assignment. Second, because value-added measures are estimated with noise, assignments based on estimated effects risk misallocation (as in Andrews et al., 2024). Thus, the cost of using models with richer heterogeneity to improve targeting is compounding misallocation risk and overstating gains. We allow the social planner to address this tradeoff directly by choosing the model of teacher effects used for assignment. By incorporating match effects, multiple outcomes, model choice, and distributional objectives, this framework characterizes a (feasible) "second-best" solution to the teacher-assignment problem.

This framework contextualizes research exploring policies that use match effects and increase test scores. For example, several recent papers study the teacher labor-market implications of teacher or administrator preferences using counterfactual school assignments based on subgroup match effects (e.g., Bobba et al., 2024; Bates et al., 2025; Laverde et al., 2025), and a few papers in applied econometrics use counterfactual class assignments to benchmark the usefulness of new methods for estimating match effects (e.g., Graham et al., 2023; Delgado, 2025; Ahn et al., 2025). These counterfactuals touch on an important public-service provision problem that we attack directly through the lens of welfare economics. We do so by considering multiple outcomes, model choice, and distributional objectives in addition to match effects.[1] Whether the resulting second-best frontier is meaningfully different from other proposals is ultimately an empirical question.

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 between 1998–2019 on math, reading, at-

[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.

tendance, 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. Our estimates are also forecast-unbiased, stable across class composition, and pass a battery of robustness checks.

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. Standard shrinkage methods control the distribution of teacher effects but 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, and (2) 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 districtwide would increase average scores by 0.061 standard deviations over third through fifth grade, or +$1,600 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 +$2,800 per student. These are 1.3× and 2.2× the gains from a benchmark 5% teacher "de-selection" 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 1.2–4.6× larger gains than math-only assignment policies (depending on the policy and the returns to non-cognitive outcomes). 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 +$2,800 versus +$1,700 for lower-achieving students. The second-best policy gives lower-achieving students an additional +$1,200 (69%), 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 +$5,000 to lower-achieving students without harming higher-achieving students on average.

Third, the second-best policy reveals that the optimal amount of student heterogeneity to estimate is modest and loads on lagged outcomes. The marginal gains from additional heterogeneity quickly diminish, while marginal misallocation risks rise. In our setting, partitions based on lagged outcomes tend to produce larger gains and less misallocation regret 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. The optimal second-best policies use above- and below-median partitions of lagged math and reading scores to maximize earnings (or 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 34% (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 (MVFP; Hendren and Sprung-Keyser, 2020) of 2.0, and payments up to $1,300 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 $625 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., 2024; Bates et al., 2025; Laverde et al., 2025), and school choice (Beuermann et al., 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). We find that our optimal assignments

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

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 misallocation risk is a quantitatively important feature of 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), misallocation risk has received less attention. In this sense, our paper is an empirical complement 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

[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.

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) highlight the core trade-offs characterizing the first- and second-best policies; (2) describe the limitations of traditional value-added measures; and (3) discuss when estimating heterogeneous, multidimensional measures of teacher value added can improve assignment policy. Although the exposition focuses on teacher value added, the theoretical insights could apply to any setting in which a social planner chooses between policies that have heterogeneous impacts.

2.1 Welfare Framework

Consider a policymaker selecting a policy $\mathcal{J}$ from a set of potential policies $\mathscr{J}$. In our application, each policy, $\mathcal{J} : i \rightarrow j$, is an assignment that maps all students to their teachers. To focus on the teacher-assignment problem, we limit our attention to policies that hold classes constant,[8] given evidence of the important roles of tracking and peer effects 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 $\mu^y_{i,j}$. Note that 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. 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.

We use welfare theory to order the welfare gains from different assignments. By assigning each student an *ex ante* welfare weight, ω_i, the social planner can use equity considerations to integrate over any effect heterogeneity. Furthermore, by defining a "score function," $S^{\mathcal{J}}_i = s(\boldsymbol{Y}^{\mathcal{J}}_i, \boldsymbol{X}_i)$, that maps 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

[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.

scalar, she can do the same for multidimensionality. Combining these approaches, the welfare of any assignment relative to the status quo can be written as follows:[10]

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

where $\mu_{i,j}$ is teacher j's effect on student i's score relative to the status quo $(S_i^{\mathcal{J}} - S_i^{\mathcal{J}_0})$ and ω_i is student i's welfare weight. While 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.

Note four restrictions implied by Equation 1. First, welfare gains are linear in score gains, consistent with relatively modest ("local") changes in each student's scores.[11] Second, student gains fully capture family preferences for assignments—one way to think of this restriction is as an incentive-compatibility constraint (families will not re-sort to new schools or classes after teachers are reassigned).[12] Third, the welfare function does not directly consider the costs of reassignment policies for teachers. This restriction can also be framed in terms of incentive compatibility: teachers must be compensated sufficiently to switch classes willingly. In Section 5, we consider the welfare effects of various policies that could ensure teacher incentive compatibility. Finally, our welfare formulation implicitly assumes away other considerations, such as school district sorting, union concerns, and the administrative costs of implementing teacher reassignments in practice. Although these considerations clearly matter in the real world, if the gains from optimal assignments are large enough, they could support interventions that alleviate these concerns.

While this formulation of welfare may preclude some cases of interest, we impose these restrictions to focus our attention on the core public-service provision problem: how to optimally assign teachers to classes. As such, readers who are critical of these assumptions could instead consider all welfare gains in partial-equilibrium terms. In any case, the strength of these restrictions depends on the nature of the optimal assignments and how different they

[10] Appendix B.1 derives Equation 1 from a social welfare function based on *ex ante* weighted lifetime utilities, $W = \sum_{i=1}^{n} \phi_i U_i^{\mathcal{J}}$ under the assumptions that $S^{\mathcal{J}}$ is an unbiased linear predictor of utility and that there are no cross-student spillovers. In this light, the score function could be thought of as a surrogate index for expected lifetime utility or earnings (see Athey et al., 2025).

[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 "no spillovers" 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).

are in practice from the status quo.

2.2 The First-Best Teacher Assignment Policy

The first-best assignment of teachers to classes is the one that produces the highest (welfare-weighted) scores. To build intuition about this assignment policy, consider two examples:

- First, let each teacher j's effect be homogeneous: $\mu_{i,j} = \bar{\mu}_j$. After ranking teachers based on 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 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): $\mu_{i,j} = \bar{\mu}_{k,j}$ 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 policies proposed in Delgado (2025) and Ahn et al. (2025) for the two-subgroup case.

In general, however, 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 putting more specialized (higher comparative-advantage) teachers in well-matched classes.

Unfortunately, this full-information first-best policy is only feasible if the social planner knows every possible match effect, $\mu_{i,j}$. In practice, the social planner must rely on the econometrician to estimate match effects. The following subsections outline the welfare implications of this practicality and expand the social planner's problem to endogenize model choice.

2.3 Assignments Using Standard Teacher Value-Added Estimates

We first consider how using standard (homogeneous) value added estimates to make assignments affects the social planner's problem. While this constant-effects model of teacher effectiveness is over-simplified, other empirical analyses often approximate welfare with average treatment effects and average welfare weights (see the argument in Hendren and Sprung-Keyser, 2020). A standard value added estimate, $\hat{\mu}_j$, is typically constructed as the (leave-year-out, jackknife-predicted) mean residual of test-score gains for students taught by teacher

j. Using these estimates, a policymaker could approximate expected welfare from an assignment as follows:

$$\widehat{\mathcal{W}}_{VA}^{\mathcal{J}} = \sum_j n_j \bar{\omega}_j \hat{\mu}_j$$

where n_j is the number of students in teacher j's assigned class under policy $\mathcal{J}$, and $\bar{\omega}_j$ is the average welfare weight of the students in that class.

Equation 2 shows how this approximation of welfare is systematically biased.

$$\mathcal{W}^{\mathcal{J}} - \widehat{\mathcal{W}}_{VA}^{\mathcal{J}} = \sum_j n_j \left[\underbrace{\bar{\omega}_j \left(\tilde{\mu}_j^{\mathcal{J}} - \bar{\mu}_j \right)}_{\text{Matching Gains}} + \underbrace{\widehat{\text{Cov}}_j(\omega_i, \mu_{i,j})}_{\text{Distributional Gains}} + \underbrace{\bar{\omega}_j(\bar{\mu}_j - \tilde{\mu}_j^0)}_{\text{External Validity}} + \underbrace{\bar{\omega}_j(\tilde{\mu}_j^0 - \hat{\mu}_j)}_{\text{Estimation Error}} \right] \quad (2)$$

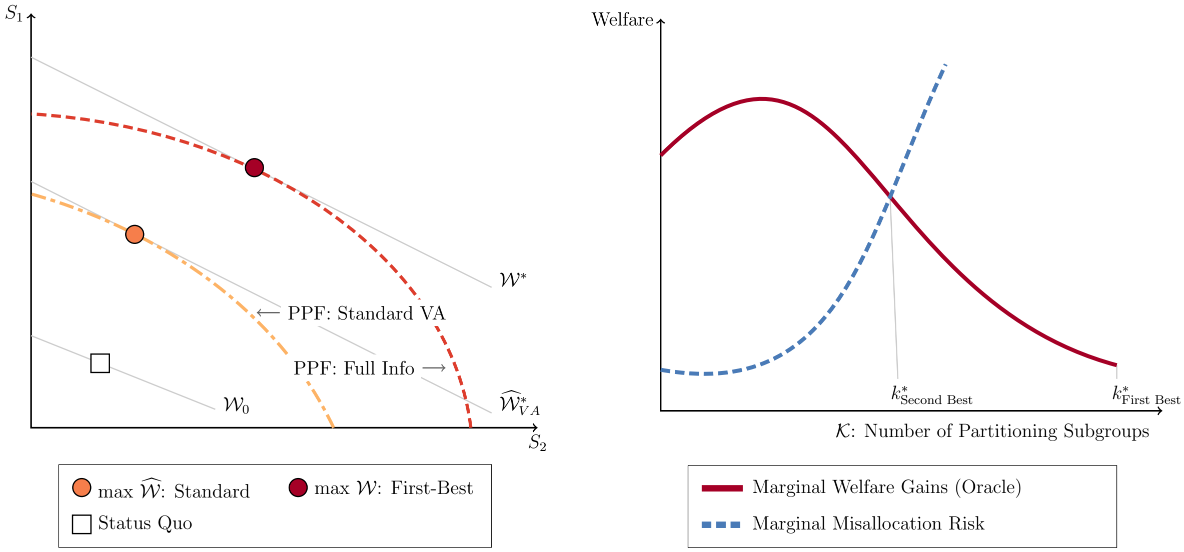
where $\tilde{\mu}_j^{\mathcal{J}}$ represents the average match effects of teacher j on students in their assigned class, and $\bar{\mu}_j$ is the teacher's average match effect in the population (or absolute advantage). $\widehat{\text{Cov}}_j(\cdot)$ reports the estimated within-class covariance of welfare weights and match effects in teacher j's assigned class. Finally, $\tilde{\mu}_j^0$ is the average match effect in the class teacher j taught in the estimation sample (see the full derivation in Appendix B.2).

There are four key differences between the true welfare resulting from an assignment and an approximation based on traditional value-added estimates. First, focusing on the average treatment effect of each teacher ignores matching gains from targeting assignments to classes with large effects (comparative advantage). Although this simplification is less costly if teacher effects are highly correlated across students, acknowledging comparative advantage is typically key to allocative efficiency.

Second, using averages ignores the distributional gains from targeting good matches to students with large welfare weights. These gains are characterized by the covariance between match effects and student welfare weights within each teacher's assigned class. This covariance will be ignorable if teachers truly have homogeneous effects, if the social planner uses uniform welfare weights, or if welfare weights happen to be uncorrelated with match effects.

Third, there is a structural difference between a teacher's average effect, $\bar{\mu}_j$, and the target parameter of value added estimates, $\tilde{\mu}_j$. This inconsistency arises because traditional value-added measures target a "treatment-on-the-treated" parameter (teacher j's effect on the students they actually taught, $\tilde{\mu}_j^0$) rather than an "average treatment effect" (their effect on average, $\bar{\mu}_j$). As such, effects estimated in one class will not be externally valid for other

Figure 1. Optimal Assignments Trade Off Match Effects and Misallocation Risk



(a) Example First-Best

(b) Feasible Second-Best

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 $\widehat{\mathcal{W}}^*_{VA}$. 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 subgroups-specific value-added estimates used to make assignments.

classes whenever the distribution of match effects varies from class to class.[13]

Finally, whatever the target parameter of a value-added estimate, there will be estimation error in finite samples. This introduces an important issue when comparing assignments in search for the optimal policy. While estimation error will bias our evaluation of any assignment rule, choosing an optimal assignment using the estimates $\hat{\mu}_j$ rather than true average match effects will compound this error, inflating anticipated gains at the optimum. This "winner's-curse" type inflation occurs because some of the presumed absolute advantage is just noise (see the analogous discussion in Andrews et al., 2024). This problem will typically become more serious the lower the number of students per class, the higher the variation in match effects within each class, and—if shrinking estimates—the fewer years

[13] 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.

each teacher is observed.[14]

The differences between realized and predicted welfare illustrate why assignments based on traditional value added do not attain the first best. If teacher effects vary, the assignment with the highest approximated welfare will likely produce gains relative to the status quo, but it will not maximize welfare in general. To illustrate this limitation, imagine a model where all heterogeneity in effects and weights occurs across two types of students. Panel (a) of Figure 1 plots a production possibility frontier of gains to students of each type (relative to the status quo). Using estimated standard value added, expected welfare increases from $\mathcal{W}_0$ to $\widehat{\mathcal{W}}^*_{VA}$, but much larger gains could be attained under the full-information first-best $\mathcal{W}^*$. Given the potential value of information about heterogeneous teacher impacts, the following subsection explores when considering heterogeneity is optimal in practice.

2.4 Assignments Using Heterogeneous Estimates

Expanding on constant-effects models of teacher effectiveness, emerging research shows that teachers do in fact have heterogeneous effects. For example, teacher value added varies across student subgroups by race (Delgado, 2025), poverty (Bates et al., 2025), achievement (Biasi et al., 2021), or all three (Ahn et al., 2025), as do other measures of teacher effectiveness (e.g., Aucejo et al., 2022; Graham et al., 2023). In this light, consider estimating subgroup-specific value added, $\hat{\mu}_{k,j}$, pertaining to type-k students (given some partition $\mathcal{K}$) to approximate welfare:

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

where $n_{k,j}$ represents the number of type-k students in teacher j's class and $\bar{\omega}_{k,j}$ their average welfare weight.[15] We consider the usefulness of this approximation in light of the four welfare components of Equation 2.

First, using richer subgroup-specific value-added estimates will improve targeting by incorporating more match effects and making assignments based on comparative advantage. In the limit, the bias in the first term goes to zero as subgroups capture additional welfare-relevant information, but these gains cannot attain the first-best until they fully characterize match effects.

Second, using subgroup-specific effects will tend to improve distributional gains related

[14] By pulling estimates toward one another, shrinkage tempers the winner's curse relative to using unbiased estimates (e.g., see Bobba et al., 2024), but may reduce the scope for matching gains.

[15] The presence of counts, 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., 2024) or absolute advantage (e.g., Ahn et al., 2025).

to the covariance between match effects and welfare weights. In the limit, subgroups devolve into cells with zero covariance. In practice, however, the covariance may increase or decrease as additional dimensions are added. Because reducing within-subgroup variability in match effects or weights reduces bias, some partitions are more useful than others. For example, effects and weights are much more likely to vary between students with more versus less academic preparation than between students with earlier versus later birthdays in a given month. Using high-variance partitions will reduce the bias in this term.

Third, using richer value-added estimates also improves external validity. The potential for gains comes from using subgroup-specific effects to make more accurate comparisons between classes with different compositions. 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, reducing bias. Although finer partitions are more likely to create empty cells from which no information can be extracted, Empirical Bayes shrinkage can predict effects from the available information.[16]

Finally, in contrast to the targeting gains on other fronts, using subgroup-specific effects will typically increase the bias term from estimation error—even when using shrinkage. Although estimation error will tend to increase with more subgroup-specific heterogeneity (as discussed in Ahn et al., 2025, in relation to overfitting match effects),[17] the most threatening mechanism for bias is not estimation error in evaluating any one assignment, but overfitting that estimation error when selecting the optimal assignment. Because the teacher-assignment problem is highly nonlinear, the risk of overfitting small amounts of noise in the estimates can compound with additional imprecisely estimated parameters. As such, the *ex post* regret ($\mathcal{W}^{\mathcal{J}} - \widehat{\mathcal{W}}^{\mathcal{J}}$) due to estimation error may be much larger in assignments made using more complicated models of match effects. Thus, model selection based on predictive fit (e.g., AIC, R^2, log likelihood) does not guarantee socially optimal assignments because fit and misallocation regret are distinct issues (Mbakop and Tabord-Meehan, 2021).

Because estimating additional heterogeneity has both costs and benefits, the optimal extent of heterogeneity to model is an empirical question. Panel (b) of Figure 1 visualizes this trade-off. It presents the marginal welfare effects from making assignments based on teacher value added using increasingly fine partitions of students. The benefits are the gains from

[16] Missing cells could generate welfare losses if policymakers are uncomfortable reassigning teachers with extrapolated effects.

[17] This is obvious for the unbiased estimates whose variability typically decreases in $n_{k,j}$, but it is also the case for shrunk estimates that use the population correlations of effects across groups and time periods as hyper-parameters. The number of these correlations to estimate with the same data increases exponentially with subgroup heterogeneity, potentially resulting in poor estimates in finer partitions.

making assignments using the (true) subgroup effects. The costs are the misallocation risk from overfitting increasingly rich empirical estimates. Obviously, the first-best assignment would use the information from the limiting partition if the match effects were known. In practice, however, the social planner should continue estimating more subgroup-specific effects only so long as the marginal benefits from better matching outweigh the marginal costs from increased misallocation risk. Using value-added estimates based on this last partition is the way to achieve the feasible second-best policy.[18]

This step of considering the optimal amount of heterogeneity to model is a key part of the social planner's problem. In a district with tens of thousands of students, there are an overwhelming number of possible partitions over which to estimate heterogeneous value added. 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). An important exception is Ahn et al. (2025), where model choice is carefully considered in the estimation step before exploring counterfactual policies. Because model overfitting and policy overfitting are conceptually distinct phenomena, endogenizing model choice as part of the assignment problem is essential for attaining the second-best assignment. In practice, the empirical distribution of teacher effects will shape both optimal teacher assignments and the magnitude of gains from optimal policy.

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 (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[19]—and two non-cognitive outcomes—standardized attendance rates and behavior GPAs computed from

[18] 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) all with heterogeneity.

[19] 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 intent of equalizing the teaching burden 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.[20] 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 measures, this section outlines our general empirical approach. To build intuition, the maintained example uses subgroups split by above- or below-median lagged outcomes. In this approach, we let $k \times s$ denote the number of subgroup-by-outcome effects a model must estimate. For example, standard value added on one outcome would have $k = s = 1$; value added between higher- and lower-achieving students would have $k = 2$ and $s = 1$' and value added between higher- and lower-achieving students on both math and reading scores would have $k = s = 2$.

Our estimation approach is adapted from Delgado (2025) and Bates et al. (2025)—see Appendix C.3 for details.[21] For each subgroup k, we separately model student outcome s in year t as

$$S_{i,s,t} = \alpha_{\mathcal{J}(i,t),s,k,t} + \beta_{s,k} X_{i,t} + v_{i,s,t} \tag{3}$$

where $\alpha_{\mathcal{J}(i,t),s,k,t}$ 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 demographics and grade-specific cubics of all four lagged outcomes or missing flags.[22]

[20] 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.

[21] We choose to follow Delgado (2025) and Bates et al. (2025) rather than 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).

[22] 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 non-cognitive outcomes

Appendix C.5 shows that results are not sensitive to alternative specifications, such as those in Chetty et al. (2014a).

After estimating Equation 3, we compute average residuals by teacher, outcome, subgroup, and year in three steps. First, we subtract the estimated effects of observed student characteristics, $\hat{\beta}_{s,k}X_{i,t}$ from the outcome $S_{i,s,t}$. Second, we project these intermediate residuals onto teacher fixed effects $\alpha_{\mathcal{J}(i,t),s}$, a teacher-experience profile $f_s(z_{\mathcal{J}(i,t),t})$, school fixed effects $\phi_{\ell(i,t),s}$, 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:

$$\bar{A}_{j,s,k,t} = \frac{1}{n_{j,k,t}} \sum_{i:\mathcal{J}(i,t)=j,k_i=k} \Big[S_{i,s,t} - \hat{\beta}_{s,k}X_{i,t} - \hat{\phi}_{\ell(i,t),s} - \hat{f}_s(z_{\mathcal{J}(i,t),t}) \Big]$$

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, $\boldsymbol{A}_j^{-t}$, we estimate shrinkage-adjusted value added and add back experience.

$$\hat{\mu}^{VA}_{j,s,k,t} = \hat{\boldsymbol{\psi}}_{j,s,k}\boldsymbol{A}_j^{-t} + \hat{f}_s(z_{j,t}) \tag{4}$$

where $\hat{\boldsymbol{\psi}}_{j,s,k}$ are reliability weights. Equation 4 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.[23] We then implement this procedure in 1,000 bootstrapped samples stratified by class to non-parametrically characterize the joint distribution of value added in each model.

This approach follows Bates et al. (2025) exactly with two deviations to deal with multidimensionality. First, we shrink value-added estimates across both the subgroups 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. 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

(thus preventing grading stringency from entering value added).

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

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 we flexibly control for grade-specific cubics in all four lagged outcomes and include 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]

Additional analyses support the plausibility of conditional independence and the validity of our estimates. First, for each outcome and student type, we document precise forecast unbiasedness. Following Chetty et al. (2014a), we regress student 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 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).

Because optimal policy considerations require evaluating counterfactual assignment policies, 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 depicts within-teacher changes in value added based on class size and composition. On the one hand, if consistently teaching larger classes or lower-achieving classes 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, Panels (a) and (b) show no evidence of this. On the other hand, if consistently teaching 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 comparative advantage and composition; however, Panels (c) and (d) show extremely precise null relationships for composition and small slopes

[24] Regularization does not impact welfare because second-best optima use relatively sparse models of heterogeneity and regularization only changes the estimates meaningfully in relatively higher-dimensional models that have larger 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.

for class size. Together, these relationships suggest that our estimates reflect fundamental teacher characteristics with strong external validity in counterfactual assignments.[26] If anything, our estimates likely underestimate the potential gains from reassignment if teachers can adapt or specialize.

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 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.

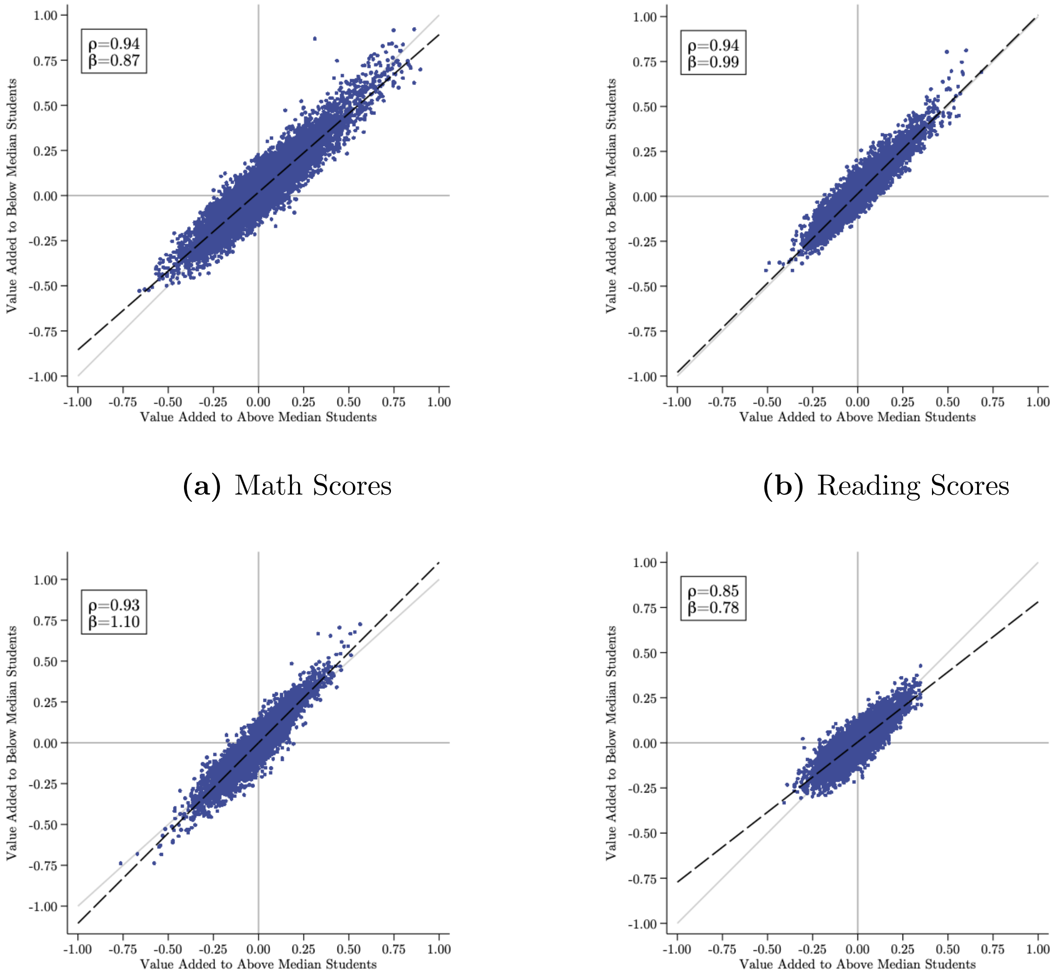
Figure 2 depicts meaningful differences in value added within *and* across teachers. Absolute advantage can be seen in the dispersion of teachers along the 45-degree line. Teachers above and to the right generate larger gains than to teachers below and to the left. Note that value added to math is much 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.[27] 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.[28] Appendix Table A.3 shows that we reject homogeneity across 1,000

[26] Because value-added estimates are predictions based on prior years, this exercise is most informative if a teacher's class size and composition are more highly correlated in recent years. To the extent that these changes are idiosyncratic across years, using prior years to predict changes in current year value added will not capture adaptation. Because Aucejo et al. (2022) find that teachers adapt teaching practices little across classrooms with different levels of prior achievement, this is unlikely to be a concern in practice.

[27] 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).

[28] This is about three times larger than the difference attributable to matches on observable student

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 both 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.

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 mean-

and teacher characteristics (Laverde et al., 2025) 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.

ingful 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.

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 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):

$$\max_{\mathcal{J}\in\mathscr{J}} \tilde{\mathcal{W}}(\mathcal{J};\boldsymbol{\omega},\boldsymbol{\mu}) = \max_{\mathcal{J}\in\mathscr{J}} \frac{1}{N} \sum_{t=2003}^{2013} \sum_{k=1}^{K} \sum_{(i,t):\, k_{i,t}=k} \omega_k \, \hat{\mu}_{\mathcal{J}(i,t),k} \tag{5}$$

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 5'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 5 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 5 as a (mixed-integer) linear programming problem (Bertsimas and Tsitsiklis, 1997)—see Appendix D.1 for details.

Four complications affect how we solve and analyze solutions to Equation 5 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.4 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.[29]

[29] Because the social objective is monotonically increasing in each teacher's value added, we choose left-tail

While we prefer the robust estimates, Appendix Table A.4 also reports results from standard optimization.

Third, predicting gains for each solution using noisy estimates as the true data-generating process will tend to be overly optimistic. Using shrunk estimates reduces this concern but does not eliminate the "winner's curse" risk given the highly nonlinear nature of the problem. We address this issue by reporting the expected gains from each assignment, $\mathcal{J}$, integrated over the joint distribution of all teacher effects: $\hat{\mathcal{W}}_{\mathcal{J}} = \mathbb{E}_{\boldsymbol{\mu}}\left[\mathcal{W}(\mathcal{J}, \omega, \mu)\right]$.[30] In principle, naively predicted gains may or may not be overoptimistic and could align with these expected gains; in practice, we find large differences. Appendix Figure A.4 depicts large differences and shows how naive estimates even create an illusion of convex returns to model complexity. Also note that this process is extremely 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 reevaluating each optimum with each set of bootstrapped estimates.

Fourth, the final challenge is comparing assignments made with estimates from different models since Equation 5 considers the optimal gains given each set of estimates. The theory in Section 2 states that the second-best optimum is an assignment made using estimates from a model that balances expected marginal matching gains against marginal increases in misallocation risk. We operationalize this trade-off with four data-driven approximations of misallocation risk. Our preferred criterion is each model's expected mean squared error relative to plausible gains from the first-best equilibrium, but we also consider *ex post* regret and heuristics based on variability and hyper-parameter behavior (see Appendix D.3 for details on all four approaches). The regret criterion for model-selection involves re-solving the assignment problem for each bootstrap, introducing another layer of computational complexity. In the end, given the empirical distribution of gains, all four approaches to model selection give similar characterizations of the second best.

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.

[30] In practice, we estimate this expectation using the bootstrapped estimates $\hat{\mathbb{E}}_{\boldsymbol{\mu}}\left[\mathcal{W}(\mathcal{J}, \omega, \mu)\right] \equiv \frac{1}{B}\sum_b \mathcal{W}(\mathcal{J}, \omega, \hat{\mu}^b)$. We prefer the bootstraps to integration 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.

[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.

4.2 Gains from (Second-Best) Optimal Assignment Policies

4.2.1 Maximizing Math Scores

We first consider the second-best problem for a social planner trying to maximize average math scores. In Equation 5, this objective implies equal (utilitarian) welfare weights on all students. To find the second-best assignment, the econometrician estimates multiple value-added models and the social planner compares expected gains from making assignments using each model to determine the optimal policy.

We consider value-added estimates including 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 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 Appendices C.2 and 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.

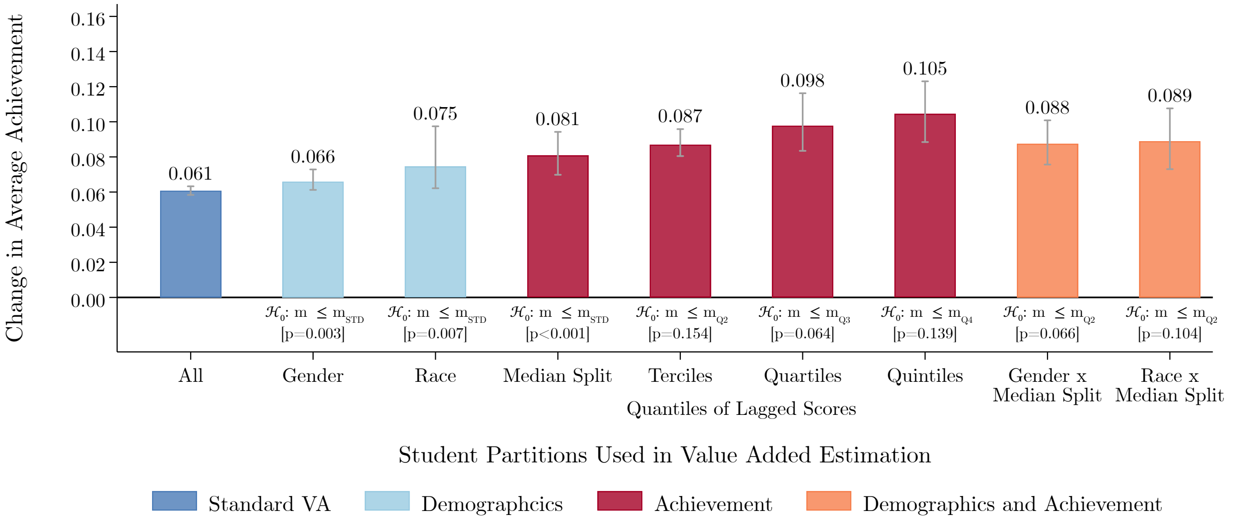
Figure 3 presents the expected gains from robust reassignments based on value-added estimates from nine models with varying complexity.[32] Bars depict the expected achievement gains from an optimal assignment relative to the status quo. We report the gains as the average cumulative effect over grades 3–5 (see Appendix D.2 for details). Each 95% confidence interval comes from 1,000 reevaluations of the robust assignments using our 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 the next most simple model,[33] where p-values report the share of bootstraps in which the simpler model produces larger expected gains than the richer model. All results are for the district-wide reassignment, but Appendix Table A.4 contains analogous within-school results. Furthermore, Appendix Table A.5 shows analogous results for reassignments maximizing reading

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

[33] This corresponds to models with one fewer quantile for achievement-based partitions and models without race and gender for demographic-based partitions.

scores and lifetime earnings.

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 5 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.

Three 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, 50% larger than a benchmark teacher "deselection" program.[34] 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 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. As predicted in Section 2, there also seem to be diminishing returns to considering additional

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

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 about 10–20%, but value added based on quantiles of lagged achievement increases gains by 32–72%. Not only are the gains from considering lagged achievement larger, but conditional on achievement, the marginal impact of considering additional demographics is even smaller. For example, using the interaction of gender or race and a median prior-achievement split generates about 9% improvements over just the above- or below-median split, compared with 21% improvements from using four achievement quartiles. 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.

Third, and finally, Figure 3 allows us to begin to compare models to choose the second best. The figure depicts expected gains from reassignments, but not misallocation risk. When we consider this trade-off directly, we find three reasons to select the assignment based on value added by achievement quartiles as the optimal second-best policy. First, assignment based on achievement quartiles has a lower expected mean squared error than all other models. While quartiles and quintiles are similar, Appendix Figure D.1 shows that other models have 1.1–2.9× higher expected MSE. Second, Appendix D.3 also shows that achievement quartiles effectively balance expected gains versus *ex post* regret and produce much more believable hyper-parameter estimates than other more complicated models. Finally, from a heuristic standpoint, using four quantiles seems reasonable given the statistical equivalence of assignments based on all larger models and the fact that these models have confidence intervals that are 14% larger on average (a heuristic measure of increasing misallocation risk).

4.2.2 Restructuring Achievement Equity

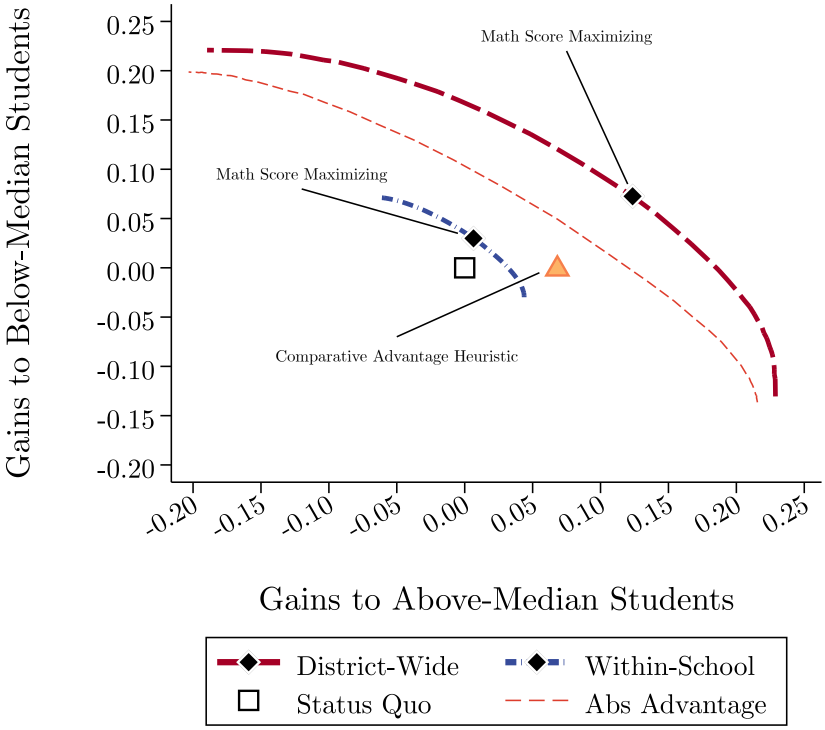
Whereas the previous results all focused on maximizing average scores, revealed preference suggests that policymakers also have distributional concerns. To operationalize this idea, we vary the welfare weights on students with below- versus above-median prior-year math scores, ω in Equation 5, from 0.0 (only care about above-median students) to 1.0 (only care about below-median students).[35] 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.

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

[35] 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 5 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.[36]

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.[37] Interestingly, the frontier of second-best gains dominates many race-focused assignments (as in Delgado, 2025). Appendix Figure A.6 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.[38]

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., 2025). 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 D.1, 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 (0.03 versus under 0.01); while the optimal district-wide assignment is

[36] 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.

[37] 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.

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

much better for both groups, it reverses the incidence, generating relatively larger gains for higher-achieving students (0.12 versus 0.07). 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.[39] 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.[40] Appendix Figure A.7 depicts the distribution of test-score changes under various policies, illustrating both harms and gains. Second, while the model cannot back out welfare weights justifying the status quo, the gains at the second best do reflect the shadow cost of relaxing current allocative constraints. In this sense, there may be large social gains from finding ways to implement low-cost reassignments, such as the within-school optimum or targeted exchanges. 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.

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 with their causal effects

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

[40] 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).

on earnings using the estimates of Chetty et al. (2014b):[41]

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

where $\hat{\mu}_{i,j}^{s}$ are the jointly estimated and shrunk effects of teacher j on outcome s 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 20 and 65; (2) gains from higher test scores apply to earnings at all ages; and (3) earnings gains are discounted at 3%.[42] 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 include returns to noncognitive gains as well, revealing that our main results tend to be conservative.

We define the welfare maximization problem as

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

which is identical to Equation 5, but optimizes over the predicted change in score $\Delta\hat{S}_{i,j}$ as a function of our jointly shrunk value-added scores on math and reading by lagged-achievement. We show in Appendix D.3 that the second-best assignment combines information from above- and below-median math value added and above- and below-median reading value added—although the expected gains from using terciles, quartiles, or quintiles are similar. In particular, Appendix Figure D.1 shows that using above- and below-median math value added and above- and below-median reading value added produces 1.2–1.6× lower MSE than other models, and Appendix D.3 shows that this model most effectively balances expected gains and *ex post* regret.

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 1989–2009, 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

[41] 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.

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

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 6 will produce lower gains than a policy that assigns effective teachers to subgroups with stronger links between test scores and earnings.

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.

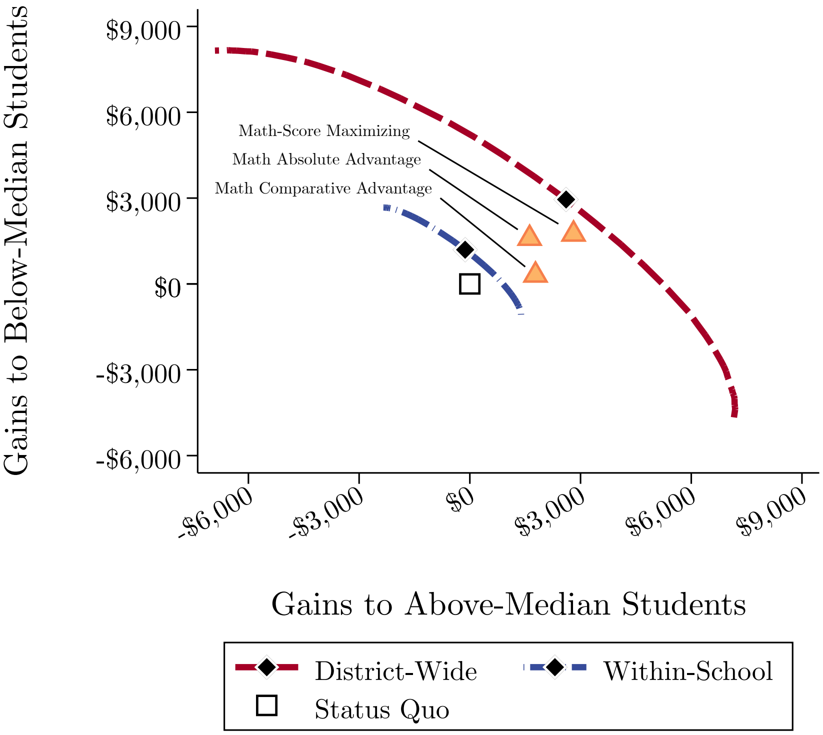
Figure 5 reveals substantial benefits from optimal teacher assignment. We estimate that the second-best optimum under district-wide assignment would generate \$3,000 in present-value earnings for below-median students and \$2,600 for above-median students (\$2,800 on average). These large gains are 2.2 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 \$5,200 without hurting high-scoring students on average. A similar comparison reveals gains of about \$1,100 from within-school assignments.[43] 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 intuition 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 2.7× larger gains (+\$1,800) than the comparative advantage heuristic proposed by Delgado (2025), 1.7× larger gains (+\$1,200) than the optimal assignment using standard value added mentioned in Bates et al. (2025), and even 1.2× larger gains (+\$500) than the district-wide

[43] 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-average 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. Each assignment solves the problem in Equation 6 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.

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 156% (about $1,800 more than teacher deselection), 86% (about $1,350) more than an optimal assignment based on standard math value added, and 70% (about $1,200) 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 $625 million. Assuming comparable gains outside of the San Diego Unified School District, implementing this policy in all U.S. public schools[44] over ten years would generate gains on the order of $100 billion.[45]

Consider two notes on interpreting these results. First, although these gains are large, using both cognitive and noncognitive outcomes produces even larger gains. 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).[46] Appendix Figure A.8 compares these assignments to the earnings-maximizing assignment in Figure 5, showing up to 60% larger gains when using noncognitive outcomes to make assignments. This suggests a second-best frontier that is 1.5–4.6 larger than other policy proposals.

Second, within-school assignments are a promising low-cost policy. In practice, the multi-billion-dollar gains 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.

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. (2025), 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

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

[45] 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.

[46] 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).

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

$$MVPF^{\mathcal{J}}(b) = \frac{\sum_i (1-t)\Delta \hat{S}_{i,\mathcal{J}(i)}}{N_j b - \sum_i t\Delta \hat{S}_{i,\mathcal{J}(i)}} \tag{7}$$

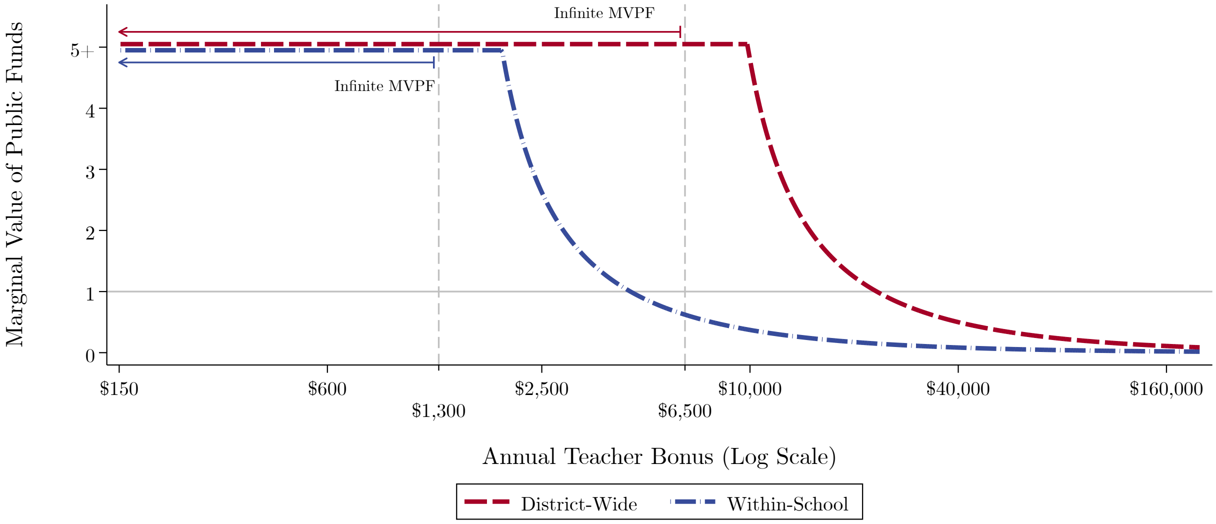
where $(1-t)\Delta S_{i,\mathcal{J}(i)}$ are the after-tax present-value monetary gains to each student from assignment $\mathcal{J}$ (given tax rate t), N_j is the number of teachers, and $t\Delta \hat{S}_{i,\mathcal{J}(i)}$ is the present-value of gains recouped as tax revenue. We focus on earnings gains, $\Delta \hat{S}_{i,\mathcal{J}(i)}$, from the second-best assignment with equal weights on all students and calibrate $t = 0.28$ with data from the Opportunity Atlas for San Diego (Chetty et al., 2018).[47]

Two key assumptions are needed to make this statistic meaningful in practice. First, the MVPF in Equation 7 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 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 $N_j b < \sum_i t\Delta \hat{S}_{i,\mathcal{J}(i)}$, have an "infinite" MVPF, as indicated in the figure.

[47] 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.

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 $6,500 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 $1,300 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 not infinite, they can be 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 reallocation 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 teachers to change schools (Glazerman et al., 2013).[48] For the within-school

[48] This experiment offered 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

reallocation, 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 $1,000 to switch classes without hesitation. The MVPF of these policies is greater than one for bonuses up to $23,000 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.

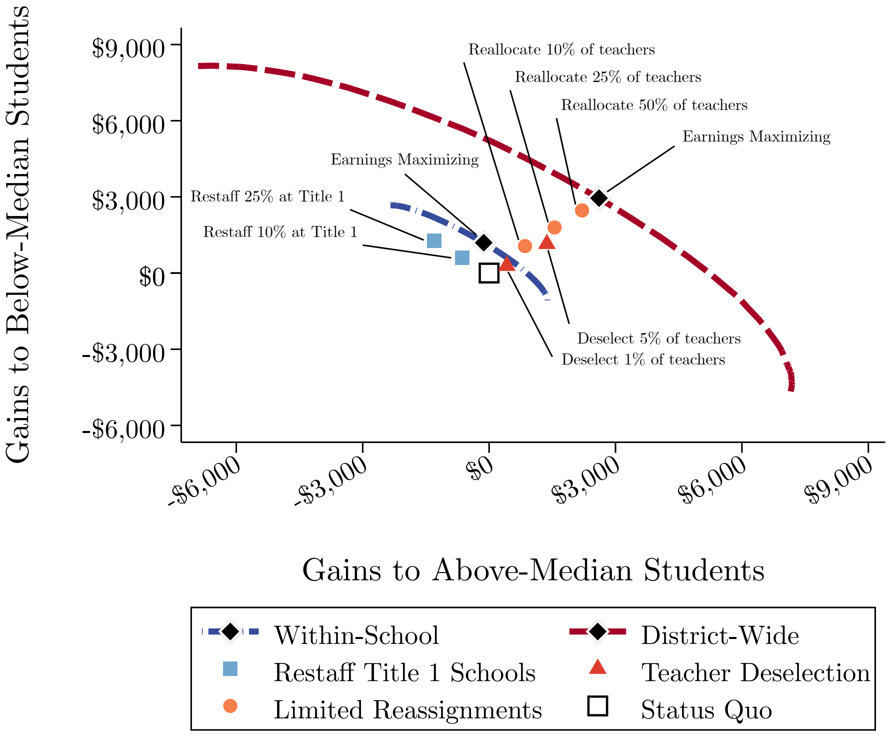
Limited Reassignments. Because it may be more valuable to reassign certain teachers than others and because there may be diminishing marginal returns to additional reassignments, we consider a set of policies that restrict district-wide assignment to a limited share of teachers. We implement this counterfactual by assuming uniform welfare weights and constraining the linear program to change the assignments of no more than X% of teachers in 1% increments. We find that reallocating (and paying bonuses to) as few as 25% of teachers achieves over 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

who transferred was $46,604 in 2010.

expected, there are sizable gains. Removing the lowest 5% (1%) of teachers would achieve gains equal to about 45% (13%) 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 $55,000 to each removed teacher would have an infinite MVPF in either case.[49]

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.

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

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

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 $600 ($1,300 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 use that approach to describe optimal second-best teacher assignment 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 value added based assignments. Back-of-the-envelope calculations suggest large welfare gains from these policies, even if they are quite expensive to implement. We also find 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 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 (2024) "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.

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.

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, 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 (2025) “Gains from Alternative Assignment? Evidence from a Two-Sided Teacher Market,” IZA 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.

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.

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

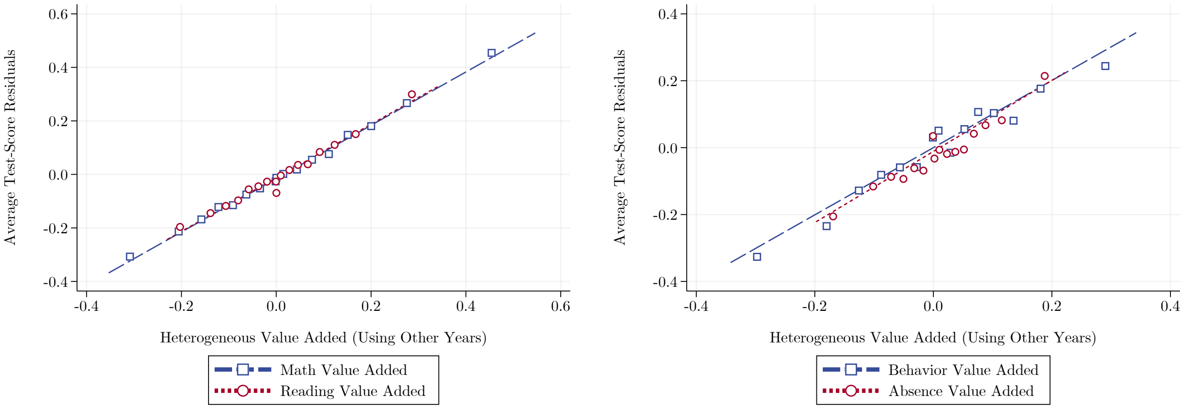
Tikhonov, Andrei Nikolaevich, Alexander V Goncharsky, Vaceslav Vasilevic Stepanov, and Anatoly Grigorevich Yagola (1995) "Numerical Methods for the Approximate Solution of Ill-Posed Problems on Compact Sets," in *Numerical Methods for the Solution of Ill-posed Problems*: Springer, pp. 65–79.

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

Online Appendix

A. Additional Tables and Figures

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

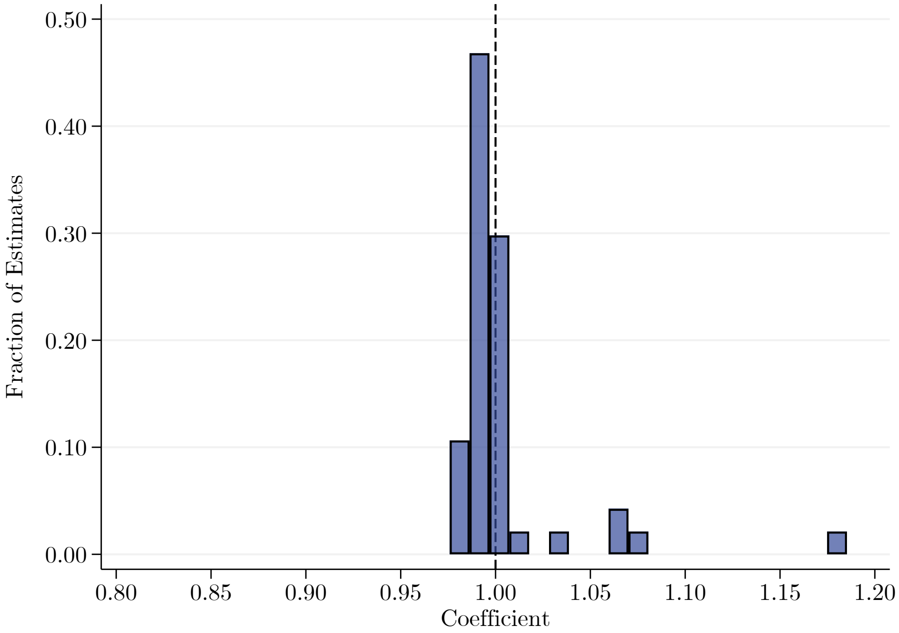


(a) Math and Reading

(b) Behavior and Absences

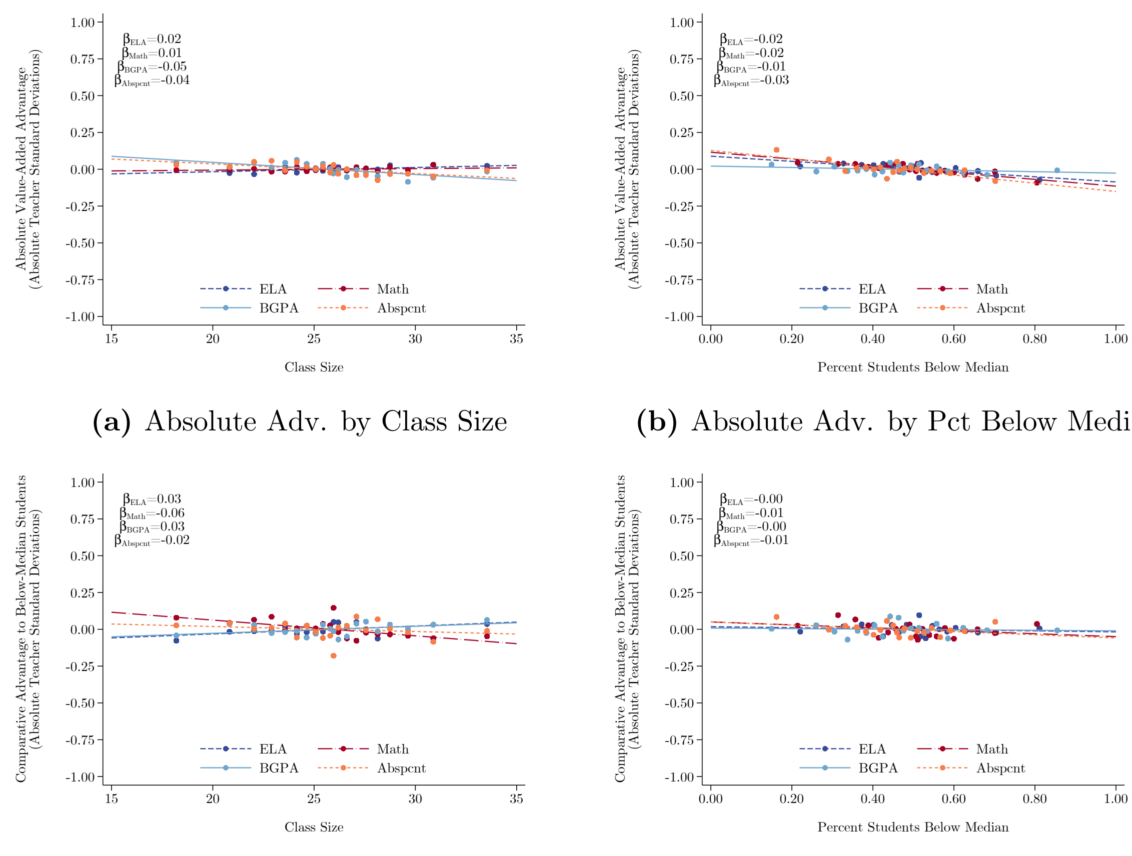
Note: This figure shows the relationship between average residuals for each teacher's class and their predicted value added (average estimated match effect on students in the class) based on prior years of data.

Figure A.2. All Models all 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 non-cognitive outcomes, i.e., standard or median split behavioral GPA value added and standard or median split attendance value added.

Figure A.3. Teachers' Value Added Doesn't Change with Class Size or Composition



(a) Absolute Adv. by Class Size

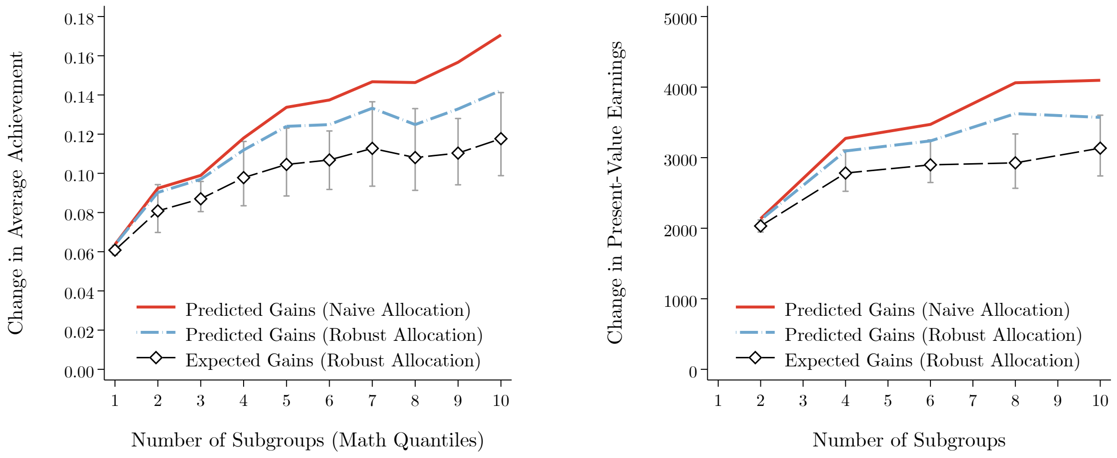
(b) Absolute Adv. by Pct Below Median

(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 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 within-teacher variation over the number of students in class where β reports the within-teacher change associated with a five-student change in class size. The panels on the right plot the within-teacher variation over the share of lower-achieving students where β reports the within-teacher change associated with a 10 percent change in the fraction of below-median students in the class.

Figure A.4. Predicted Gains Are Far too Optimistic

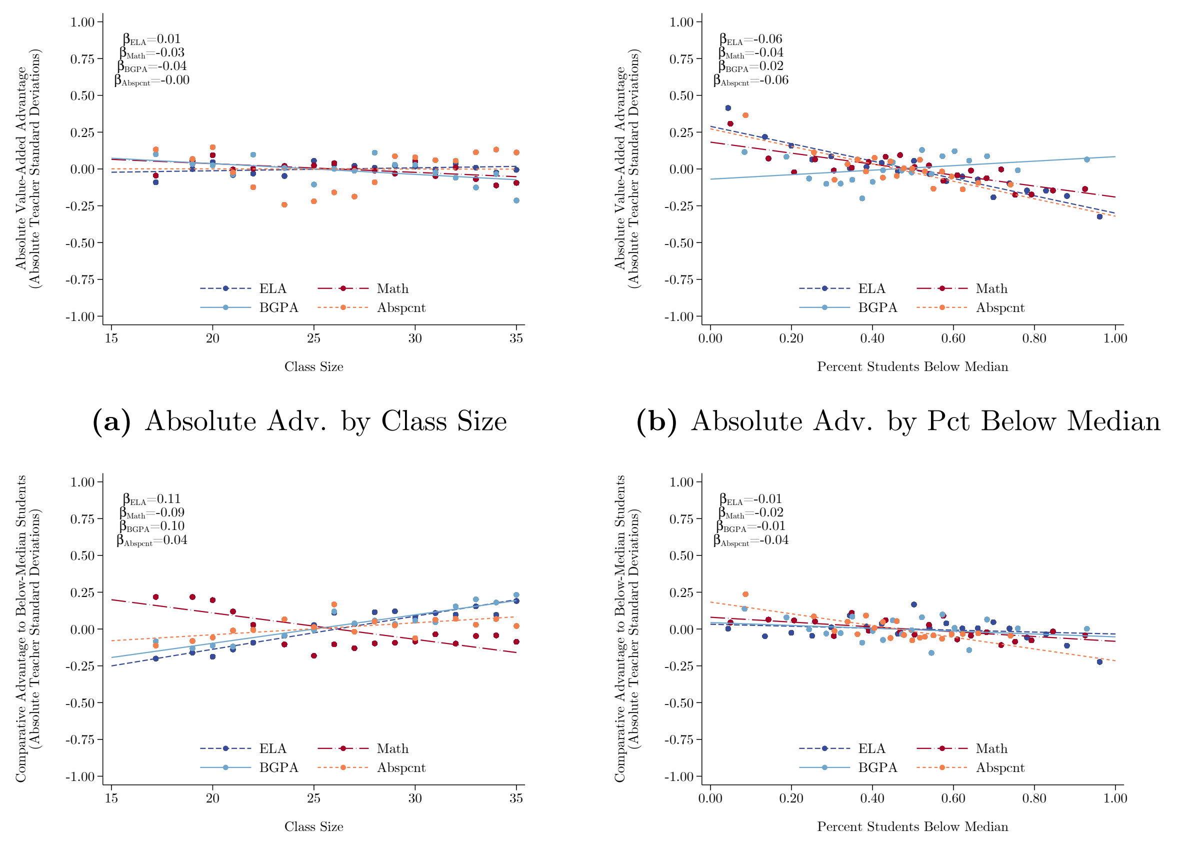


(a) Math Scores

(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. Status Quo Assigns Lower-Achieving Students Worse Teachers



(a) Absolute Adv. by Class Size

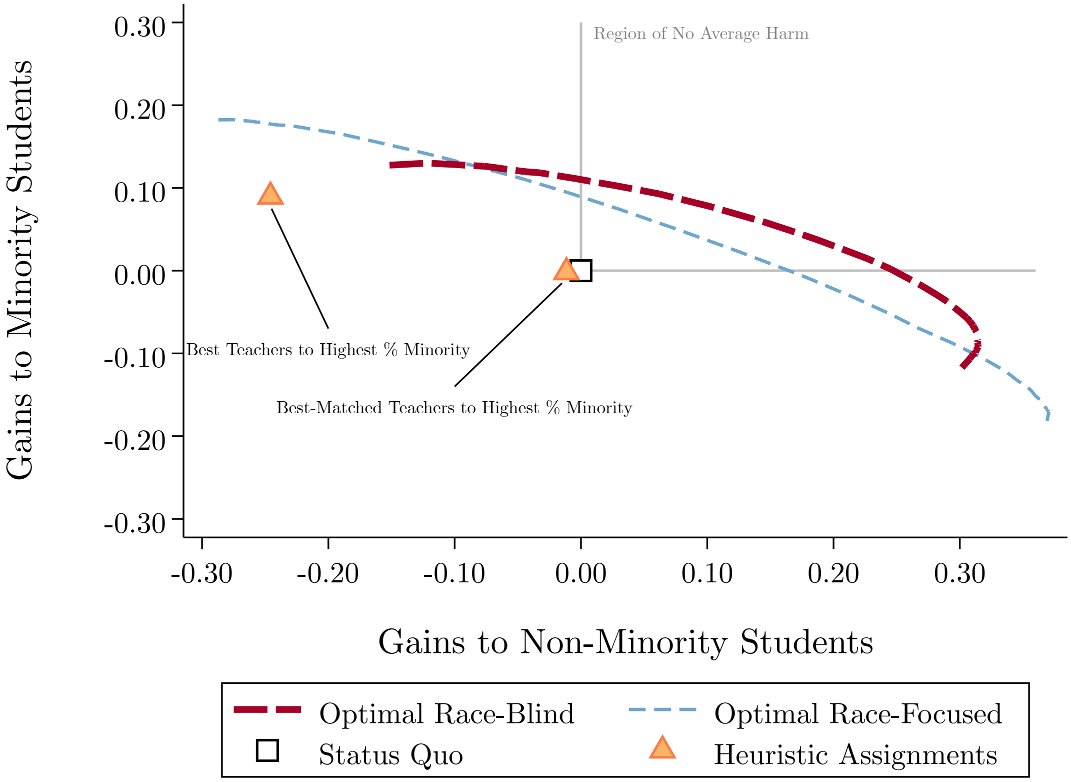
(b) Absolute Adv. by Pct Below Median

(c) Comparative Adv. by Class Size

(d) Comparative Adv. by Pct Below Median

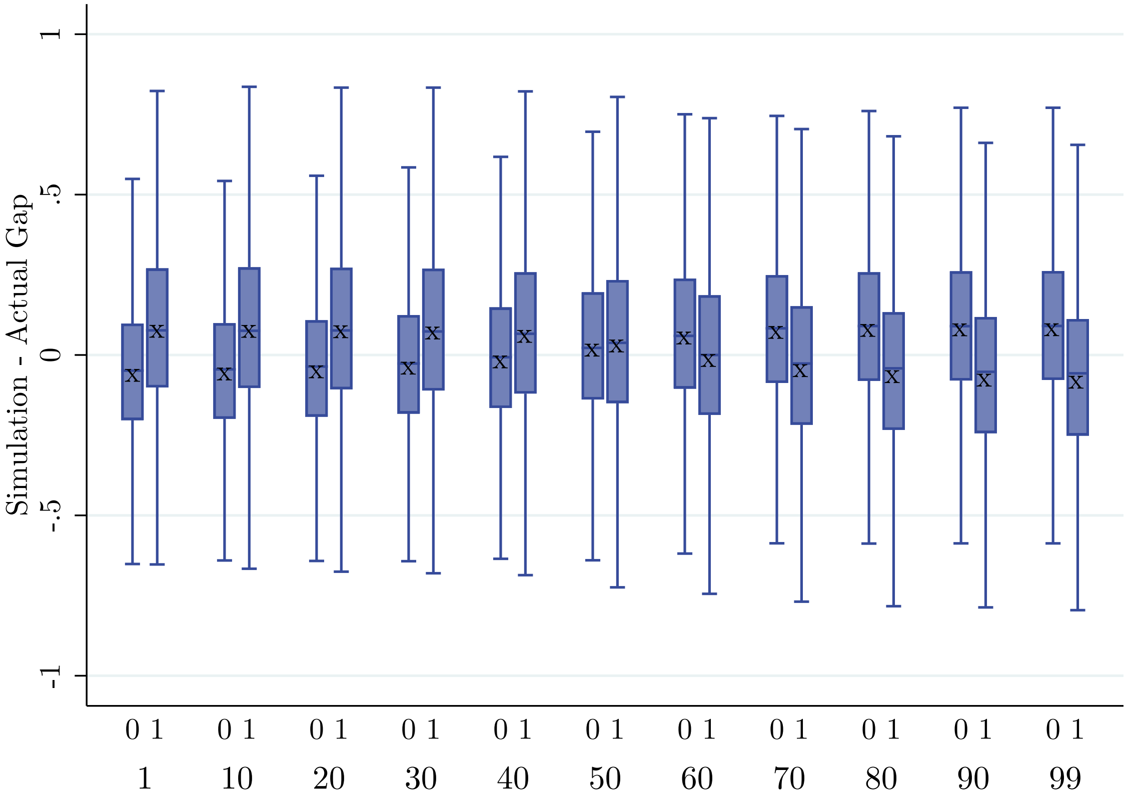
Note: This figure shows how our heterogeneous estimates of teacher value-added on reading, math, behavior, and absences relate to status quo class size and class composition. The top panels show teacher absolute advantage (average value added among higher- and lower-scoring students) and the bottom panels 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.6. Race-Blind and Race-Conscious Reassignments



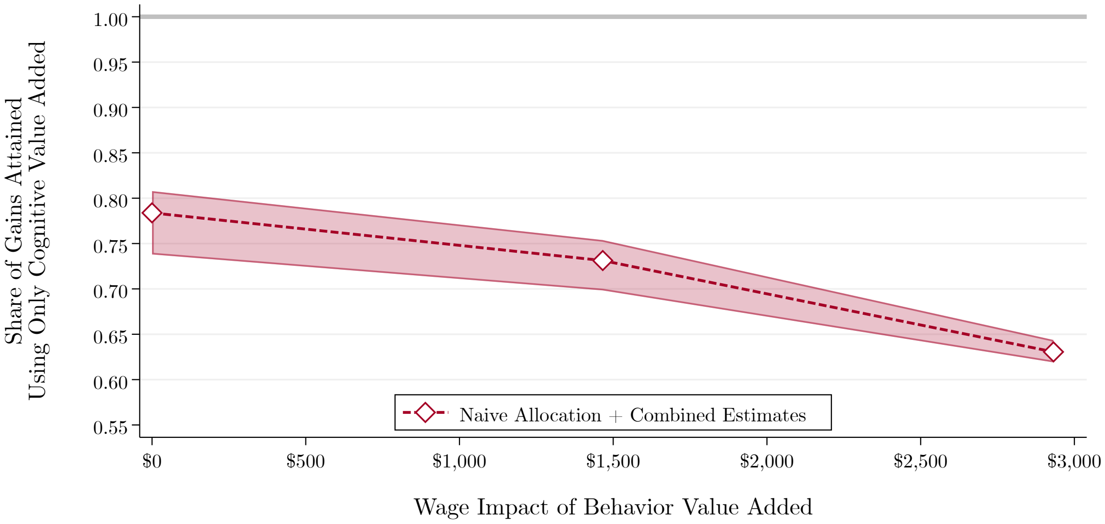
Note: This figure compares the gains from reassignments by student race using two racial groups: (1) Black and Hispanic students and (2) other students. We report two sets of results from optimal reassignments: one using value added by achievement and one using value added by race. We also report two 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.7. 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.8. Percent of Gains from Assignments Ignoring Non-Cognitive Skills



Note: This figure reports the overall share of gains attained by a naive assignment ignoring non-cognitive 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 non-cognitive scores on earnings.

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

	Math	Math	Reading	Reading	Behavior	Behavior	Absences	Absences
	Below	Above	Below	Above	Below	Above	Below	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. Naive and Robust Reassignment Results

	Math Standard	Gender	Race	Math Median	Math Terciles	Math Quartiles	Math Quintiles	Gender x Median	Race x Median
				Split				Split	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.5. 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.

B. Theory Appendix

B.1 Derivation of Equation 1

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, $\boldsymbol{X}_i$. Assume that an individual's outcomes affect only the utility and welfare weights of that individual i. Then the expected welfare $\mathcal{W}$ under policy $\mathcal{J}$, given $\boldsymbol{Y}$ and the scores $\boldsymbol{S}^{\mathcal{J}}$ can be written as

$$\begin{aligned}
\mathbb{E}[W^{\mathcal{J}}|\boldsymbol{S}^{\mathcal{J}}] &= \sum_{i=1}^{n} \mathbb{E}[\phi_i U_i^j | S_i^{\mathcal{J}}] \\
&= \sum_j \sum_{i:\mathcal{J}(i)=j} \mathbb{E}[\phi_i U_i^j | S_i^{\mathcal{J}}] \\
&= \sum_j \sum_{i:\mathcal{J}(i)=j} \frac{\mathbb{E}[\phi_i U_i^{\mathcal{J}} | S_i^{\mathcal{J}}]}{S_i^{\mathcal{J}}} S_i^{\mathcal{J}} \\
&\equiv \sum_j \sum_{i:\mathcal{J}(i)=j} \omega_i^{\mathcal{J}} S_i^{\mathcal{J}}
\end{aligned}$$

where the first equality follows from the no-spillovers assumption, the second follows from the surjectivity of $\mathcal{J}$ (all students get taught), the third by multiplying and dividing by $S_i^{\mathcal{J}}$, and the fourth by defining $\omega_i^{\mathcal{J}} = \phi_i \frac{\mathbb{E}[U_i^{\mathcal{J}} | S_i^{\mathcal{J}}]}{S_i^{\mathcal{J}}}$.

Furthermore, if $s(\cdot)$ is additively separable into a student- and a match-specific component $(\mu_i + \mu_{i,j})$ with the score match effect for the status quo, $\mu_{i,\mathcal{J}_0(i)}$, normalized to zero, and if it is an unbiased, strictly linear predictor of U, then the expected difference between any assignment and the status quo can be written as

$$\begin{aligned}
\mathcal{W}^{\mathcal{J}} &\equiv \sum_j \sum_{i:\mathcal{J}(i)=j} \omega_i^{\mathcal{J}} S_i^{\mathcal{J}} - \sum_j \sum_{i:\mathcal{J}_0(i)=j} \omega_i^{\mathcal{J}_0} S_i^{\mathcal{J}_0} \\
&= \sum_j \sum_{i:\mathcal{J}(i)=j} \left[\omega_i(\mu_i + \mu_{i,j}) - \omega_i(\mu_i + \mu_{i,\mathcal{J}_0(i)})\right] \\
&= \sum_j \sum_{i:\mathcal{J}(i)=j} \omega_i \mu_{i,j}
\end{aligned}$$

where the first equality leverages the fact that $s()$ is additively separable and that if $s()$ is unbiased and strictly linear, then $\mathbb{E}[U|S] = \alpha S$, so $\forall \mathcal{J}\ \omega_i^{\mathcal{J}} = \phi_i \alpha \equiv \omega_i$. The second equality follows from the normalization of the status quo match to zero.

B.2 Derivation of Equation 2

Let expected welfare be $\mathcal{W}^{\mathcal{J}}$ as defined in Equation 1, let $\hat{\mu}_j$ be an estimate of teacher j's value added, and let n_j is the number of students in the class to which teacher j is assigned. Consider the bias we would introduce by approximating welfare as $\widehat{\mathcal{W}}^{\mathcal{J}}_{VA} = \sum_j n_j \bar{\omega}_j \hat{\mu}_j$

$$
\begin{aligned}
\mathcal{W}^{\mathcal{J}} - \widehat{\mathcal{W}}^{\mathcal{J}}_{VA} &\equiv \sum_j \sum_{i\colon \mathcal{J}(i)=j} \omega_i \mu_{i,j} - \sum_j n_j \bar{\omega}_j \hat{\mu}_j \\
&= \sum_j n_j \left[\left(\frac{1}{n_j} \sum_{i\colon \mathcal{J}(i)=j} \omega_i \mu_{i,j} \right) - \bar{\omega}_j \hat{\mu}_j \right] \\
&= \sum_j n_j \left[\frac{1}{n_j} \sum_{i\colon \mathcal{J}(i)=j} \omega_i \mu_{i,j} + \bar{\omega}_j \left(\tilde{\mu}^{\mathcal{J}}_j - \tilde{\mu}^{\mathcal{J}}_j + \bar{\mu}_j - \bar{\mu}_j - \hat{\mu}_j \right) \right] \\
&= \sum_j n_j \left[\underbrace{\bar{\omega}_j \left(\tilde{\mu}^{\mathcal{J}}_j - \bar{\mu}_j \right)}_{\text{Matching Gains}} + \underbrace{\widehat{\text{Cov}}_j(\omega_i, \mu_{i,j})}_{\text{Distributional Gains}} + \underbrace{\bar{\omega}_j(\bar{\mu}_j - \tilde{\mu}^0_j)}_{\text{External Validity}} + \underbrace{\bar{\omega}_j(\tilde{\mu}^0_j - \hat{\mu}_j)}_{\text{Estimation Error}} \right]
\end{aligned}
$$

The first equality comes from factoring out n_j. The second from adding and subtracting $\tilde{\mu_j}^{\mathcal{J}}$ and $\bar{\mu}_j$, where $\bar{\omega}_j$ and $\tilde{\mu}^{\mathcal{J}}_j$ represent the average welfare weights of students in teacher j's assigned class and the average match effect of teacher j on those students, and $\bar{\mu}_j$ is their population average match effect (absolute advantage). The third comes from adding and subtracting $\bar{\omega}_j \tilde{\mu}^0_j$, 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 $\widehat{\text{Cov}}_j(\omega_i, \mu_{i,j}) \equiv \frac{1}{n_j} \sum_{i\colon \mathcal{J}(i)=j} \omega_i \mu_{i,j} - \bar{\omega}_j \tilde{\mu}^{\mathcal{J}}_j$).

Expositional Note: For pedagogical clarity this equation is depicted as if welfare is measured in levels, but recall that formally we defined the welfare of assignment $\mathcal{J}$ in differences from the status quo assignment, $\mathcal{J}_0$. As such, $\tilde{\mu}^0_j$ has technically been normalized to zero. Because all other μs have been similarly normalized, this doesn't change the direction or magnitude of bias, but we feel that including $\tilde{\mu}^0_j$ in the equation is useful for building intuition for the core trade-offs.

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

Non-Cognitive 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 3rd-5th grade classrooms from the 1998–1999 through 2018–2019 school years and make five restrictions intended to target typical teachers in SDUSD teaching traditional 3rd-5th 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 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 reasons 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:

$$S_{i,s,t} = \mu_{\mathcal{J}(i,t),s,k,t} + \beta_{s,k} X_{i,t} + \phi_{\ell(i,t),s} + \phi_{s,t} + \theta_{\mathcal{C}(i,t),s,k} + \varepsilon_{i,s,t}$$

where $S_{i,s,t}$ is the score from student i for outcome s in year t, $\mu_{\mathcal{J}(i,t),s,k,t}$ is the value added from the assigned teacher j for students of type k, $X_{i,t}$ are the observed student characteristics, $\phi_{\ell(i,t),s}$ are all school-specific factors impacting student achievement on outcome s from the student's assigned school $\ell(i,t)$, $\phi_{s,t}$ are outcome specific time effects, $\theta_{\mathcal{C}(i,t),s,k}$ are outcome-specific classroom shocks to students of type k for assigned classroom $\mathcal{C}(i,t)$, and $\varepsilon_{i,s,t}$ 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 1. **(Joint Stationarity)** Assume that subgroup-specific teacher effects, classroom shocks, and student-level shocks follow a stationary process.

$$\begin{aligned} \mathbb{E}[\mu_{\mathcal{J}(i,t),s,k,t}|s,k,t] = \mathbb{E}[\theta_{C(i,t),s,k,t}|s,k,t] = \mathbb{E}[\varepsilon_{i,s,t}|s,k,t] &= 0 \\ Cov(\mu_{\mathcal{J}(i,t),s,k,t}, \mu_{\mathcal{J}(i,t),s,m,t+h}) &= \sigma_{\mu_k,\mu_m,h} \\ Cov(\theta_{C(i,t),s,k,t}, \theta_{C(i,t),s,m,t}) &= \sigma_{\theta_k,\theta_m} \\ Cov(\varepsilon_{i,s,t}, \varepsilon_{i,s,t+h}) &= \sigma_{\varepsilon_{s,k},h} \\ Cov(\varepsilon_{i,s,t}, \varepsilon_{i,s',t}) &= \sigma_{\varepsilon_s,\varepsilon_{s'}} \end{aligned}$$

Assumption 2. **(Fixed-Effect Independence)** Let $\bar{\alpha}_{j,s,k}$ 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.

$$(\alpha_{j,s,k,t} - \bar{\alpha}_{j,s,k}) \perp \phi_{\ell(j,t),s} \forall k$$

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 domain, s.

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

$$S_{i,s,t} = \beta_{s,k} X_{i,t} + \alpha_{\mathcal{J}(i,t),s,k,t} + v_{i,s,t}$$

where i indexes students, s 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.[50]

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

$$\hat{v}_{i,s,t} = S_{i,s,t} - \hat{\beta}_{s,k} X_{i,t}$$

and project these residuals separately for each outcome s onto teacher ($\alpha_{j,s}$), school ($\phi_{\ell(i,t),s}$), and year fixed effects ($\phi_{s,t}$), as well as a teacher experience function $f_s(z)$. 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 for across outcomes s.

$$\begin{aligned}\hat{v}_{i,s,t} &= \sum_{e=1}^{6} \delta_s^e \mathbb{1}\{z_{\mathcal{J}(i,t),t} = e\} + \delta_s^7 \mathbb{1}\{z_{\mathcal{J}(i,t),t} > 7\} + \alpha_{\mathcal{J}(i,t),s} + \phi_{\ell(i,t),s} + \phi_{s,t} + \eta_{i,s,t} \\ &= f_s(z_{\mathcal{J}(i,t),t}) + \alpha_{\mathcal{J}(i,t),s} + \phi_{\ell(i,t),s} + \phi_{s,t} + \eta_{i,s,t}\end{aligned}$$

Third, we form a second set of student-level residuals in which we remove the school

[50] 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.

effects and teacher-experience effects:

$$\begin{aligned} A_{i,s,k,t} &= \hat{v}_{i,s,t} - \left(\hat{f}_s(z_{\mathcal{J}(i,t),t}) + \hat{\phi}_{\ell(i,t),s}\right) \\ &= S_{i,s,t} - \hat{\beta}_{s,k} X_{i,t} - \hat{f}_s(z_{\mathcal{J}(i,t),t}) - \hat{\phi}_{\ell(i,t),s} \end{aligned}$$

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

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

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

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

The matrix $\boldsymbol{\psi}_j'$ is a teacher-specific set of reliability weights with dimensions $M \times M(t-1)$, where M is the number of outcome-subgroup divisions.[51]. 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 $\sigma_{\epsilon_s,\epsilon_s'}$ 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

[51] 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 $M = 5$)

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 $\sigma_{\mu_k,\mu_m,8}$ for any k, m. Our final estimate for teacher j's value added in year t on outcome s for students in subgroup k is as follows:

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

where $\hat{\boldsymbol{\psi}}_{s,k}$ 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 hyper-parameters. 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).

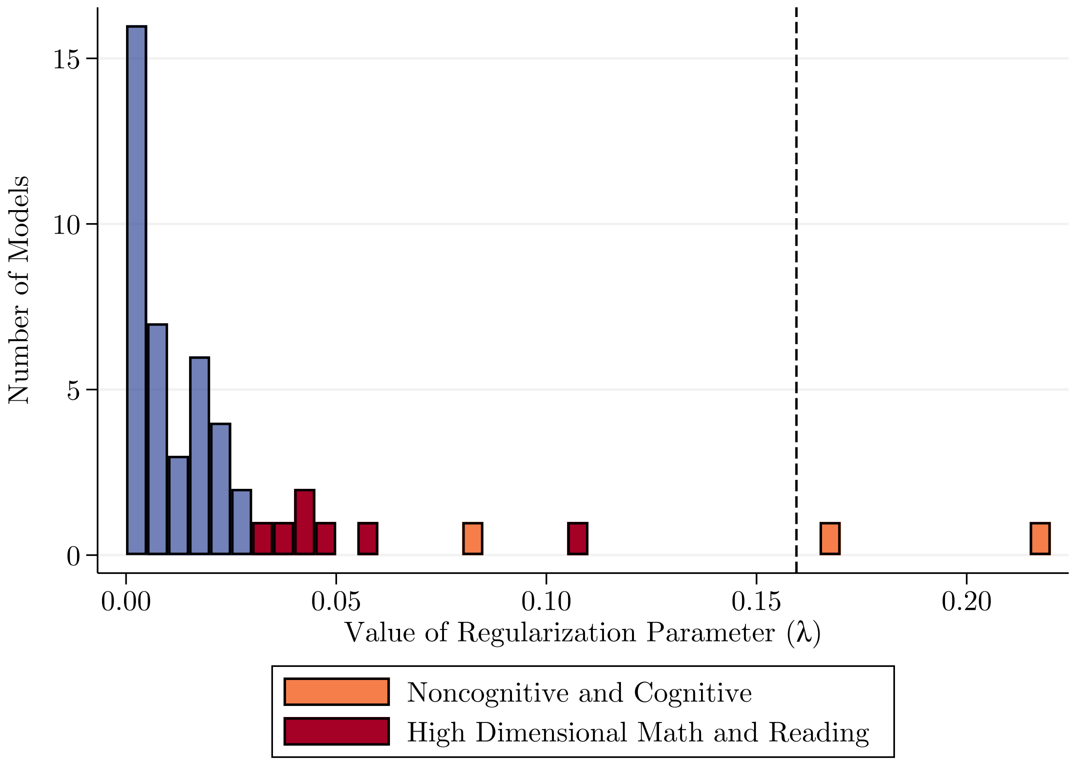
In practice, we choose a unique λ for each model by minimizing the forecast bias $(1-\beta)$ for the estimates (an approximate out-of-sample prediction[52]). We use a standard optimization routine, initializing at $\lambda = 0.005$, calculating the implied estimates, estimating the forecast bias, and 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 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.

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.

[52] Forecast bias is not precisely an out-of-sample prediction because data from all years are used to estimate the hyper-parameters—including the year being forecast. However, given the large number of teachers, years, and students, leakage is small.

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.

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 and 0.7. This suggests a strong relationship between our preferred estimates and those obtained using this alternative specification.

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.

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.

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. We do so in the following way. First, when possible, we use the current-year score rather than the lagged score because it is the closest proxy for lagged scores. Second, if the current-year score is also missing, we use the student's modal subgroup for the relevant outcome across all years in which we observe the student. Third, we use the subgroup of a randomly drawn peer of the same gender, race, school, year, and grade. Finally, in a very small number of cases, we use a random peer from students in their school in that same year or, as a last resort, from students in their school at any time and assign the student to a subgroup. This procedure ultimately classifies all unassigned students and allows us to account for them properly when assigning teachers across classes. Additionally, we include teachers in the reassignment exercise only if we have estimates for their value added on the relevant subgroups and outcomes. We retain teachers without value-added estimates in the same class they actually taught (effectively omitting that class from the reassignment exercise). Ultimately, because we impute value added for missing subgroups or outcomes using the Empirical Bayes predictions, this restriction makes little empirical difference in our main specification, but it drives some of the differences between rows in Appendix Table A.4.

We formulate the mixed-integer linear programming problem in the following way. We first assign teachers and classes to reassignment "cells", i.e., the set of classes that are allowable assignments for those teachers. Each cell is effectively its own optimization problem since we allow for no assignments across cells, so for the remainder of the section we will treat the problem as if there were only one cell. For within-school swaps, these cells are school-grade-year cells, and for across-school swaps they are grade-year cells. We then represent each possible assignment as a binary variable in the optimization problem. For example, $x_{j,c}$ takes the value of one if teacher j was assigned to class c and zero otherwise. In each case, the status quo class to which teacher j is assigned to is class $c = j$.

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 $C : (i, t) \rightarrow c$ 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:

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

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).[53] This means that if we are unable to

[53] 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

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.[54] We then sum these grade-specific averages into the total expected three-year gain for students of each type.

D.3 Comparing Models and Identifying the Second Best

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.

For many of these criteria, the following definitions will be helpful.

Definition 1. (Predicted Gains) Denoting estimate b (typically from B bootstrapped replications) of model m as $\hat{\boldsymbol{\mu}}^{m_b}$, the predicted gains from policy $\mathcal{J}$ are then

$$\hat{\theta}^{m_b}_{\mathcal{J}} = \frac{1}{N} \sum_{j} \sum_{i:\, \mathcal{J}(i)=j} \hat{\mu}^{m_b}_{i,j}$$

Furthermore, denoting the assignment chosen by the social planner based on the original value-added estimates, m_0, as $\mathcal{J}^*_m$, the predicted gains from an optimal assignment are $\hat{\theta}^{m_0}_{\mathcal{J}^*_m}$, and the bootstrapped evaluations of this assignment are contained in the set $\{\hat{\theta}^{m}_{\mathcal{J}^*_m}\}$.

Definition 2. (Oracle Gains) For any model m, take the set of estimates, b, as a true oracle of the effects. The set of oracle gains is $\{\hat{\theta}^{m}_{\mathcal{J}^b_m}\}$, where $\mathcal{J}^b_m$ is the optimal assignment under the oracle.

D.3.1 Mean Squared Error

One intuitive criterion for trading off matching gains and misallocation risk is the mean squared error. A model with greater matching gains will have less bias relative to the first-best optimum, and models with less misallocation risk will have lower variance. We adapt the MSE criterion to this setting as follows:

assigned to in the status quo.

[54] 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.

Definition 3. (Adapted MSE) Following Definition 1, the true gains from any policy are $\theta_{\mathcal{J}} = \frac{1}{N}\sum_j \sum_{i:\, \mathcal{J}(i)=j} \mu_{i,j}$ and the gains from the first-best assignment are $\theta^* = \max_{\mathcal{J}} \theta_{\mathcal{J}}$. The mean squared error of model m relative to the first-best assignment is then

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}^m) &= \mathbb{E}\left[\left(\hat{\theta}^m_{\mathcal{J}^b_m} - \theta^*\right)^2\right] \\ &= \mathbb{V}(\hat{\theta}^m_{\mathcal{J}^b_m}) + \left(\mathbb{E}[\hat{\theta}^m_{\mathcal{J}^b_m}] - \theta^*\right)^2\end{aligned}$$

Unfortunately, defining the MSE alone is insufficient to compare models and assignments because the gains from the first-best assignment are unknown. To overcome this limitation, we draw plausible values for θ^* and calculate the expected MSE for each model m:

$$\overline{\mathrm{MSE}}(\hat{\theta}^m) = \int_{\theta'} \mathbb{V}(\hat{\theta}^m_{\mathcal{J}^b_m}) + \left(\mathbb{E}[\hat{\theta}^m_{\mathcal{J}^b_m}] - \theta'\right)^2 \mathrm{d}\theta'$$

In practice, we estimate $\overline{\mathrm{MSE}}(\hat{\theta}^m)$ using the sample analogues $\hat{\mathbb{E}}[\theta^m_{\mathcal{J}^b_m}] = \frac{1}{B}\sum_b \hat{\theta}^{m_b}_{\mathcal{J}^b_m}$ and $\hat{\mathbb{V}}(\theta^m_{\mathcal{J}^b_m}) = \frac{1}{B-1}\sum_b \left(\hat{\mathbb{E}}[\theta^{m_b}_{\mathcal{J}^b_m}] - \hat{\theta}^{m_b}_{\mathcal{J}^b_m}\right)^2$. We also draw θ' 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 MSE across models is shown in Figure D.1. 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.

D.3.2 Misallocation Risk as Oracle Regret

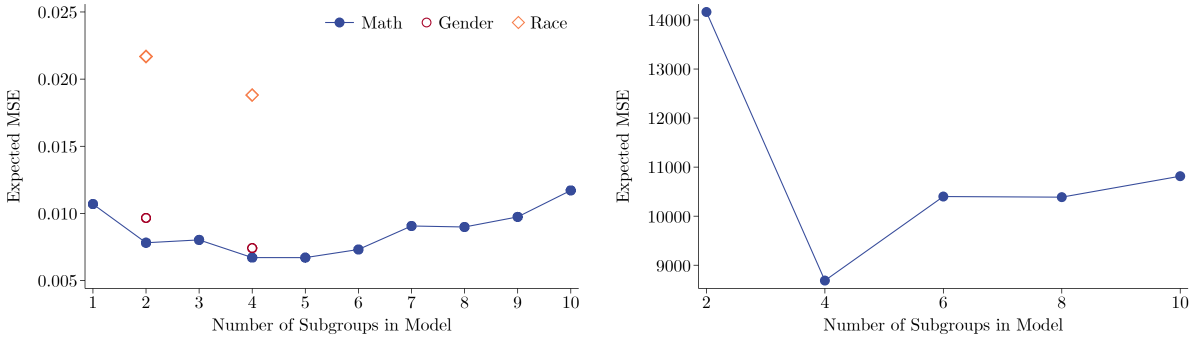
Definition 4. (Oracle Regret) We characterize the distribution of regret under each model as the foregone gains in assignment $\mathcal{J}^*_m$ relative to the oracle optima $\mathcal{J}^b_m$. For each oracle, the oracle regret is defined as

$$\mathrm{Regret}(m_b) = \hat{\theta}^{m_b}_{\mathcal{J}^b_m} - \hat{\theta}^{m_b}_{\mathcal{J}^*_m}$$

A maximization problem that trades off matching gains and misallocation risk can be derived from regret if the social planner chooses a model to maximize expected outcomes subject to an arbitrary constraint on regret, R:

$$\max_m \mathcal{L}(m) = \max_m \mathbb{E}[\hat{\theta}^m_{\mathcal{J}^*_m}] - \kappa\, \mathbb{E}[\mathrm{Regret}(m_b) - R]$$

Figure D.1. MSE across Models with Differing Complexity



(a) MSE across Math Models

(b) MSE across Earnings Models

Note: Each point in the figure represents the average MSE 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" include male versus female value added in the two subgroup case and male/female value added by above- and below-median math achievement in the four-subgroup case. Those labeled "Race" include Black and Hispanic/Non-Black and Non-Hispanic value added in the two subgroup case and Black and Hispanic versus Non-Black and Non-Hispanic value added by 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.

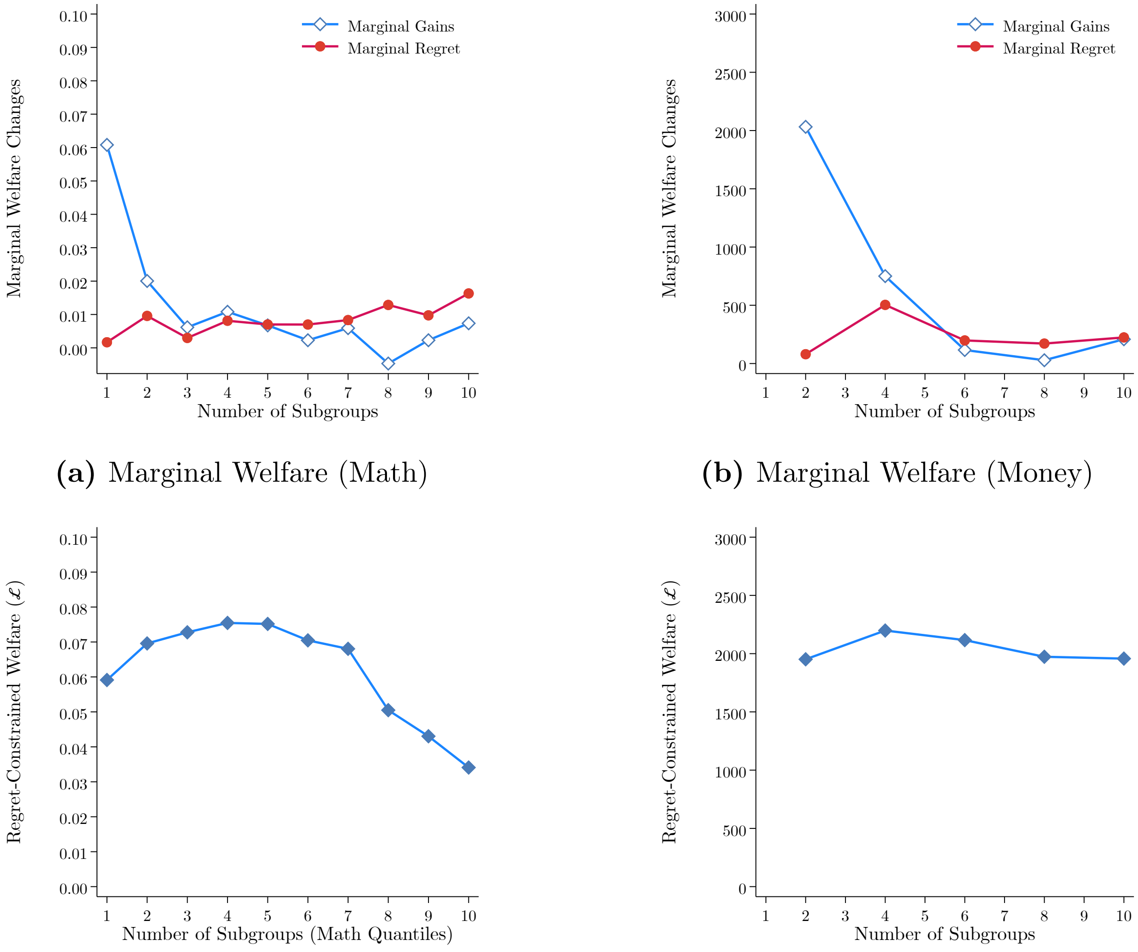
where κ is the shadow value of relaxing the regret constraint. Although the shadow price κ is unknown, we can characterize the optimal model under any given value of κ.

In the data, expected gains to both math scores and lifetime earnings are concave in model complexity, whereas regret is convex. Appendix Figure D.2 uses this objective function to construct an empirical analogue of Panel (b) from Figure 1, assuming $\kappa = 1$. We choose $\kappa = 1$ because expected gains and regret are already in the same units; however, the results are similar for math and identical for earnings as long as $\kappa \in (0.6, 1.5)$. Panels (a) and (b) depict the marginal increase in expected gains from added model complexity against the marginal increase in misallocation risk measured by regret. Panels (c) and (d) plot the objective function for all models over the distribution of model complexity. This approach indicates that the best model for math uses achievement quartiles and that the best model for lifetime earnings uses above- and below-median for both math and reading.

D.3.3 Statistical Heuristics

Because the expected gains tend to be concave in model complexity whereas measures of variability continue to increase, a simple criterion for model choice is to select the simplest

Figure D.2. Model Selection with Oracle Regret



(a) Marginal Welfare (Math)

(b) Marginal Welfare (Money)

(c) Constrained Welfare (Math)

(d) Constrained Welfare (Money)

model from the equivalence class of models with the highest expected gains.

In practice, we identify equivalence classes by testing the one-sided null hypothesis that the richest model m in the set has lower gains than each simpler model m', and then grouping estimates that are not distinguishable at a given level of significance. Then we select the simplest model in that equivalence set. For example, if for a given level of statistical significance the gains from making assignments using models with 4–9 subgroups were all statistically indistinguishable from the gains from making assignments using 10 subgroups (the richest model we estimate), we would choose the 4-subgroup model since it is the simplest. This process results in achievement quartiles under $[p \leq 0.01]$ and achievement quintiles under $[p \leq 0.05]$.

D.3.4 Hyper-Parameter Validation

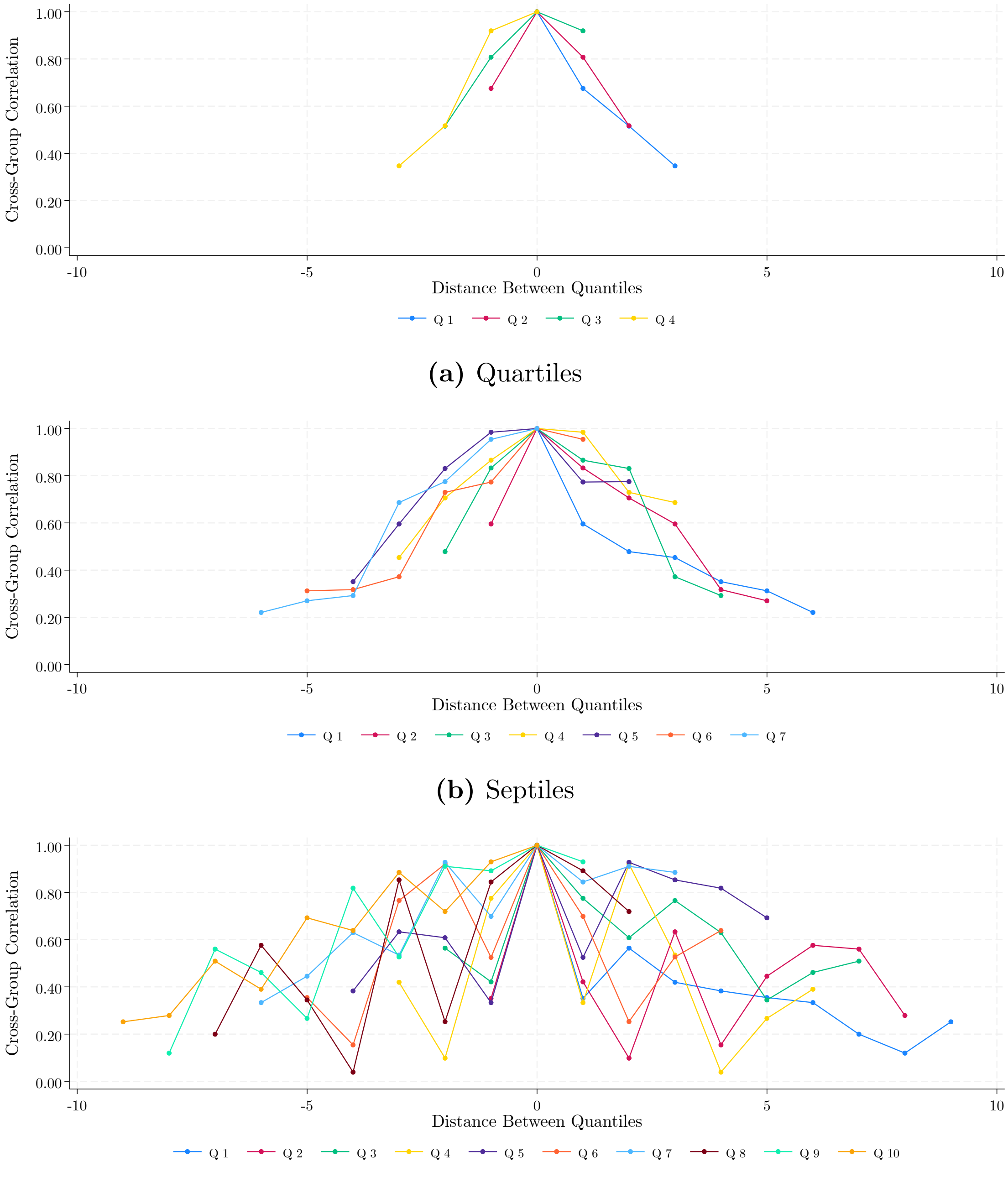
We also use model hyper-parameters to assess model quality. The stability and monotonicity of the estimated hyper-parameters from our Empirical Bayes estimation provides another basis for comparison between models. These hyper-parameters include the correlations and intertemporal (auto)-correlations of class-subgroup residuals. Heuristically, we expect these correlations to be higher for more similar groups and time periods. For example, teacher effects on the ninth and tenth deciles should be more highly correlated than those on the ninth and second deciles. However, because increasing the number of effects to estimate from k to $k+1$ introduces an additional $(k+1) \times (T-1)$ hyper-parameters from the same data, the risks of misestimation increase as we split the sample more finely.[55]

Figure D.3 plots three examples from math heterogeneity by achievement, illustrating intuitive patterns that break down as models become too complex. Panel (a) shows the contemporaneous correlations between subgroup residuals over the distance between quantiles.[56] 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 over 0.9 and the furthest quartiles correlated at just over 0.35. Increasing model complexity complicates these patterns. In Panel (b) of Figure D.3, we observe the cross-group correlations across septiles of prior math achievement. While typically monotonic in the distance between quantiles, these parameters are much less consistent and appear 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.

[55] Because hyper-parameters are identified by teachers who teach students from both subgroups at time t, small sample size bites twice—once through noisier average subgroup residuals, and again through fewer teachers teaching both types of students at time t.

[56] For example, for the $Q1$ line, +1 denotes quartile 2 and +3 denotes quartile 4.

Figure D.3. Hyper-Parameter Stability Decreases with Model Complexity



(a) Quartiles

(b) Septiles

(c) Deciles