Welfare Added?
Optimal Teacher Allocations from Value Added

Tanner S Eastmond* Nathan J Mather† Michael David Ricks‡
Julian Betts§

February 2025

Abstract

This paper considers the allocation of teachers to classes of students and the role that value-added measures of teacher effectiveness can play in optimal policy considerations. After characterizing the trade offs implicit in using heterogeneous, multidimensional measures of teacher value added to approximate welfare, we quantify the importance of these considerations in the allocation of teachers to elementary school classes. Using data from the San Diego Unified School District, we estimate heterogeneity in teacher value added between students with above- and below-median test scores and behavior. Although using heterogeneous estimates can improve outcomes as much as twice as much as allocations based on standard value added estimates, increasing the number of student groups beyond a small number lowers welfare due to incorrect allocations induced by estimation error. Based on external estimates of the labor-market returns, we estimate that optimally allocating teachers in third through fifth grade could increase present-value earnings by $2000 per student. Targeted bonuses paying teachers to be reallocated would have an infinite marginal value of public funds for bonuses up to 27% of baseline pay.

*Department of Economics, Brigham Young University: teastmond@ucsd.edu
†Secretariat: nmather@secretariat-intl.com
‡Department of Economics, University of Nebraska - Lincoln: mricks4@unl.edu
§Department of Economics, University of California, San Diego and National Bureau of Economic Research: jbetts@ucsd.edu

This research is the product of feedback and comments from many people including Ash Craig, Jim Hines, Gordon Dahl, Itzik Fadlon, Lars Lefgren, Peter Hull, Nathan Hendren, Andrew Bacher-Hicks, William-Delgado, Amy Finklestein, Yifan Gong, Andrew Simon, and researchers at the Education Policy Initiative, Youth Policy Lab, and SANDERA as well as with seminar participants at the University of California - San Diego, the University of Michigan, Brigham Young University, CAL Labor Summit, and Boston University. Thanks also to Andy Zau who facilitated the data access and to Wendy Ranck-Buhr, Ron Rode, and others at the San Diego Unified School District for their interest and feedback.

1. Introduction

Every year 50 million students in the United States are assigned public schools teachers who will shape their life trajectories for decades to come. The impacts of these teachers have traditionally been evaluated using "value-added measures," regression-adjusted averages of student scores. Teachers with higher value added not only impact their student's later-life outcomes (Chetty et al., 2014b), but mounting empirical evidence reveals that teacher effects are both *heterogeneous*—e.g., each teacher's impact varies by student race, gender, or socioeconomic status (e.g., Delgado, 2023; Bates et al., 2024)—and *multidimensional*—e.g., each teacher influences cognitive, behavioral, and social outcomes (Jackson, 2018; Petek and Pope, 2023). Although these differences suggest that in theory, value-added information could help allocate teachers by comparative advantage, in practice, there are theoretical, statistical, and computational challenges impeding such allocations.

This paper theoretically and empirically examines the optimal allocation of teachers to classes. We identify two theoretical trade-offs in a simple welfare-theoretic model where a social planner allocates teachers to a set of classes based on teachers' value added (holding class composition constant to focus on teacher allocation). First, if the social planner knew how each teacher would impact each student, the first-best allocation would trade off placing highly effective teachers (absolute advantage) in classes with more students against placing teachers who are relatively better for certain types of students (comparative advantage) in classes with higher shares of these students.[1] The fact that teacher effects are not known and must be estimated introduces a second trade-off regarding how much heterogeneity to consider. Specifically, the second-best allocation also trades off the matching gains from estimating teacher effects with richer heterogeneity against misallocation costs from increasingly imprecise estimates. The relative magnitude of these two trade-offs and the optimal amount of heterogeneity to estimate will depend on the empirical distribution of heterogeneous teacher value-added effects.

We estimate heterogeneity in teacher value-added in the San Diego Unified School District, the second largest district in California. Adapting the methods pioneered by Delgado (2023), we estimate the shrinkage-adjusted value-added of all third- through fifth-grade teachers who taught in SDUSD between 1998-2019. We first focus on math to demonstrate how value-added heterogeneity affect allocations before considering multidimensionality. We initially focus on heterogeneity by prior student achievement, estimating separate value added on students who had above- and below-median test scores the previous year.

[1] For those who would rather not call these absolute and comparative advantage, these concepts could be called comparative advantage by class size and comparative advantage by class composition.

Although these measures of value-added are correlated with standard, homogeneous value-added measures, we find substantial heterogeneity. The average within-teacher difference in value-added across groups (i.e., comparative advantage) is 48% of a standard deviation in mean value-added. We also find these estimates are forecast unbiased and pass a battery of standard robustness checks. We use this and successively more complicated methods to quantify four main results.

Our first main finding is that in this simple case with one outcome and heterogeneous effects across higher- and lower-achieving students, we find large gains from making assignments based on comparative advantage. Because the optimal allocation trades off allocations based on class size and student shares within the classroom, there is no simple decision rule for the optimal allocation. Instead, we characterize the allocation problem as a mixed-integer linear programming problem and compute the optimal teacher allocation given the estimates of each teacher's effects.[2] We find that over the course of third to fifth grade, using heterogeneous measures of value-added to improve district-wide teacher assignments could cumulatively raise average math scores by 0.09 student standard deviations. These reallocative gains are equivalent to the gains students would experience from an improvement in *every* teachers' value-added by 18% of a (teacher) standard deviation. Furthermore, these gains are much larger (39%) than the gains that could be attained using standard value added measures that ignore heterogeneous effects.

Second, we quantify the effects that value-added estimation error has on the the allocation problem. To do this we calculate the expected gains from reallocations with simulated estimates such that teachers do not affect students differently across student types, or across teachers, but value-added is still estimated with error (based on teacher-specific posteriors). Any "gains" from this reallocation will be based on noise proportional to the noise in our actual estimates. This approach quantifies significant bias. The naive estimates predict gains of 0.13 of a student standard deviation over three years. This is 43% above the corrected estimate (0.09). The core issue is that even though the estimates are forecast unbaised on average, individual estimates are estimated with error. As a result, the solution to the mixed-integer linear programming problem makes allocations based on both signal and noise, resulting in a winner's curse. Directly calculating this bias provides an important cautionary note that readers of the fast-growing literature on heterogeneous value added (e.g., Delgado, 2023; Bates et al., 2024; Bau, 2022; Biasi et al., 2021) should be careful not to apply insights from this literature uncritically.[3]These inights apply beyond value added:

[2]In all reallocations the assignment of students to classes is still held constant, as is the grade in which the teacher teaches.

[3]1

Aucejo et al. (2022) and Graham et al. (2023) show heterogeneous teacher effects on other measures of teacher impact as well.

Third, because considering heterogeneity offers potential benefits and costs, we directly by compare allocations made using different dimensions of heterogeneity in the value-added estimation, and find that the above-below median split achieves the second-best allocation for math scores. For this exercise we estimate the naive predicted gains and corrected gains under the optimal allocation using nine different value-added models based on achievement, demographics, and their interactions. Considering additional heterogeneity has significant benefits (especially heterogeneity by achievement), but they are quickly swamped by the efficiency costs of estimation error and common support problems. The policy with the largest corrected gains uses value added on above- and below-median achieving students. Rather than treating this estimation error as an unwelcome side-effect of heterogeneity, our theoretical framework allows us to trade-off these costs with matching gains in search of the feasible second-best allocation.

Finally, returning to the problem of multidimensionality, we consider the welfare-maximizing allocation in general and find that it would produce roughly $40.3 million in present value gains to families. To do so, we estimate heterogeneous effects on math scores, reading scores, attendance, and behavior GPA. Although it is not obvious *ex ante* how to address this multidimensionality, our theory suggests combining outcomes based on how they affect long-term outcomes of interest. By pricing value added gains on these dimensions using causal estimates of teacher (or school) value added on lifetime earnings (from Chetty et al. (2014b), Hanushek et al. (2024), and Lavy and Megalokonomou (2024)), we can essentially optimize over causal changes in earnings. Back-of-the-envelope calculations suggest that the optimal district-wide allocation of teachers would generate $2,07 in present valued earnings per student or $40.3 million in total.[4] Accounting for multidimensionality is quantitatively important. Whereas many education and economic interventions focused on math scores (Chetty et al., 2014a; Delgado, 2023; Bates et al., 2024; Ricks, 2022), we show that accounting for the separate (correlated) marginal effects of math, reading, and behavioral outcomes, generates 34% larger wage impacts (value-added of $5 million) relative to focusing on math alone.

Our paper makes three main contributions. First, we extend the conversation on value added and value added heterogeneity to consider optimal policy. Existing work on value added has generally been empirical. After seminal work on the credibility and import of value added measures (Hanushek, 2011; Chetty et al., 2014a,b) this literature has extended

[4]Here present valuation is discounted at 3% following back to age 10 following Krueger (1999) and Chetty et al. (2014b).

to other outcomes (Jackson, 2018; Petek and Pope, 2023) and heterogeneity cuts (Delgado, 2023). Rather than focusing on estimation, our contribution to this literature is using value added methods to inform extraordinarily complex teacher allocation problems. In this sense, our research is similar to other work that estimates teacher or school value and uses it as an input to answer other economically important questions such as school competition (Bau, 2022), teacher labor markets (Bates et al., 2024), and school choice (Beuermann et al., 2023). By considering how to use estimates of teacher effectiveness to inform public service provision, we also provide a theoretical framework that could apply to a variety of settings where match effects are relevant to an allocation problem, such as the between workers and firms or doctors and patients (Dahlstrand, 2022).

More broadly, our results have implications for the discussion of using value-added when compensating teachers, doctors, hospitals, prosecutors, or others and extend our understanding of the welfare implications of such policies.[5] Motivated by the large predicted earnings gains from reallocations, we explore the welfare implications of using lump-sum transfers to compensate teachers for the possibility of being reallocated. We consider varying sizes of bonus payments to all teachers and find enormous gains measured in the marginal value of public funds (or MVPF (Hendren and Sprung-Keyser, 2020)). The MVPF of bonuses in the district-wide reallocation is infinite for up to $8300 per teacher (roughly 14% of salary for a SDUSD teacher with 10 years of experience). We also consider smaller (and arguably more feasible) within-school-grade reallocations—which have smaller gains but which should be all but costless to teachers—we find that the MVPF is infinite for bonuses of up to $2200. Similarly, smaller, targeted reallocations can achieve 50% of the gains with 20%of the costs. These ideas combine insights from two literatures on teacher labor markets: one focusing on dismissal (Hanushek, 2009; Staiger and Rockoff, 2010; Chetty et al., 2014a) but sometimes ignoring teacher supply decisions (as pointed out in Rothstein, 2010) and the other characterizing teacher demand (Johnson, 2021) but sometimes ignoring teacher impacts on students (as addressed in Bates et al., 2024, where both are combined). Our contribution is characterizing the welfare effects of policies that use teacher value-added but compensate teachers for the possible disutility of the resulting allocation. Similar exercises could apply to value added in any other setting.

Finally, our results also highlight the first-order importance of considering heterogeneity in empirical welfare analyses. Just as with teacher allocations, there may be many settings in

[5]Many individuals and institutions are evaluated with value-added measures. For example, in addition to teachers, consider schools (**?**); doctors, hospitals, and nursing homes (Chandra et al., 2016; Doyle et al., 2019; Hull, 2020; Einav et al., 2022; Chan et al., 2022); and even judges, prosecutors, and defense attorneys (Abrams and Yoon, 2007; Norris, 2019; Harrington and Shaffer, 2023; Landon, 2024). Similar patterns of multidimensionality and heterogeneity have been found in health-related value-added (e.g. Hull, 2020).

which the gains possible from allocations based on heterogeneous effects are much larger than those based on means only. We document this empirically in our setting where considering just one dimension of heterogeneity increases test score gains by 34-97% relative to gains from reallocation that uses only the standard value-added measure. This connection between effect heterogeneity and comparative advantage complements a growing understanding that heterogeneity is a key consideration for allocating scarce resources according to a social objective. This idea of targeting has been explored theoretically in Kitagawa and Tetenov (2018) and Athey and Wager (2021) and is reflected in a recent explosion of empirical inquiry targeting treatments as varied as social safety programs (Alatas et al., 2016; Finkelstein and Notowidigdo, 2019), costly energy efficiency interventions (Ito et al., 2021; Ida et al., 2022), promoting entrepreneurship in developing countries (Hussam et al., 2022), and even resources to reduce gun violence (Bhatt et al., 2023). Our results highlight how ignoring the implications of heterogeneity or estimation error may have serious welfare ramifications and highlight approaches for assessing those trade-offs in practice for targeting-like analyses.

This paper is organized into 6 sections. Section 2 introduces our framework for welfare and value-added with the implications of heterogeneity. Section 3 presents our data, background, and value-added estimation procedure. Section 4 presents optimal allocations using heterogeneity, focusing only on math scores. Section 5 extends the analysis to consider multidimensionality and welfare. Section 6 concludes.

2. Welfare and the Trade-offs of Teacher Allocations

This section considers the optimal policy of allocating teachers to classes including the implications of using traditional value-added measures in welfare analyses. We highlight trade-offs implicit in the first-best policy, characterize the biases that arise from relying on traditional value-added measures, and discuss whether estimating heterogeneous, multidimensional measures of teacher value added can improve allocative efficiency. Although the exposition focuses on teacher value added, the theoretical insights could apply to any setting where a social planner is choosing between policies that have heterogeneous impacts.

2.1 Set-Up and First-Best Policy

Consider a policy maker selecting a policy $\mathcal{J}$ from a set of potential policies $\mathscr{J}$. In our application each possible policy $\mathcal{J} : i \to j$ is an allocation that maps all students to their teachers. We denote the match effect of teacher j on student outcome y_i as $\mu^y_{i,j}$. To focus on the teacher allocation problem, we only consider the subset of policies that hold classes

constant ($\mathscr{J} = \{\mathcal{J} : \mathcal{J}(i) = \mathcal{J}(i') \iff \mathcal{J}'(i) = \mathcal{J}'(i')\}$)[6] given evidence of the important roles of tracking and peer effects in educational settings.[7]

The first challenge with this allocation problem is that the match effects μ only partially order allocations. 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 in all affected classes. This condition will never be met if each teacher has homogeneous effects on all students, and it is unlikely in general given the dispersion of student needs within classes and of teacher effectiveness within the district.

Fortunately, welfare theory allows us to create an strict ordering of allocations. If a social planner can assign each student a welfare weight, $\gamma^{\mathcal{J}}_i$, and use a "score function," $S^{\mathcal{J}}_i = s(\boldsymbol{Y}^{\mathcal{J}}_i, \boldsymbol{X}_i)$, that maps observable outcomes $\boldsymbol{Y}_i$ (like test scores or annual earnings) and characteristics, $\boldsymbol{X}_i$ (like lagged scores or poverty status) into a welfare-relevant scalar, then the welfare function can be written:[8]

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

where and $\mu_{i,j}$ is teacher j's effect on student i's score. In practice the policy maker may just want to maximize average scores $\gamma_i = 1\ \forall i$, but revealed preference suggests that policy makers care particularly about gains to certain students. For example, the federal No Child Left Behind law enacted in 2001 focused on low-achieving students by penalizing schools with subgroups that were not meeting standards.

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

- First, let teacher j's effects be homogeneous, $\mu_{i,j} = \bar{\mu}_j\ \forall i,\ j$. Then first-best allocation would simply rank teachers based on their absolute advantage and would assign the highest ranked teachers to the classes with the largest (welfare-weighted) number of students.[9]

- Second, let teacher j have constant effects on two subgroups of students, $\mu_{i,j} = \bar{\mu}_{k,j}\ \forall i$:

[6] i.e., if any two students i and i' have the same teacher under allocation $\mathcal{J} \in \mathscr{J}$, they must also have the same teacher under allocation $\mathcal{J}' \in \mathscr{J}$.

[7] To the extent to which manipulating either tracking or peer effects could generate additional gains, our optimal allocations will serve to give a lower bound on the true potential social gains.

[8] Appendix B.1 outlines the assumptions necessary for this derivation and discusses the interpretation of the welfare weights γ given a social welfare function based on life time utilities $W = \sum_{i=1}^n \phi^{\mathcal{J}}_i U^{\mathcal{J}}_i$. Under this setup the score function could be thought of as a surrogate index for expected lifetime utility or earnings (see Athey et al., 2019).

[9] Teachers with an absolute advantage in teaching all students therefore have a comparative advantage in teaching larger classes.

$x_i = k$, and let class sizes be identical. Then the first-best allocation would rank the teachers by their comparative advantage at teaching group k and would assign the highest ranked teachers to the classes with the largest (welfare-weighted) share of group k students.

In general when match effects and class sizes vary, the first best allocation trades off the marginal gains from putting teachers with absolutely better skills in larger classes with putting teachers in classes with comparatively better skills in well-matched classes.

2.2 Allocations Using Traditional Teacher Value Added Estimates

The first-best policy is not feasible whenever individual welfare weights and match effects are unobserved. Empirical analyses often simplify the weights and treatment effects to means in order to measure welfare (see the intuition in Hendren and Sprung-Keyser, 2020). In the teacher allocation context, these average effects correspond to traditional measures of teacher "value added" (as in Rivkin et al. (2005) or Chetty et al. (2014a)). A traditional value added estimate, $\hat{\mu}_j$, is typically constructed as the (jackknife-predicted) mean residual of test-score gains for students taught by teacher j. Using these estimates, a policy maker could approximate welfare as

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

where n_j is the number of students in the class to which teacher j is assigned under policy $\mathcal{J}$.

Equation 7 shows how using this approximation of welfare to evaluate allocations is systematically biased (see derivation in Appendix B.2).

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

where $\gamma^{\mathcal{J}}_c$ and $\mu_{c,j}$ represent the average welfare weights of students in teacher $j's$ assigned class and the average match effect of teacher j on those students; $\widehat{\text{Cov}}_j(\cdot)$ reports the estimated within-class covariance for the class teacher j is assigned to under policy $\mathcal{J}$, and $\tilde{\mu}_j$ is the average match effect in the class teacher j actually taught in the estimation sample.

There are four key differences between the true welfare resulting from an allocation and the approximation based on traditional value added estimates. First, focusing on the average treatment effect of each teacher ignores all possible gains from match effects and

comparative advantage. Although this simplification may be benign if teacher effects are highly correlated across students, acknowledging comparative advantage is typically central to efficient economic allocations. As such, it's important to note that estimates based on traditional value added estimates will be biased in general.

Second, using averages ignores the welfare gains from good matches for 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. If teachers truly have homogeneous effects, this covariance will be zero. It will also be zero if the social planner has equal welfare weights on all students or if they happen to be uncorrelated with the match effects.

Third, there is a structural difference between a teacher's average effect, $\bar{\mu}_j$, and the target parameter of a value added estimatnd $\tilde{\mu}_j$. This inconsistency arises because traditional value-added measures target a "treatment on treated" parameter (the effect of the teacher on the students they were actually taught in the data, $\tilde{\mu}_j$) rather than an "average treatment effect" (the effect of the teacher on average, $\bar{\mu}_j$). As such, effects estimated on one class will not be externally valid whenever the distribution of match effects varies from class to class and $\hat{\mu}_j$ will not be consistent for $\bar{\mu}_j$.[10]

Finally, whatever the target parameter of a value added estimate, estimation error will lead to differences in finite samples. As a result, allocations using value added estimates $\hat{\mu}_j$ rather than true (average) match affects will result in smaller than anticipated gains because some of the presumed absolute or comparative advantage will just be noise. This "winner's curse" type problem will typically become more serious the lower the number of students per estimation class, n_j, the higher the variation in match effects within each class, and (if jacknifing estimates) the fewer years each teacher is observed.

Because of the differences between welfare and the constant-effects approximation, allocations based on traditional value added will not attain the first best. If there is enough variation across teachers in mean value added, the allocation with the highest approximated welfare will likely produce gains relative to the status quo, but it will not maximize welfare in general. Panel (a) of Figure 1 illustrates these twin insights with a simple model where all heterogeneity in effects and weights occurs across two types of students. Gains to students of each type (relative to a status-quo policy) are plotted on a production possibility frontier. Using only mean effects, welfare increases from $\mathcal{W}^0$ to $\mathcal{W}^*_{VA}$, but much larger gains could be attained under the first-best $\mathcal{W}^*$.[11] Figure 1 highlights the potential for allocative gains

[10] Any non-random sorting of students into classes will produce this problem, but even under random assignment, finite-sample randomization error will typically lead to idiosyncratic variations in match effects across classes.

[11] Note that the welfare-maximizing allocations do not necessarily maximize average scores because the

from estimating heterogeneous effects in a simple setting, but whether considering additional dimensions of heterogeneity is optimal in practice may in fact be a quantitative question.

2.3 Allocations Using Heterogeneous Estimates

Instead of relying on constant-effects models where each teacher has the same impact on all students, emerging research has identified what anecdote has long known: that teachers do in fact have heterogeneous effects. For example, teacher value added varies across student subgroups of race (Delgado, 2023), poverty (Bates et al., 2024), and achievement (Biasi et al., 2021), 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 $\mathcal{K}$). Using these estimates, a policy maker could produce a new approximation of welfare based on these conditional average treatment effects.

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

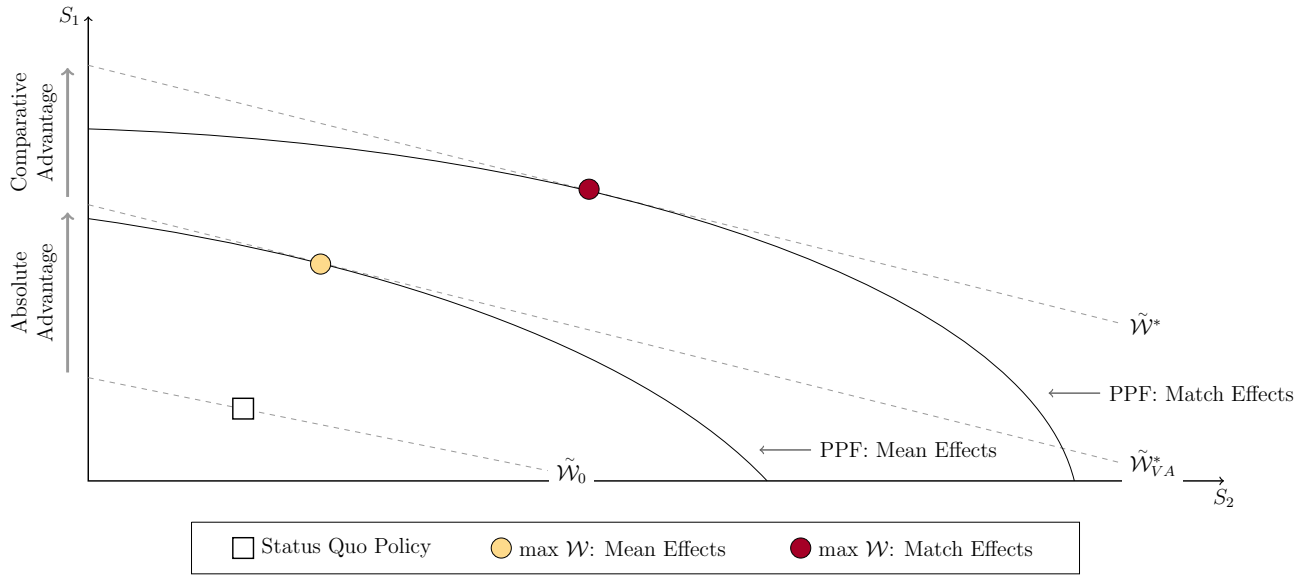
where $n_{k,j}$ represents the number of type-k students in teacher j's class. This new approximation presents several benefits relative to one based on traditional value-added measures, but it also presents some new challenges. We consider the four components of bias from equation 7 in turn.

First, using subgroup-specific value added estimates will always reduce the first bias term related to ignored matching gains. These gains occur because estimating subgroup-specific effects increases the scope for matching based on comparative advantage. In the limit, the bias in the first term goes to zero as the subgroup effects and subgroup weight capture all of the welfare-relevant information. These are the gains illustrated in panel (a) of Figure 1 and are made possible by allocating teachers based on subgroup effects rather than mean effects. Although in contrast to Figure 1, these gains will not approach the first-best until match effects are homogeneous within each type k.

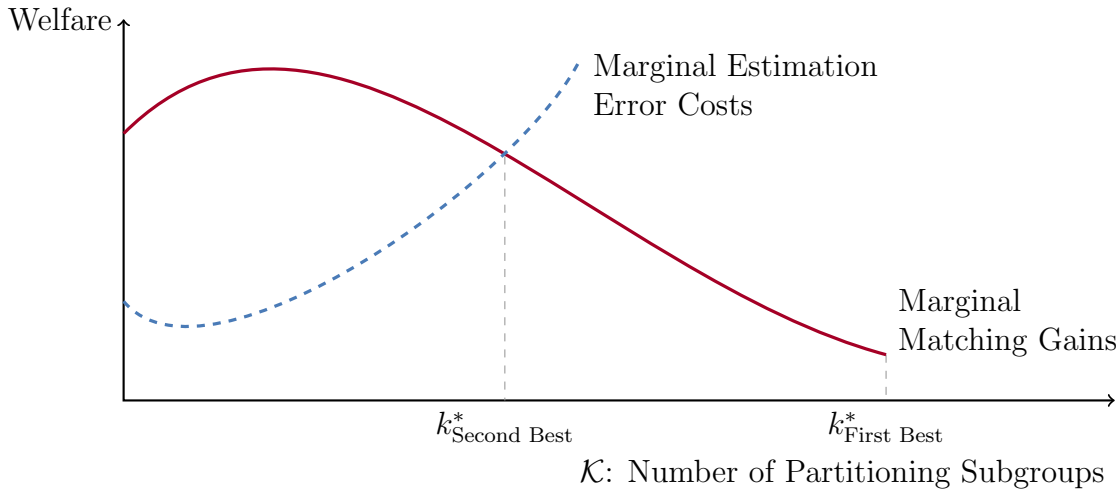
Second, using more subgroup-specific effects will tend to reduce the second bias term related to the covariance of match effects and welfare weights. In the limit, the subgroups will devolve into single-student cells, and the within-subgroup covariance will be zero. In practice, however, the covariance may increase or decrease as additional dimensions are added along the way. Because reducing the within-subgroup variability in match effects or weights will reduce the bias, some partitions are likely more useful than others. For example, effects and weights are much more likely to vary between student with more and less academic

policy maker may value gains to type-one students more than gains to type-two students. This is represented by the sloped indifference curves at $\mathcal{W}^0$, $\mathcal{W}^*_{VA}$, and $\mathcal{W}^*$.

Figure 1: True Match Effects and Estimation Error Both Affect the Best Feasible Allocation



(a) Absolute Advantage, Comparative Advantage, and Social Preferences in the First-Best



(b) Trading off Match Effects and Estimation Error in the Second-Best

Note: This figure illustrates the welfare gains from using value-added information to make teacher allocations. Panel (a) presents a stylized depiction of the benefits of considering heterogeneity. The two axes present the outcome score of interest, S, for individuals of two types. The graph contains two production possibility frontiers and three iso-welfare curves. The interior production possibility frontier is attained by reallocating based off of average match effects (like traditional value-added measures). These mean estimates enable welfare gains from reallocating based on (welfare weighted) absolute advantage. The second, dominant frontier is attained by considering effect heterogeneity and, thus, comparative advantage. The indifference curves show the welfare value of the status quo, (2) the welfare maximizing allocation using mean effects, and the welfare maximizing allocation using heterogeneous effects. Panel (b) extends this intuition to depict not only the benefits of considering heterogeneity, but also the costs. Here the horizontal axis depicts the number of subgroup-specific value added estimates the social planner uses to make allocations. The figure depicts both the marginal benefits from considering richer heterogeneity (as visualized in panel (a), but also the increasing costs due to estimation error. In the first best, match effects are known and should all be used, but in the second best, the social planner must trade off the matching gains against the costs of estimation error.

preparation than between students with earlier or later birthdays within a given month. Using high-variance partitions will reduce the bias in this term.

Third, using subgroup-specific value-added estimates has a limited potential to reduce the third bias term related to external validity. The potential for gains comes from using subgroup-specific effects to make more accurate comparisons between classes with different compositions (i.e., improving the external validity). 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 the bias). The potential to reduce bias is limited, however. The key challenge is that using subgroup-effects to improve external validity requires common support, and finer partitions are more likely to create empty cells from which no information can be extracted. For example, while estimating separate effects for each teachers on above- and below-median students may be feasible, using 100 percentile groups would be infeasible because too few teachers teach students from every single percentile of the achievement distribution each year.

Finally, using subgroup-specific effects will typically increase the last bias term related to estimation error. Because estimation error is related to the variance of scores and the number of students taught, splitting the sample to estimate subgroup-specific effects will tend to increase estimation error.[12] This estimation error poses significant challenges to the allocation problem because making assignments using a noisy estimate of comparative advantage may not actually result in any welfare gains. As a result, the optimal extent of heterogeneity to model is an empirical question. While shrinkage-correcting value-added measures may help reduce the variability of estimates, estimation error will be a problem in general unless the generative model is perfectly specified.[13]

Because estimating additional heterogeneity to inform the allocation problem has both costs and benefits, the optimal set of effects to utilize is a quantitiative question. Panel (b) of Figure **??** visualizes this tradeoff. It presents the marginal welfare costs and benefits from making allocations based off of teacher value-added on increasingly fine partitions of students. Obviously the first best allocation would use the information from the limiting partition if those match effects were known, but in practice, the social planner should continue estimating more subgroup-specific effects until the marginal benefits no longer outweigh the marginal costs. Using value added estimates based off of this last partition is how to achieve the

[12] It will do so whenever the within-subgroup variance is comparable to the overall variance (and as long as the conditional variance of the residuals is not decreasing faster than $\sqrt{\frac{n}{k}}$ across finer partitions).

[13] In practice the most common methods for shrinkage use *ad hoc* jacknife-prediction style shrinkage which is not likely to satisfy this condition.

feasible second-best policy.

These trade-offs highlight two rules of thumb for finding the partition which will generate the second-best allocation. First, when comparing two similar partitions (for example above and below median achievement or low-income and higher-income families), better partitions will typically capture greater variation in weights and match effects while maintaining common support for all teachers. For example, while there may be heterogeneity in weights and match effects by both achievement or race, they may be much less common support across classes for race. Second, when comparing a coarser partition with a slightly finer partition (for example income terciles versus quartiles or income versus income-by-sex), better partitions will capture variation in weights and match effects without producing small or empty cells for most teachers.

The extent to which these considerations bias welfare approximation and the importance of each trade-off will depend on the empirical distribution of teacher value added. To quantify the benefits and costs of estimating heterogeneous, multidimensional effects, we need to estimate different measures of value added. Only then can we consider reallocations and determine the optimal (second-best) allocation of teachers to classes.

3. Estimating Heterogeneous Value-added for Teachers in San Diego Unified

This section sets the groundwork for estimating heterogeneity in teacher value-added and whether it matters for the allocations of teachers to elementary school classes. To that end, we describe the data from the San Diego Unified School District, present our estimation strategy for value-added, and summarize patterns in value-added—including the extent of comparative advantage and its distribution across classes in the status quo.

3.1 Background and Administrative Data

We use administrative data on the universe of students attending schools in the San Diego Unified School District (SDUSD). For our main analyses we focus on 6,001 teachers who are the main instructors in third, fourth, or fifth grade classes in the 1998-1999 through 2018-19 school years.[14] We link all teachers to their students each year and we restrict our attention to student-year observations with math, reading, behavior, or attendance outcome in two consecutive years. This leaves us with 756,281 student-year observations in 36,053 class-year groups. The administrative data also contain relevant information about student demographics and academics as well as long-term outcomes. We provide more descriptive

[14]Because there was no state exam in 2014 we cannot estimate value added in 2013-14 or 2014-15 school years.

statistics and information about the current allocation of teachers to classes in Appendix E.

3.2 Estimation, Identification, and Validation

We use the data from San Diego Unified to evaluate the importance of heterogeneity in optimal policy. While we consider additional dimensions over which we could estimate heterogeneous effects in Section 4, we frame the exposition of our estimation approach in the context of lagged student outcomes. To begin we will focus on students with above- or below median outcomes in the prior year. We consider four multidimensional outcomes, math scores, reading scores, behavior GPA, and attendance. For each outcome we estimate each teacher's value-added on the of students with below-median outcomes (e.g., lower-scoring students) and above-median outcomes (e.g., higher-scoring students).

Our general appoach to estimating value added consists of three steps. The details are presented in Appendix C, and we follow similar estimation approaches whether estimating average or subgroup-specific teacher value added. First, for each subgroup k we estimate the effects of student i's characteristics in year t, $X_{i,t}$, on outcome s, $S_{i,s,t}$ (measured in student standard deviation units), in a regression of the form:

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

where $\alpha_{\mathcal{J}(i,t),s,k}$ and $\phi_{\mathcal{S}(i,t),s,k}$ are subgroup-specific teacher and school fixed effects, and X includes student-, class, and school-level controls like demographics and cubics of lagged achievement.[15]

Second, for each subgroup we obtain the average of the residuals implied by $(\boldsymbol{\alpha}_{s,k}, \boldsymbol{\beta}_{s,k}, \boldsymbol{\phi}_{s,k})$ by class and year:

$$\bar{A}_{j,s,k,t} = \frac{1}{n_{j,k,t}} \sum_{i:\mathcal{J}(i,t)=j,k_i=k} \left[S_{i,s,t} - \hat{S}_{i,s,t} \right]$$

Finally, we estimate leave-year-out (jackknife) measures of teacher impact by predicting $\bar{\boldsymbol{A}}_{s,k}$ with the residuals in all other years.

$$\hat{\mu}^{VA}_{j,t,s,k} = \hat{\boldsymbol{\psi}}_{s,k} \bar{\boldsymbol{A}}_{j,-t,s,k} \tag{3}$$

This prediction step shrinks each value added score toward the value added in other years

[15] For most regressions we use outcomes in year t and lagged scores from year $t-1$, but following Petek and Pope (2023) we use behavior GPAs from year $t+1$ and lagged outcomes from year $t-1$. Because behavior in year t is reported directly by the teacher (rather than being centrally assessed) it is subject to measurement error from idiosyncratic subjectivity (e.g., a teacher who rates all kids highly would appear to have a high value added if relying on year t outcomes.

while allowing drift over time. Because the distribution in teacher value-added must be stationary across time, we also demean $\hat{\mu}$ within each grade-year-type cell in our results (see details in Appendix C).

We implement these steps following Chetty et al. (2014a) and (Stepner, 2013) for each subgroup, noting that this procedure is the special case highlighted in Delgado (2023) where information about type-k' value added is not used to predict type-k value added. We make this assumption regarding heterogeneity to ensure internal consistency with how we treat multi-dimensionality. The maintained assumption in Chetty et al. (2014a) and Delgado (2023) regarding multi-dimensionality is that there is no covariance between the specific idiosyncratic components of teacher value-added across outcome domains. For consistency we also ignore any covariance between the type-specific idiosyncratic components of teacher value-added on each outcome in the shrinkage step.

Interpreting our estimates as causal effects requires that individual- and class-type-level shocks are conditionally independent from teacher assignment. Students certainly are not randomly sorted to teachers, but rich information about lagged outcomes at the student, class, and school level seem to capture the key variation in unobserved determinants of outcomes (as confirmed in in Chetty et al. (2014a)). This is why we flexibly control for grade-specific cubics in lagged student achievement in all outcomes, school fixed effects, and the cubics of class and school averages of all outcomes. Our other controls include ethnicity, reported sex, age, past special education, past English-language-learner status, class means of these demographics, class size, grade indicators, and year indicators.[16] Note that including school fixed effects implies that our model is identified based off quasi-experimental changes in test scores when teachers switch schools.

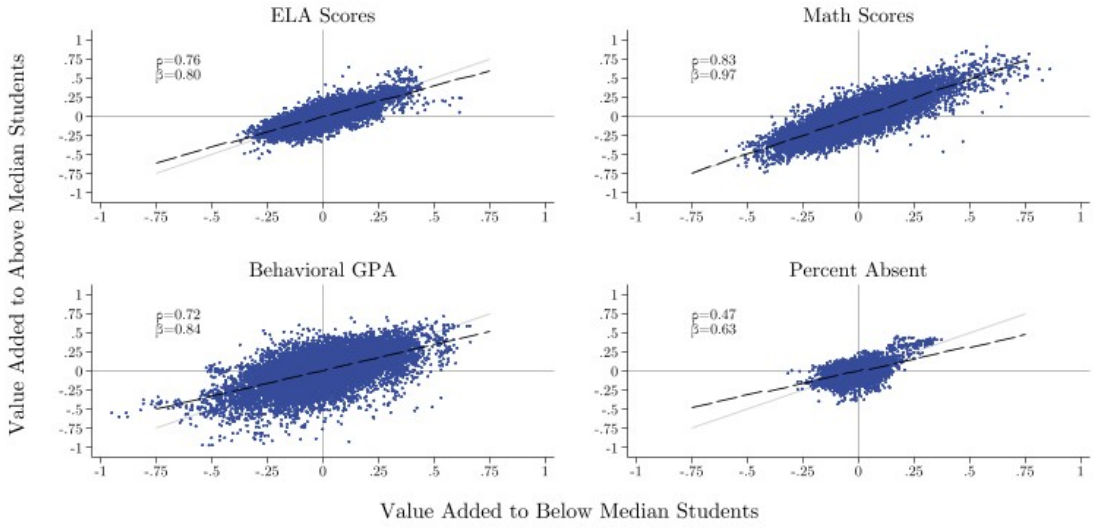
To assess the plausibility of these identifying assumptions and their implications for the external validity of our estimates to counterfactual simulations, we present additional checks on estimate validity. First, for each outcome and student type we document precise forecast unbiasedness. Appendix Figures A.2 plots the average residuals in classes taught by each teacher over ventiles of subgroup-specific value added estimates predicted based on other years. Each relationships has a slope of almost exactly 1 with tight standard errors (cluster corrected at the teacher level). Second, Appendix Figure D.10 reports high patterns of persistence over time. For example, over 40% of teachers have a comparative advantage for teaching one group of students in *all* years, and the year-to-year correlation in comparative advantage is between 0.78-0.90 for all estimates.

[16] In addition to the school fixed effects, the other difference between our controls and those in Chetty et al. (2014a) is that they include free and reduced price lunch, which we omit based on SDUSD's interpretation of federal regulations regarding the use this information.

3.3 Heterogeneity Highlights the Importance of Comparative Advantage

Using this procedure, we estimate the heterogeneous effects of 6,000 teachers and depict the main value-added results in Figure 2. This Figure reports a scatter plot of each outcome (math scores, reading scores, behavior GPA, and attendance). Each point represents one teacher-year observation where their value-added on higher-scoring students is plotted on the y-axis over their value-added on lower-scoring students on the x-axis. Each plot also presents the correlation coefficient between the value-added on the two student groups as well as a slope coefficient for the line of best fit between the two.

Figure 2: Value-added varies significantly within and across teachers



Note: This figure shows our heterogeneous estimates of teacher value-added on both reading scores, math scores, behavior GPA, and attendance. Each dot represents one teacher-year estimate of value-added on students with above- or below-median lagged outcomes. Each correlation coefficients, 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 line with slope one is plotted in the background.

Figure 2 depicts substantial differences within *and* across teachers, suggesting we should reject the standard "constant effects" model of value in favor of one with appreciable comparative advantage. Differences across teachers, or absolute advantage, can be seen in the broad dispersion of teachers along the 45-degree line. Teachers above and to the right generate larger testing gains compared to teachers below and to the left. It is also interesting to note that value-added to math is much more dispersed than value-added to reading (consistent with evidence from (e.g., Chetty et al., 2014a)), but both are less dispersed than effects on behavior GPA and more dispersed than effects on attendance.

Comparative advantage can also be seen visually. Teachers with dots above the gray

45-degree line have a comparative advantage in teaching students with above-average lagged outcomes, and teachers with dots below that line have a comparative advantage in teaching those with below-average outcomes. The the average comparative advantage is large: 53% the size of the cross-teacher standard deviation in standard teacher value-added for reading and 48% for math. By bootstrapping our estimation procedure 1000 times we also reject the null of constant effects for 80% of teachers at the [p=0.10] level or higher. Interestingly, teachers' value-added is more highly correlated across achievement groups for math than for reading. This is also consistent with absolute advantage being more important and variable with Math teaching than with reading teaching. Value added less correlated for attendance, which seems to be because of ceiling affects on students in the above-average attendance subgroup.

These differences within and between teachers highlight the potential for gains from reallocation. We estimate that teacher value-added to higher- and lower-scoring students is correlated at 0.76 for reading and 0.83 for Math. The fact that this correlation is less than one allows for gains from allocating teachers by comparative advantage. Even though the correlations are high, there are still significant margins for gains. For comparison, our cross-group correlations are lower than those by socioeconomic status (0.9 for math in Bates et al., 2024) but larger than those by race (0.7 for math and 0.4 for reading in Delgado, 2023). Additionally, to consider the multidimensionality, we note that cross-subject correlations are fairly low. For example, Figure A.1 shows that the cross-subject, cross-group correlation in math and reading are both around 0.6.

4. Efficiently Allocating Teachers to Classes

With estimates of the heterogeneous teacher effects, we use our theory to consider the allocation teachers to classes. This section defines the allocation problem and presents the predicted gains under the naive optimal allocations. Then it quantifies the bias attributable to estimation error and compares the true expected gains from using different models of value-added. To focus on the trade-offs related to heterogeneity, this section focuses only on math scores, but multidimensionality will be added back in in Section 5.

To build intuition consider a social objective $\widetilde{\mathcal{W}}$ focusing on higher- and lower- scoring students. Let $\mathcal{J} : (i,t) \to j$ be an allocation function, telling us which teachers teach each student in each year. We define the following optimization problem for weighted test score gains in a given subject (s, subject subscripts suppressed as we begin with math):

$$\max_{\mathcal{J}\in\mathscr{J}} \widetilde{\mathcal{W}}(\mathcal{J};\omega) = \max_{\mathcal{J}\in\mathscr{J}} \frac{1}{N_{i,t}} \sum_{(i,t)} \omega_L \, L_{i,t} \, \hat{\tau}_L^{\mathcal{J}(i,t)} + (1-\omega_L)\,(1-L_{i,t})\,\hat{\tau}_H^{\mathcal{J}(i,t)} \tag{4}$$

where $\omega_L \in [0.0, 1.0]$ represents the weight on lower-scoring students in the social objective, $L_{i,t}$ is an indicator for whether student i is lower-scoring, and $\hat{\tau}_H^j$ and $\hat{\tau}_L^j$ are our estimates of heterogeneous value-added. Analogous functions can be defined with more granular welfare weights and estimated effects or using traditional value-added measures.

The optimal policy is chosen from the set $\mathscr{J}$ of feasible allocations. In our setting, we focus only on reallocating teachers to existing classes in the grade they actually taught without changing the composition of those classes in any way. We do this to avoid introducing peer-effect biases into our welfare estimates. The single-ω parameterization of welfare imposes linear indifference curves that trade off performance for lower- and higher-scoring students where the weight on each group reflects the degree to which the social planner wishes to target gains to one group of students relative to the other. It also assumes that the social planner only values gains to students in the given subject—both assumptions we relax in Section 5.

This allocation problem captures three distinct trade-offs that have been mentioned in the value-added literature but never fully addressed together. First, the optimal allocation must account for the *comparative advantage* of teachers because of differences in *class composition* (as pointed out in Delgado, 2023). Second, the optimal allocation must also account for the *absolute advantage* of teachers because of differences in *class size*. This crucial detail has been accounted for at the school level (see Bates et al., 2024), but class size and class composition vary both across *and* within schools. Because of these differences, we are interested in both within-school and district-wide reallocation exercises. Finally, the optimal allocation must account for possible heterogeneity in the social value of gains to different types of students—something unique to our paper.

To begin, we solve this allocation problem for two sets of possible reallocations: within-school and district-wide. For the within-school reallocation we restrict $\mathscr{J}$ to require that teachers do not change schools. This within-school problem can be solved easily by iterating over finite combinations within school-grade(-year) cells, but the district-wide reallocation problem has over 8×10^{473} allocations to search over each year. Because the optimal policy depends on both absolute and comparative advantage when both class sizes and class compositions vary, this problem cannot be solved by simply assigning teachers to classes with large shares of students they have a comparative advantage in teaching or simply assigning the best teachers to the largest classes. We characterize equation 4 as a mixed-integer linear programming problem and solved using the COIN-OR Branch and Cut solver implemented by the Python library Gurobi (Gurobi OptimizationLLC, 2024).

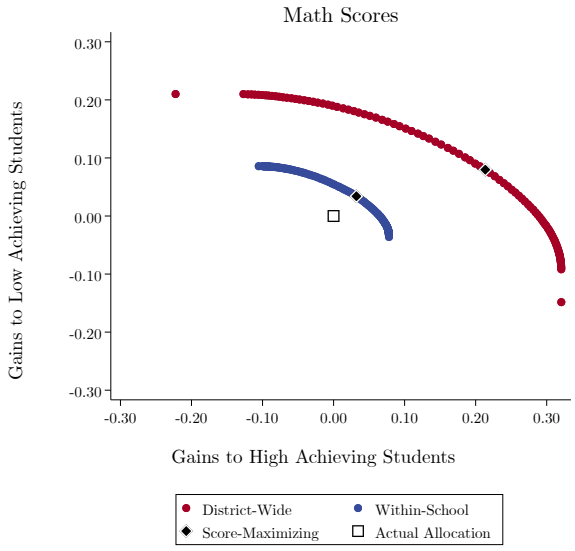
Before proceeding, we want to note three caveats in considering these reallocations. First, note that because we do not change class composition, these gains could be significantly

larger in a district that employs class-level tracking because of greater variance in class composition. Second, the district-wide reallocations might be infeasible. For example, in SDUSD the union contract gives teachers with seniority higher priority in hiring. Teachers also have strong preferences over locations (Boyd et al., 2005a) and schools (Bates et al., 2024) that could make some allocations from being incentive incompatible. Third, the new allocations must be interpreted in the light of partial equilibrium, barring families re-sorting to classes (via requests), schools (via school choice), or districts (via in- or out-mobility).

4.1 Allocations Incorporating Heterogeneous Impacts Increase Test Scores

We create a production-possibility frontier (PPF) for the predicted gains to each group from the within-school and district-wide reallocations. To do this, we solve the optimization problem in Equation 4 for 101 different values of the social weights ω_L ranging from 0.0 to 1.0. We then recover the average value-added received by lower- and higher-scoring students and calculate the gain beyond the status quo. By comparing the optimal allocations under different weights, this analysis characterizes the trade-offs between gains gains to lower- and higher-scoring students, illustrated in the PPFs.

Figure 3: Optimal Allocations Can Create Large Gains to High- and Low-scoring Students



Note: This figure shows the predicted test score gains from optimal allocations relative to the status quo. Two production possibility frontiers are presented, one for reallocating teachers within school-grade cells and one reallocating teachers across schools (still within grade). Each PPF is constructed by finding the optimal allocation by linear programming given relative weights on lower- and higher-scoring students [0.0,1.0]. Gains are reported as average changes in scores measured in student standard deviations over the third through fifth grade.

We depict these production-possibility frontiers for math and reading in Figure 3. We plot the PPF for change in reading scores on the left and Math scores on the right. Each point presents the average one-year change in lower-scoring students' test scores in the optimal allocations (on the y-axis) over average change for higher-scoring students (on the x-axis), all relative to the status quo (noted with the square marker). Allocations that would reduce a group's average scores relative to the status quo are denoted with negative numbers. Allocations above and/or to the right of the status quo are preferred by the social planner. The lighter (blue) PPF denotes the within-school reallocations and the darker (red) PPF the district-wide reallocations. Unsurprisingly, the district-wide reallocations produce gains that are further out in both dimensions.

Figure 3 reveals large predicted gains from reallocation. For example, in the district-wide reallocation a social planner seeking to raise average scores (i.e., $\omega_L = \omega_H = 0.5$) could increase lower-scoring student scores by 0.07 student standard deviations and higher-scoring students' scores by 0.20 student standard deviations over the course of third to fifth grade. The simpler within-school reallocation could raise reading and math scores for both groups by just over 0.04 standard deviations. These are large gains—equivalent to the gains that would result from improving the value-added of *every teacher* in the district by one teacher standard deviation (but retaining status quo assignments) for one year, and triple the gains from proposed teacher screening programs that "deselect" (i.e., fire) teachers with the lowest 5% standard value-added (as considered in Hanushek, 2009, 2011; Chetty et al., 2014b).

The curvature of the PPFs also demonstrates the value in explicitly considering the distributional goals of a policymaker. We consider extensive policy focus on low-achieving students a revealed-preference argument that the social planner may not prefer the test-score maximizing allocation. In this case a policy maker may value gains to some students more than others. These gains are dependent on the extent to which distributional goals deviate from the mean scores objective but are large for more extreme distributional goals. For example, a reallocation could raise lower-achieving students' predicted scores by about 0.2 standard deviations without harming higher-achieving students on average or help higher-achieving students by about 0.3 without harming lower-achieving students on average. To quantify these welfare gains, Appendix Figure A.4 presents the difference in an "Atkinson index" measure[17] for each allocation on the comparative advantage frontier, showing larger welfare impacts when the social planner favors one group more or less.

Comparison with Traditional Value Added. While reallocations predict large gains from teacher reallocations, some of these gains would be possible without knowing hetero-

[17]This measure reports an increase in average scores that would make the social planner would be indifferent with the test-score maximizing allocation.

geneous teacher effects. To evaluate the importance of heterogeneity, we compare the allocations made based off of standard (absolute advantage) and heterogeneous estimates of value-added (absolute and comaprative). In practice, we implement this by replacing the heterogeneous estimates in Equation 4 withe standard value-added scores and re-solving the allocation problem. To build intuition we consider only teachers who have both standard and heterogeneous value-added scores and we compute true gains using the (forecast-unbiased) heterogeneous effects because standard estimates may fail to predict the effect of sending teachers to very different classes.[18]

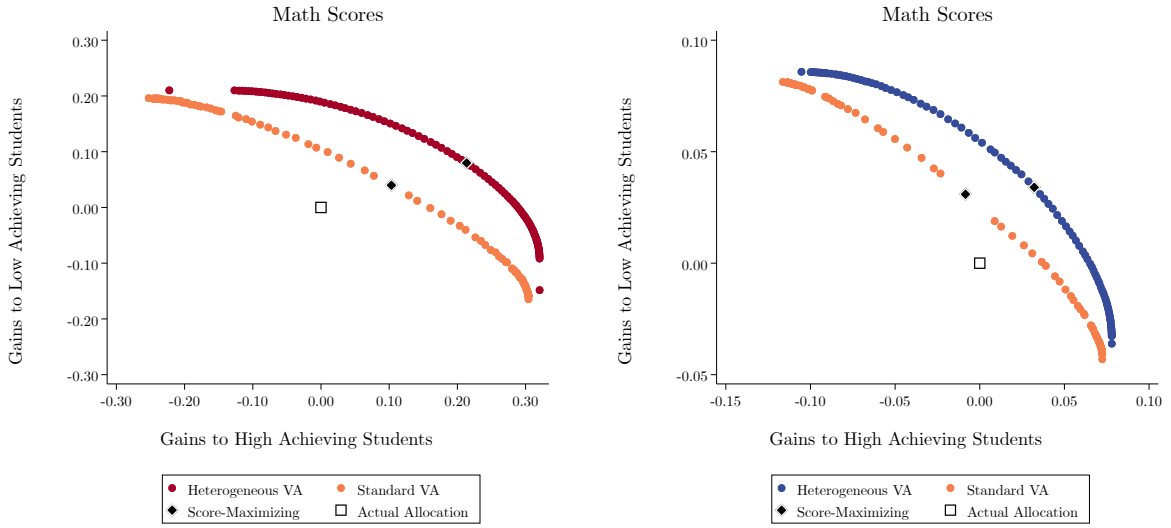
Making allocations based off of absolute advantage does raise predicted test scores, but not nearly as much as using the heterogeneous estimates to account for comparative advantage. Figure 4 depicts the production-possibility frontiers for group achievement gains using each value-added model. Again each point presents the average change in lower-scoring students' test scores in the optimal allocations (on the y-axis) over average change for higher-scoring students (on the x-axis) relative to the status quo (noted with the square marker). Panel (a) presents the results from the district-wide reallocation, Panel (b) presents those from the within-school reallocation. These figures also mark the allocations that maximize test scores with a black diamond for reference. These results are analogous to the theoretical depiction in Figure 1.

Comparing the optimal allocations reveals that using information about comparative advantage can as much as double the achievement gains from reallocations. In the district-wide reallocation, allocations using comparative advantage improve math scores 122.7% more than allocations using only absolute advantage. These are large gains: 0.07 standard deviations in math for all students in the district would be an impressive policy victory. The differences are both economically and statistically meaningful. Using a 1000 sets of bootstrapped estimates, we calculate the variability of predicted gains around the optimum identified by our test-score maximizing reallocation. We find that the asymmetric 95% confidence intervals of [0.130,0.137] for math. Not one of the bootstrapped replications predicts gains at the optimum below those experienced at the allocation based on traditional value added.

Consider four interesting implications of these results for our understanding of teacher allocations. First, we find that modeling heterogeneous effects substantially increases predicted test-score gains. This should not be surprising given our strong intuition about the importance of comparative advantage. Second, however, predicted test-score gains may not accurately reflect the true gains from a reallocation. Since each teacher effect is measured

[18]In this sense the comparison reports the predicted matching and external validity gains without (yet) considering any losses from empty cells or estimation error.

Figure 4: Using Heterogeneous Estimates Produces Larger Gains from Reallocation



(a) District-Wide Reallocation (b) Within-School Reallocation

Note: This figure shows the test score gains from optimal allocations relative to the status quo. In each panel two production possibility frontiers are presented, one for reallocating teachers based on our estimates of value-added (absolute and comparative advantage) and one reallocating teachers only based on traditional value-added (absolute advantage). Panel (a) displays the result for reallocating teachers across schools and panel (b) the results for reallocating teachers within schools (both always keep teacher in the same grade). Each PPF is constructed by finding the optimal allocation given relative weights on low- and high-scoring students [0.0,1.0] by solving the optimal mixed-integer linear programming problem. Gains are reported as average changes in scores measured in student standard deviations per school year that the reallocation is performed.

with error, the optimized allocation risks generating a winners curse by making assignments based off of noise rather than signal. Shrinkage will help with but not alleviate this issue, so we address it in the following subsection. Third, the predicted test-score gains to each group are different, with higher-achieving students standing to gain more from reallocations. This asymmetry highlights the welfare implications of structural features of the education system such as the fact that higher-scoring students tend to be in larger classes. We consider implications to equity in Section 4.3. Finally, because teachers have distinct value added on different outcomes, allocations that optimize math scores, reading scores, and behavioral outcomes will each be distinct.[19] As such, the gains highlighted in papers that do reallocations using one subject at a time like Delgado (2023) and Bates et al. (2024) only give a lower bound to the gains from using information from multidimensional teacher output. This will motivate our analyses in Section 5 where we aggregate gains over multidimensional

[19]That said, for one-subject reallocations like Bates et al. (2024), it is better to focus on math because of the variance in teacher value-added is larger, generating larger predicted gains from reallocations to math than to reading.

outcomes.

4.2 Quantifying Estimation Error to Identify the (Second-Best Allocations

To this point we have focused on one partition of students—heterogeneity by above- and below- median lagged outcomes. This comparison allows us to consider the trade-off between making allocations based off of absolute and comparative advantage, but ignores the trade-off between estimating richer heterogeneity and introducing estimation error. To compare possible partitions and identify the second-best allocation we outline our approach for quantifying each component of welfare bias then compare the results of many different value added models.

4.2.1 Quantifying Welfare Biases

In Section 2 we outlined four biases that can arise when trying to make teacher allocations using value added scores and how each will be affected by estimating teacher effects with additional degrees of heterogeneity. This section decomposes our predicted test-score gains into these different components to separate true matching gains from misallocation based on estimation error and empty cells.

- **(Naive) Predicted Gains.** Allocations based off of value-added measures take them as the ground truth and optimize over those predicted gains. These transparent predictions are what the optimization suggest should occur and are reported in Section 3.1. This (admittedly naive) assumption ignores the other sources of bias in Equation 7 that could result in welfare losses.

- **Empty Cells.** Estimating additional degrees of heterogeneity imposes stronger and stronger requirements about common support. In practice this means that models with richer heterogeneity will identify effects on smaller and smaller subsets of teachers. We attempt to bound the welfare losses from this support problem using the following exercise. First, after estimating an additional level of heterogeneity, we identify the teachers who can no longer be reallocated.[20] Then we re-estimate the predicted gains using the simpler value-added model first using the full set of teachers and then restricting to teachers who have effects in the more finely partitioned model. The losses from excluding these teachers in the simple model offer an intuitive bound on the losses from excluding them in the richer model.

[20] At least, they cannot be reallocated without making strong assumptions. For example, using predictions based on the known relationship between effects for other teachers mechanically imposes that there is no comparative advantage anyway.

- **Estimation Error.** We account for the role of estimation error using a placebo reallocation in four steps. First, we use our 1000 sets of bootstrapped estimates to characterize the joint distribution of (posterior) value-added scores. Second, assuming that all teachers truly have no effect on students, we draw noisy "estimates" around zero based on the bootstrapped variance of each teacher's effects. Third, we re-solve the allocation problem. As researchers we know there should be no gains, but this placebo reallocation will predict gains because of the estimation error. Because the amount of estimation error is directly based on the data, the size of predicted gains in the placebo quantifies the expected losses from over-fitting.

- **(True) Expected Gains.** To recover the gains a policymaker should actually anticipate based off of simulated re-allocations, we take the predicted gains (which already account for losses from empty cells) and subtract the predicted gains from the placebo reallocation. This difference represent the sophisticated correction to the naive predicted gains. In the context of Equation 7 it captures the combined gains from match effects and improved external validity.

The combination of these approaches will allow us to separately report the true expected gains from reallocations (net of losses caused by modeling decisions) and to identify the second best policy.

To make this concrete, consider the comparison between the standard value added and heterogeneous value added by two student groups - above and below median math achievement. The (naive) predicted gains from the allocation based off of heterogeneous effects were about 0.13 standard deviations. This is 0.04 standard deviations greater than the naive predicted gains from the allocation based off of the standard value added holding the sample constant. We estimate the losses due to empty cells at around 0.01 standard deviations. We find that estimation error accounts for 0.04 out of 0.13 standard deviations of the gains in the heterogeneous value added case and 0.02 out of 0.08 standard deviations of the gains in the standard value added case. In other words the true expected gains from a reallocation using the standard value added are 0.07 standard deviations, compared to 0.09 using the heterogeneous value added. This comparison clarifies that although it is unambiguously better to use information based on heterogeneity, a sophisticated social planner would expect smaller gains than the naive estimates would suggest.

4.2.2 Identifying the Second-Best Allocation

The estimates for these different quantities under various student groupings are displayed in Figure 5. In panel (a), we show the gains and predicted losses when moving from Standard

VA measures to 2 bins and ultimately to 4 bins. Notably, the common-support losses when splitting students into terciles and quartiles are large enough that even the naive gains decrease from the 2 bin case. Furthermore, the estimated error in each specification increases with the number of bins. This means that the robust gains from reallocation are the largest when splitting students into two achievement groups.

Furthermore, we see that the gains from estimating heterogeneous value added for each teacher across two achievement groups are larger that other possible student groupings, though estimating heterogeneous value added across racial/ethnic groups comes close. This is shown in panel (b). Splitting students into two gender groups shows small losses to missing cells and estimation error, but also relatively small achievement gains overall. Once we take two gender groups times two achievement bins, however, we see quite large gains. Losses to empty cells play a much larger role when splitting students into racial/ethnic groups, especially once we take these groups times two achievement groups.

These results suggest that the costs to increasing the amount of estimated heterogeneity outweigh the benefits beyond two student groups in our setting. However, this need not be true in every setting. For example, if a district had broad academic *and* racial/ethnic diversity across classrooms, then it is possible that allowing heterogeneity across two achievement bins by two racial/ethnic student grouping would provide the maximum gains.

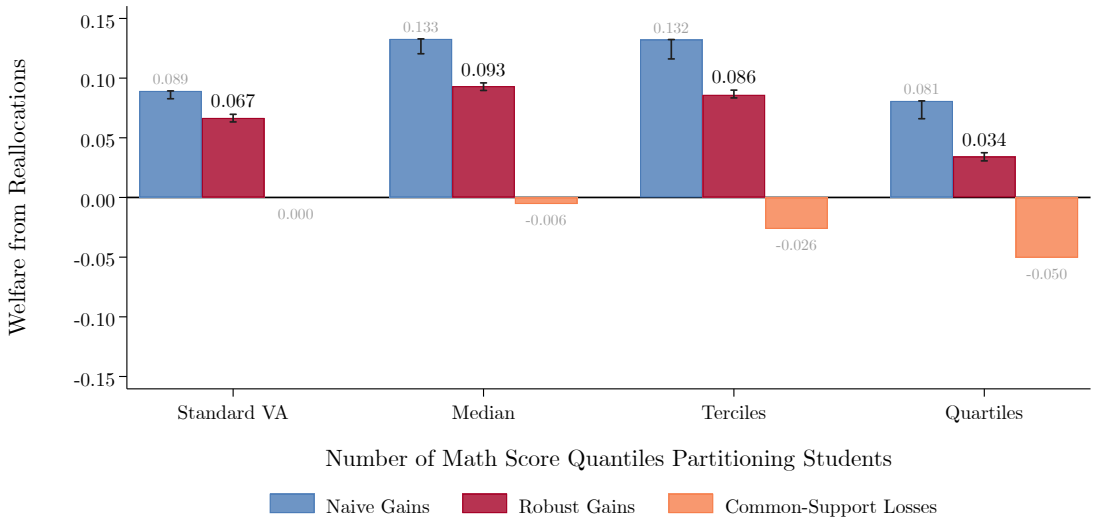
4.3 Equity Implications from Reallocations

Having described the optimal reallocations and decomposed the welfare gains from them, we also explore the equity implications that the proposed reallocations would have. Specifically, we study how our reallocations affect overall achievement gaps and racial achievement gaps, and we describe how certain allocations that generate gains on average still create significant heterogeneity for winners and losers masked by that average.

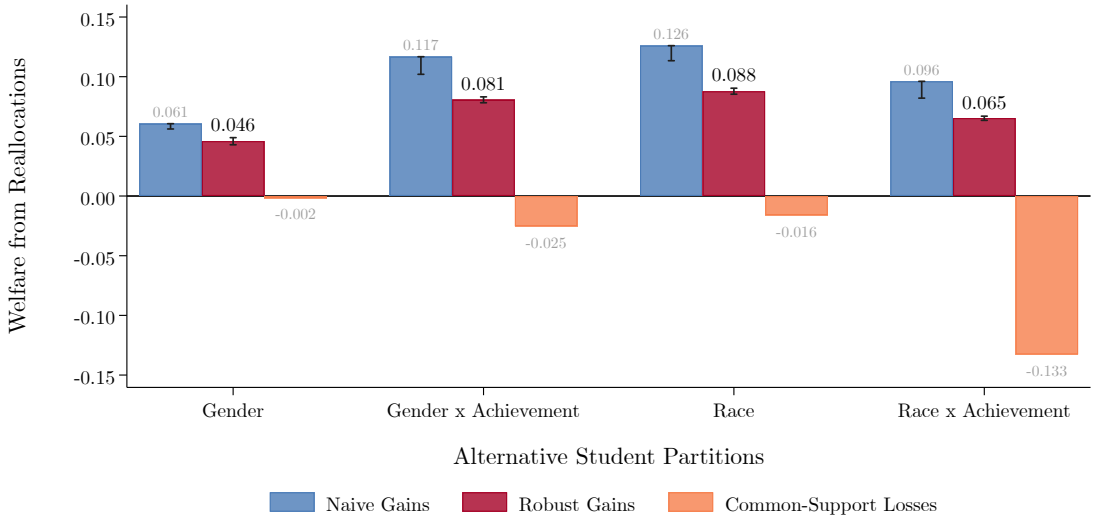
Many education policies—including those that motivated our welfare theory—propose interventions that will lower the achievement gaps between lower- and higher-scoring students. Appendix Figure A.7 shows how the optimal within-school and district-wide reallocations would change the achievement gap between students who performed above and below median in the previous year and between Black and Hispanic students versus White and Asian students. The main takeaway from these analyses is that a social planner who cares about gaps can partially control the size of the gaps by making allocations that are on the efficiency frontier based on comparative advantage. In fact, our race-blind allocations can reduce average racial test score gaps just as much as the race focused reallocations in Delgado (2023).

As noted in the theory, it is unlikely that any reallocation–even one creating large average gains—is a Pareto gain in the sense that it helps, or leaves unaffected, all students. Despite

Figure 5: Gains from More Heterogeneity Are Quickly Swamped by Noise



(a) Achievement Quantiles



(b) Other Characteristics

Note: This figure shows the effects of using heterogeneous estimates of teacher value-added to make allocations. Panel (a) reports gains from using different achievmenet quantiles. Panel (b) explores demographics.

the net gains from matching teachers to their comparative advantages and putting stronger teachers in larger classes, reallocations will assign some students to less effective teachers or to teachers who are a worse match for them (despite the teacher being a better match for their class).[21] Appendix Figure A.8 shows that just as some students experience lower test score growth because of the year-to-year allocations of teachers in the status quo, some also receive lower value-added teachers in our reallocations.

5. From Value Added to Welfare Added

Thus far this paper has considered the trade offs embedded in a simple model of teacher allocations, estimated the relevant parameters, and demonstrated the test score gains from reallocations along a single subject. Our empirical findings in Section 4 make statements about a popular outcome of interest, test scores, but this section considers how teacher reallocations might affect welfare. This requires a few additional assumptions beyond what was required for identification of the teacher effects. In general any reallocation will have implications for students, families, and teachers.

First, we need to make an assumption about family preferences and their behavior in light of our policy change. We assume that families—the main decision-makers for students—value the average achievement of the school they enroll in. This means that students will not resort to new schools after we have rearranged teachers within a school. This is obviously restrictive as parents may value many aspects of education, some idiosyncratic, like having a teacher an older sibling took classes, and others more systematic, like sociability and non-cognitive value-added (e.g., Jacob and Lefgren, 2007; Petek and Pope, 2023; Beuermann et al., 2023). Nevertheless, the vast majority of families do not request specific teachers, and even when they do, not all requests are honored. Given extensive evidence that families do not respond to information about value-added in school choice (Abdulkadiroğlu et al., 2020) or housing markets (Imberman and Lovenheim, 2016), we think this assumption is not too restrictive. Readers critical of this assumption should consider all welfare gains in partial equilibrium terms.

Second, because reallocations affect teachers, we need to consider the welfare component attributable to teachers' disutility from the reallocations. Section 2 deals with this by imposing that teacher have equal surplus in any assignment. Instead, we can treat teacher's preferences as an incentive compatibility constraint and assume they will need to be compen-

[21] This trade-off is not unique to teacher allocations based on value added, so in that sense the "harms" presented here should be benchmarked by the fact that in the status quo students are regularly assigned to less effective teachers, and for these students, the average "loss" (relative to the expectation) is about 0.10 student standard deviations in their scores on tests of each subject.

sated enough to willingly switch classes. Using a revealed preference argument, if teachers willingly move, they will have been made better off. Assuming all teachers must be compensated for changing assignments will likely overstate the cost to teachers because many may prefer their new assignments,[22] but the main challenge is how to price this disutility. Some papers have attempted to price the disutility to teachers from various policies (e.g., Rothstein, 2015; Bates et al., 2024), but highly structured wages in teacher labor markets often make this difficult in practice. Instead, we will focus on the marginal value of public funds (MVPF, Hendren and Sprung-Keyser, 2020) for a set of hypothetical, incentive-compatible teacher bonuses for reallocating.

Note that by restricting our focus on families and teachers in this way, we implicitly assume that other considerations like union concerns or the administrative costs of performing the reallocations are negligible. While these considerations are likely important, we argue that welfare gains of a large enough magnitude could allow transfers or interventions to alleviate these concerns or pay these costs.

5.1 Students: Earnings Implications of Reallocations

We begin with the welfare implications for students under the assumptions outlined above. Even under these two assumptions, allocations are still only partially ordered because teachers affect reading scores, math scores, behavior, and attendance. The optimal reallocation policy would consider the impact on all of these outcomes simultaneously. To combine both of these subjects into a single score function, we map achievement gains to lifetime earnings, which we do using the subject-specific estimates from Chetty et al. (2014a). Given (surprising) evidence that long-term outcomes are mainly captured in test score value added (Hanushek et al., 2024; Lavy and Megalokonomou, 2024), we focus on math and reading. This subsection demonstrates our approach for finding the optimal achievement gains for students' lifetime earnings and performing allocations that maximize those income gains.

5.1.1 Choosing an Income-Optimal Score Function

Because there are numerous allocations, all of which would generate different earnings outcomes, our first objective is choosing a welfare "score" function to maximize income. To do so we use the subject-specific estimates of the effects of value-added in math or reading on student earnings from Chetty et al. (2014a). They estimate that a one standard deviation increase in reading scores in elementary school generates an additional $1,524 in earnings in early adulthood and that the corresponding gains in math are $650.

[22]For example, some teachers will be sent to schools they would like to teach at but cannot because of opening and union tenure requirement.

Because of the fundamental trade-off between the facts that our reallocations generate larger gains in math, but gains to reading matter more for earnings, we take a principled approach to defining the income-optimal allocation. We consider the following set of utilitarian score functions that take into account value-added in two subjects, s, reading and Math.[23]

$$\tilde{\mathcal{W}}(\mathcal{J};\omega) = \frac{1}{N_{i,t}} \sum_{(i,t)} \sum_{s} \omega_s \left[L_{i,s,t}\, \hat{\tau}_{L,s}^{\mathcal{J}(i,t)} + (1 - L_{i,s,t})\, \hat{\tau}_{H,s}^{\mathcal{J}(i,t)} \right] \tag{5}$$

where ω_s represent the weight on each subject and $\sum_s \omega_s = 1$. And now $L_{i,s,t}$ indicates whether the student is low scoring in that particular subject.

Solving the optimization problem for a range of $\omega_{ELA} \in [0.0, 1.0]$ generates a production possibility frontier similar to those in the reallocation exercises in Section 4. Whereas the previous PPF plotted the trade-offs of possible gains between higher- and lower-scoring students, the PPF in Panel (a) of Figure 6 presents the trade-offs between gains to average Math and average reading scores. For example, an allocation focused entirely on Math scores could raise average math scores by 0.058σ (0.016σ within schools). Because Math and reading value-added are somewhat correlated, this allocation would also raise reading scores by 0.019σ (0.005σ within schools). The focus on math scores only, however, forgoes large reading gains. This could be particularly problematic as reading gains are nearly 2.5 times more important for earnings.

We combine the information on possible gains with the estimates of the subject-specific income effects of those gains to calculate the weight each subject that maximizes income gains. The estimates from Chetty et al. (2014a) create relative "prices" of gains to scores in each subject measured in earnings. As such, the income-maximizing weight sets the marginal rate of substitution between reading and math scores equal to the relative price. We illustrate this graphically in Panel (a) of Figure 6 using a dashed line with a slope of the relative price. This line is tangent to the within-school PPF at $\omega_{ELA} = 0.71$ and to the district-wide PPF at $\omega_{ELA} = 0.70$. These values favor reading gains, but do not focus exclusively on reading value-added because the value of marginal gains to reading scores from increasing ω_{ELA} beyond 0.71 are smaller than the value of the larger gains to increasing math scores.

The combination of gains from both subjects significantly increases the income gains from students. The facts that math value-added scores have higher variance and result in larger achievement gains from reallocations might motivate a social planner to focus only on math scores in their objective function. In fact, this intuition plays out in the policy experiments considered in Delgado (2023) and Bates et al. (2024) which both focus only

[23] We will soon relax the assumption about a utilitarian social planner.

on math. Surprisingly, our results overturn this intuition. A district-wide allocation that focuses only on math scores increases average present-valued earnings by $1030. The insight that we can incorporating information about both math and reading optimally generates gains of $1390 per student. This $360 (34%) gain is large and is costless once one allows the social planner to optimally weight value-added to both test scores.

5.1.2 Characterizing Possible Income Gains

With information about the income-optimal score function in hand, we return to the question of optimal policy with heterogeneous social preferences. Combining all of the pieces we define a new social welfare function to optimize

$$\begin{aligned}\tilde{\mathcal{W}}(\mathcal{J};\omega) = \frac{1}{N_{i,t}} \sum_{(i,t)} \omega_L \left[\omega_{\text{ELA}}\, L_{i,\text{ELA},t}\, \hat{\tau}^{\mathcal{J}(i,t)}_{L,\text{ELA}} + (1-\omega_{\text{ELA}}) L_{i,\text{Math},t}\, \hat{\tau}^{\mathcal{J}(i,t)}_{L,\text{Math}}\right] \\ +(1-\omega_L)\left[\omega_{\text{ELA}}\,(1-L_{i,\text{ELA},t})\,\hat{\tau}^{\mathcal{J}(i,t)}_{H,\text{ELA}} + (1-\omega_{\text{ELA}})\,(1-L_{i,\text{Math},t})\,\hat{\tau}^{\mathcal{J}(i,t)}_{H,\text{Math}}\right]\end{aligned}$$

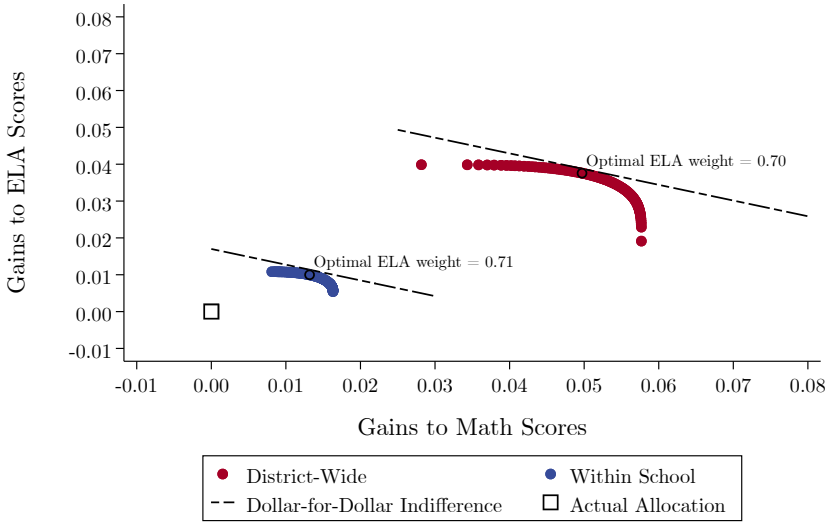
where now we explicitly sum test score gains over both subjects and both student types with their respective weights. Because this formulation exponentially increases the dimensionality of ω, we use our evidence about income-optimal weights to choose $\omega_{\text{ELA}} = 0.75$ and $\omega_{\text{Math}} = 0.25$ in this section. To the extent to which the optimal ω^*_{ELA} varies over ω_L, our results provide a lower bound on the true earnings gains.[24]

After calculating the efficient allocations for each ω, we use the process in Chetty et al. (2014a) to map the test score improvements into the present value of lifetime earnings. We outline our approach as follows. First, we assume that individuals may choose to work between the ages 20 and 65. We also assume that the average income gains implied from test scores apply to all of these earning. Finally, we assume that families discount these earnings gains at a 3% (i.e., with a 5 percent discount rate partially offset by 2 percent wage growth) back to age 10, the average age of students in our sample. Empirically this implies a multiplier of 15.5 on the baseline gains implied from test scores.

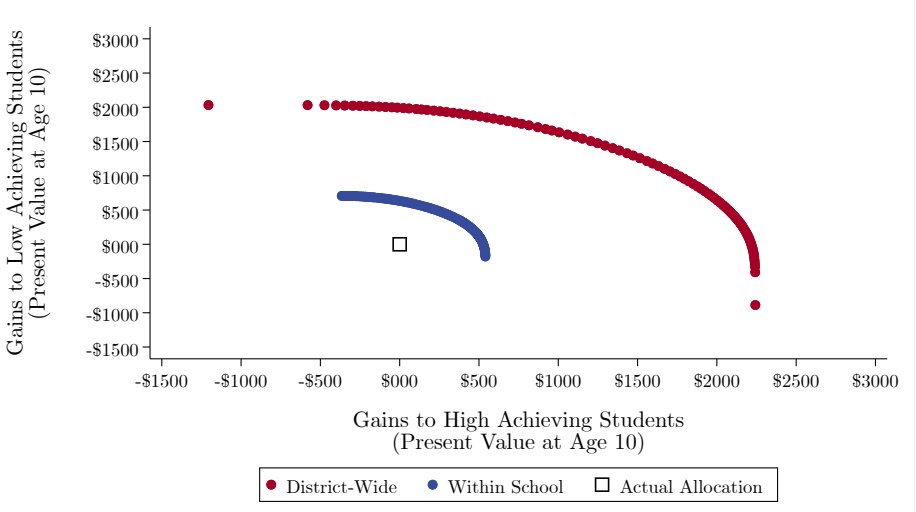
The results, depicted in Panel (b) of Figure 6 show that optimally reallocating teachers could create millions of dollars of gains per year. Based on our calculation, the income-maximizing district-wide allocation would generate over $1140 in present valued earnings for low scoring students and over $1630 for high-scoring students. Since there are 10,150

[24]Note that because not all students are low scoring in Math and reading the achievement weight ω_L may not apply uniformly to each student. In practice this means that there are four implicit weights generated by this welfare function. One conceptually simple way to think of this function is treating each student's score as a different student and then weighting the welfare from gains to that "student" by both their achievement and which test it is.

Figure 6: Reallocations Can Shrink Persistent Gaps in Student Performance



(a) Choosing the Wage-Maximizing Score Function



(b) Present Value Income Gains

Note: This figure shows how we combine math and reading scores to estimate the frontier of possible earnings gains. Panel (a) displays the PPF of math versus reading gains (assuming equal weights). The tangent lines are those implied by the subject-specific estimates of Chetty et al. (2014a). Panel (b) shows the implied effect on lifetime earnings from reallocations with a score of S =0.75 reading + 0.25 Math (present valued at age 10).

students of each type each year (on average), this implies the value of a reallocation across all students is $27.9 million each year. While smaller, the gains from the within school reallocations are not insignificant: over $400 for lower-scoring students and over $300 for high-scoring students, implying $7.4 million across the district for each year of reallocations.

Policy makers concerned about inequality can also create large redistributive gains. For example in the district-wide reallocation, a social planner could increase the present value of lower-scoring students' earnings by $1990 without hurting high scoring students on average. A similar comparison reveals gains of $600 from within school reallocations. Compounded year-over year gains like these could be powerful tools at reducing not only achievement, but also earnings inequality among students coming out of the district. In Appendix Figure A.9, we compare these results to those of a social planner with continuous CES preferences across students rather than discrete preferences across groups and show similar patterns.

Taken together the gains from this policy are enormous. Even if the $27.9 million gain is infeasible because of teacher or union preferences, the within-school reallocation is an essentially costless program generating nearly quarter of those gains. This underscores the power of using information about comparative advantage to improve policy. Furthermore, if there are ways to make the $27.9 million gains attainable, a discussion of how to do so is of first-order importance. The following subsection provides one such discussion.

5.2 Teachers: Welfare Value of a Teacher Bonus Program

We now turn to the welfare implications for teachers. Rather than trying to price teacher disutility, we focus on a teacher bonus thought experiment. One advantage of considering this experiment is that it allows us to separately consider welfare and incentive compatibility. Our estimates reflect the welfare attainable for each policy and would allow policymakers to choose the optimal one based on their understanding of the incentive constraints (e.g., teacher supply, wages, amenities, seniority, unions, etc.).

Imagine a policy that paid all teachers a certain bonus for participating in a reallocation. Teachers would be paid this bonus whether or not their school or class assignment changed. If the bonus was sufficient to ensure incentive compatibility, then one way to characterize the welfare under the resulting allocation would be the marginal value of public funds (MVPF, Hendren and Sprung-Keyser, 2020). This characterizes a lower bound on an envelope of possible incentive programs that could be improved by targeting bonuses the teachers with the highest impacts from reallocation or by relaxing the requirement to participate in the reallocation (for example, for teachers with very strong preferences to their current assignment.

The MVPF is a "bang-for-the-buck" measure of the bonus program, calculated as the

present value of the total program benefits divided by the net cost of implementing it. Specifically, for a bonus of size b the MVPF of allocation j is

$$MVPF^j(b) = \frac{\sum_i (1-t)\Delta S_i^j)}{N_j b - t\Delta S_i^j} \tag{6}$$

where $(1-t)\Delta S_i^j$ are the after-tax present-value monetary gains to each student from allocation j (given tax rate t), N_j is the number of teachers and $t\Delta S_i^j$ is the present-value of gains recouped as tax revenue. The key assumption required for this statistic to be meaningful in this policy thought experiment is internalizing the fiscal externality of the district's policy. For example, this could be interpreted as the national value of the district administering the reallocation policy. Although it is possible to compare national and local MVPFs (e.g., see Agrawal et al., 2023), we focus on this simplified case as in other work (Hendren and Sprung-Keyser, 2020).(Hendren and Sprung-Keyser, 2020).[25]

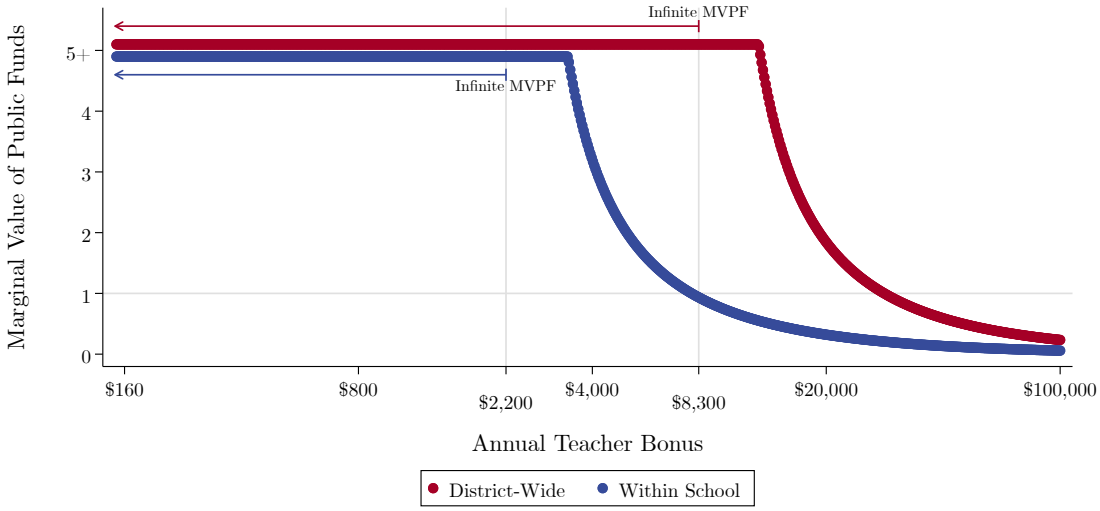
We combine our estimates of present-value monetary gains with data from the Opportunity Atlas (Chetty et al., 2018) to calculate these MVPF empirically. For the changes in earnings, we focus on the utilitarian, earnings-maximizing, within-school and district-wide reallocations as described in the previous subsection. To compute the tax rate, we note that 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, implying $t = 0.28$ for in equation 6.

We present the results in Figure 7. Figure 7 plots the Marginal Value of Public Funds over a broad support of possible bonus sizes (using a log scale on the x-axis). The two series represent the MVPF of a bonus program of a given size for the district-wide or within-school reallocations. The curve showing the value of bonuses for the within-school reallocations is lower because those reallocations produce smaller gains. For each point, the MVPF can be interpreted as dollars of social benefit produced for each dollar spend on the teacher bonus program. Values of the MVPF above 5 are reported at the same height on the y-axis.

The main takeaway from Figure 7 is that for a broad range of bonus sizes the policy of reallocations and bonuses has an infinite MVPF. An infinite MVPF occurs when the net cost of the program is negative and the benefits are positive. in other words, the district would be *making* money by paying to reassign teachers—and would be increasing student earnings in the process. For the district-wide reallocation, the MVPF is infinite for a bonus of up to $8,300, and it is infinite for bonuses up to $2,200 for the within-school reallocation. This

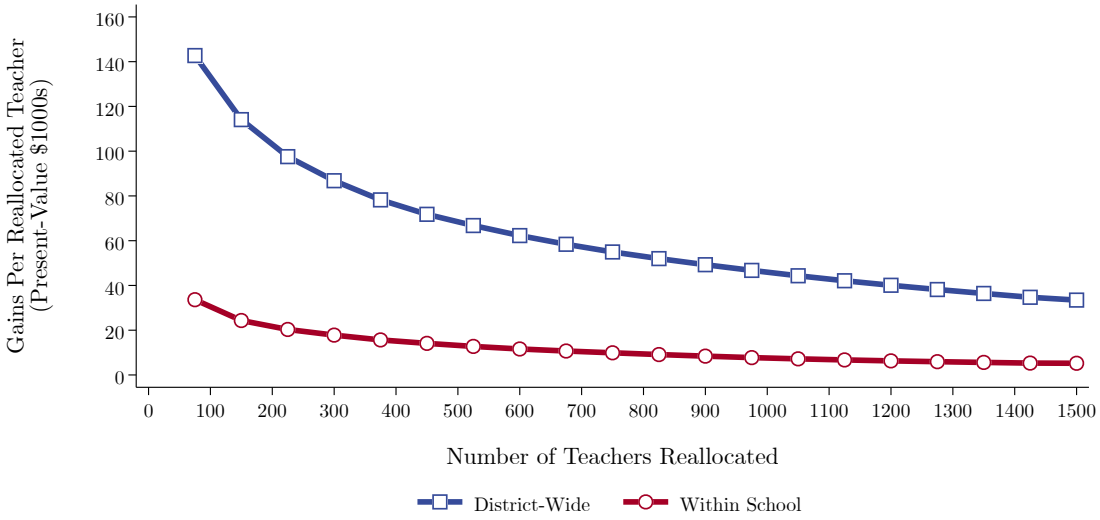
[25]Note that the two could be equivalent if the state and federal governments were to transfer the marginal tax revenue generated by the policy back to the SDUSD.

Figure 7: Compensating Teachers for Reallocations Could Have Enormous Welfare Impacts



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

Figure 8: Gains Per Teacher Are Even Larger Under Targeted Reallocations



Note: This figure shows the expected gains from reallocations of only some teachers. Number of teachers reallocated per year is plotted on the horizontal axis and the present value gains are depicted on the vertical axis.

second number is particularly striking because despite being noninvasive the within-school reallocation is still generating substantial gains.

A second important insight from Figure 7 is that even when the MVPF is not infinite it is still large even for very costly bonus programs. For example, for the district-wide reallocation, a bonus program of paying *every teacher* in the district $20,000 to participate in the reallocation would still have an MVPF of roughly 2. In other words, it would generate $2 of present valued earnings gains for every dollar spent on bonuses. This is a marked pay increase – equivalent to a one-third salary increase for a teacher in the 2010-11 school year with 10 years of teaching experience and the middle tier of education in the district's collective bargaining agreement.

Note that some of these bonus policies may not be incentive compatible, but other research suggests that reallocations with large and even infinite gains could be attainable. For example, while $20,000 may sounds enormous, it amount was shown to be more than enough inducing teachers to move to very low performing schools in a large randomized controlled trial (Glazerman et al., 2013). On the other hand, it's likely that almost all of the within-school reallocations are incentive compatible for most bonuses. First this is because teachers seem to care much more about which school they teach at than which class they teach—in large part because of commuting (Bates et al., 2024)—and this is not affected in the within-school reallocation. Furthermore, in the within-school reallocation most teachers do not even switch classes, suggesting that the utility impact of the reallocation would be particularly small.

Gains are even larger if you restrict switches to only some teachers. For example, our simulations suggest that 50% of the gains are attainable by making 20% of the district-wide swaps. In this case, a targeted incentive-compatible teacher bonus program could be 2.5 times larger and still have the same MVPF as the a district wide policy (e.g., . Within school the gains are even steeper, with 50% of the gains coming from the first 10% of switches. Figure 8 illustrates the total gains produced by each marginal switch as a function of the number of teachers reallocated.

Taken together the teacher bonus thought experiment suggests that the large gains from reallocations are more than an impossibility. Although some teachers would be worse off because of certain reallocations, generating structures that appropriately compensate them for teaching to their comparative advantage could generate tremendous gains. In fact, many of the policies we explore generate large enough earnings gains to students to justify lavish teacher bonuses on the grounds of added tax revenue alone.

6. Discussion and Conclusion

This paper set out to answer two questions: When does heterogeneity matter for maximizing a social objective in general? And how large are the welfare gains from using heterogeneous estimates for refining education policy in particular? We employed and extended tools from public finance to think about aggregating teacher effects on multidimensional outcomes and heterogeneous student types into welfare relevant statistics and implemented them in the context of a large urban school district. In reallocation exercises, using information about both multidimensionality and heterogeneity produce up to double the gains for test scores or for later-life outcomes relative to using standard measures that assume teachers have homogeneous impacts on students, and which focuses on one student outcome rather than two. This highlights the importance of incorporating such information into welfare considerations and policy.

We conclude by exploring three policy trade-offs that our results highlight and discussing possible directions for continued inquiry.

In the specific context of education value-added, our results highlight the power of comparative advantage relative to other policy proposals. Historically researchers have benchmarked the importance of teacher value-added with the a policy "deselecting" (i.e., firing) low-performing teachers (see Hanushek, 2009, 2011; Chetty et al., 2014b; Delgado, 2023). Although deselecting 5% of teachers with the lowest value-added could produce large gains, there are concerns about the ethics of mistakes (Staiger and Rockoff, 2010) and the implications for teacher labor markets (Rothstein, 2015), in the sense that it is not obvious who the replacement teachers will be, and their own teaching effectiveness. An interesting implication of our results, however, is that by relaxing the traditional assumptions of constant effects and equal class sizes we can reallocate rather than release teachers. In our setting a district-wide reallocation would produce gains more than three times larger than the gains from deselecting 5% of teachers. Furthermore, because deselection using standard value-added penalizes teachers who happen to be allocated to worse-matched classes, reallocations prevent incorrect dismissals—16-19% of those targeted. A reallocation-based policy would be less costly to teachers and more beneficial. A within-school reallocation would be even less costly and would still generate 50% of the gains from deselection. In other words, our results suggest that in some, and perhaps many, cases, teachers in the bottom 5-10% need not be deselected, but rather provided an assignment that better matches their comparative advantage. In other cases, where absolute advantage is extremely low, deselection could still be an option.

A second, more general, policy-insight is that our theory can show policymakers how

mean evaluations of existing policies may (or may not) apply to new policy considerations. For example, we show that mean-based welfare estimates can be biased when based on estimates that are not externally valid, or when there is a covariance between welfare weights and treatment effects. While our results clearly indicate the value of considering heterogeneity, even without information beyond the means, policymakers can use these conditions to assess the severity of the bias. For example, using estimates from an expansion of Medicaid to beneficiaries similar to those who are eligible in another state may be very reasonable, whereas assuming that both welfare weights and the elasticity of taxable income are homogeneous along the income distribution may not be. Furthermore, policy can be further improved by conditioning on the relevant dimensions of heterogeneity. Admittedly, using characteristics to condition the estimates often reduces precision—although this type of tradeoff between bias and variability is hardly unique to our setting.

A final policy consideration can be taken from our results at large. Since value-added and other mean evaluations are useful in so many contexts, we hope many practitioners will extend the use of heterogeneous estimates. As they do our research can provide a framework for the gains from adding heterogeneity and which dimensions of heterogeneity and multidimensionality to add and which to ignore. While our results highlight striking patterns in how value-added heterogeneity specifically may affect the long-term outcomes of students, we note that assessing the optimality of reallocation policies in the long run will depend on heterogeneity in the long-term effects. We think an important next step in this literature is directly assessing the effect of multi-dimensional measures of teacher quality on various life-long outcomes and particular the heterogeneity in these relationships across groups.

Taking a step back, our results also highlight the value of testing for and estimating heterogeneous estimates of teacher impacts, and of causal effects more broadly. Whether it is allocating teachers to classes, assessing racial health disparities in care, comparing possible social services, or measuring the effects of firms on earnings growth, the mean is rarely enough to characterize the full question of interest. Although estimating and implementing these evaluations can be costly, researchers have their own comparative advantage in such analyses, and our results suggest enormous gains from finding ways to leverage that knowledge to improve allocation in public programs of many types.

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.

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.

Athey, Susan, Raj Chetty, Guido W Imbens, and Hyunseung Kang (2019) "The Surrogate Index: Combining Short-term Proxies to Estimate Long-term Treatment Effects more Rapidly and Precisely,"Technical report, National Bureau of Economic Research.

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.

Bates, Michael D, Michael Dinerstein, Andrew C Johnston, and Isaac Sorkin (2024) "Teacher Labor Market Policy and The Theory of the Second Best," forthcoming, Quarterly Journal of Economics.

Bau, Natalie (2022) "Estimating an equilibrium model of horizontal competition in education," *Journal of Political Economy*, Vol. 130, No. 7, pp. 1717–1764.

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 (2023) "Predicting and Preventing Gun Violence: An Experimental Evaluation of READI Chicago,"Technical report, National Bureau of Economic Research.

Biasi, Barbara, Chao Fu, and John Stromme (2021) "Equilibrium in the market for public school teachers: District wage strategies and teacher comparative advantage," working paper, National Bureau of Economic Research.

Boyd, D., H. Lankford, S. Loeb, and J. Wyckoff (2005a) "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.

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.

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, 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.

Dahlstrand, Amanda (2022) "Defying Distance? The Provision of Services in the Digital Age,"Technical report.

Delgado, William (2023) "Disparate Teacher Effects, Comparative Advantage, and Match Quality."

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.

Einav, Liran, Amy Finkelstein, and Neale Mahoney (2022) "Producing Health: Measuring Value Added of Nursing Homes,"Technical report, 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.

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.

Gurobi OptimizationLLC (2024) "Gurobi Optimizer Reference Manual," URL: `https://www.gurobi.com`.

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," working paper, National Bureau of Economic Research.

Harrington, Emma and Hannah Shaffer (2023) "Estimating Prosecutor Effects on Incarceration and Reoffense,"Technical report, Working Paper.

Hendren, Nathaniel and Ben Sprung-Keyser (2020) "A Unified Welfare Analysis of Government Policies," Vol. 135, No. 3, pp. 1209–1318, URL: `https://academic.oup.com/qje/artjournal={QuarterlyJournalofEconomics}`, DOI: 10.1257/jep.34.4.146.

Hull, Peter (2020) "Estimating Hospital Quality with Quasi-Experimental Data,"Technical report, 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,"Technical report, National Bureau of Economic Research.

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 (2021) "Selection on Welfare Gains: Experimental Evidence from Electricity Plan Choice,"Technical report, National Bureau of Economic Research.

Jackson, C Kirabo (2018) "What do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes," *Journal of Political Economy*, Vol. 126, No. 5, pp. 2072–2107.

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.

Johnson, Andrew C. (2021) "Preferences, Selection, and the Structure of Teacher Pay," working paper.

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.

Krueger, Alan B (1999) "Experimental Estimates of Education Production Functions," *The Quarterly Journal of Economics*, Vol. 114, No. 2, pp. 497–532.

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

ment," working paper, National Bureau of Economic Research.

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

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.

Ricks, Michael David (2022) "Strategic Selection Around Kindergarten Recommendations,"Technical report.

Rivkin, Steven G, Eric A Hanushek, and John F Kain (2005) "Teachers, schools, and academic achievement," *econometrica*, Vol. 73, No. 2, pp. 417–458.

Rothstein, Jesse (2010) "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *The Quarterly Journal of Economics*, Vol. 125, No. 1, pp. 175–214.

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

Staiger, Douglas O and Jonah E Rockoff (2010) "Searching for Effective Teachers with Imperfect Information," *Journal of Economic Perspectives*, Vol. 24, No. 3, pp. 97–118.

Stepner, Michael (2013) "VAM: Stata module to compute teacher value-added measures," Statistical Software Components, Boston College Department of Economics, October, URL: `https://ideas.repec.org/c/boc/bocode/s457711.html`.

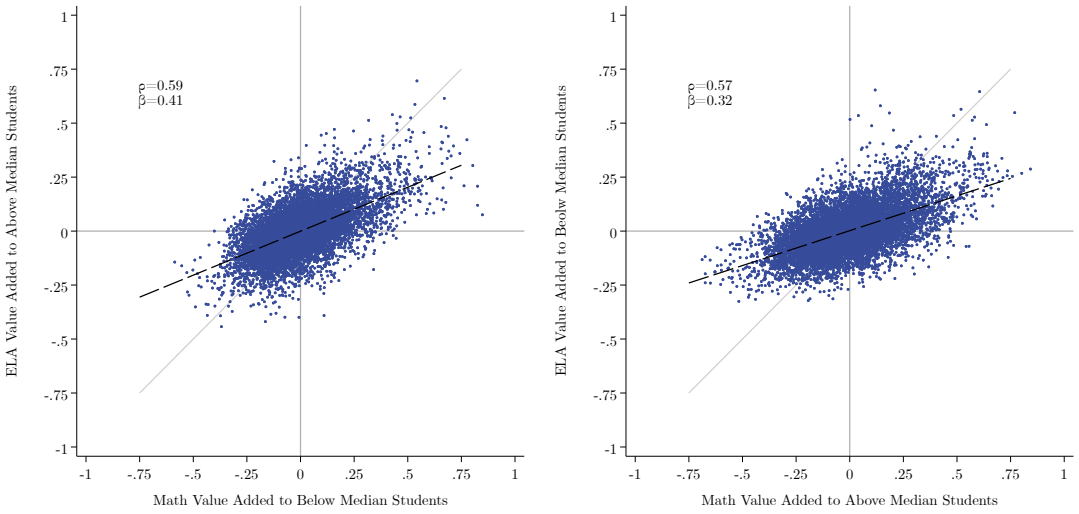
A. Additional Tables and Figures

Table A.1: The Standard Deviation of Class Size and the Share of Students in the Class Who Are High-Scoring in reading and Math

AFTER CONTROL FOR:	STD. DEVIATION CLASS SIZE	STD. DEVIATION SHARE OF CLASS ABOVE MEDIAN, ELA SCORES	STD. DEVIATION SHARE OF CLASS ABOVE MEDIAN, MATH SCORES
Grade*Year	3.68	0.50	0.50
School*Grade*Year	1.71	0.46	0.46

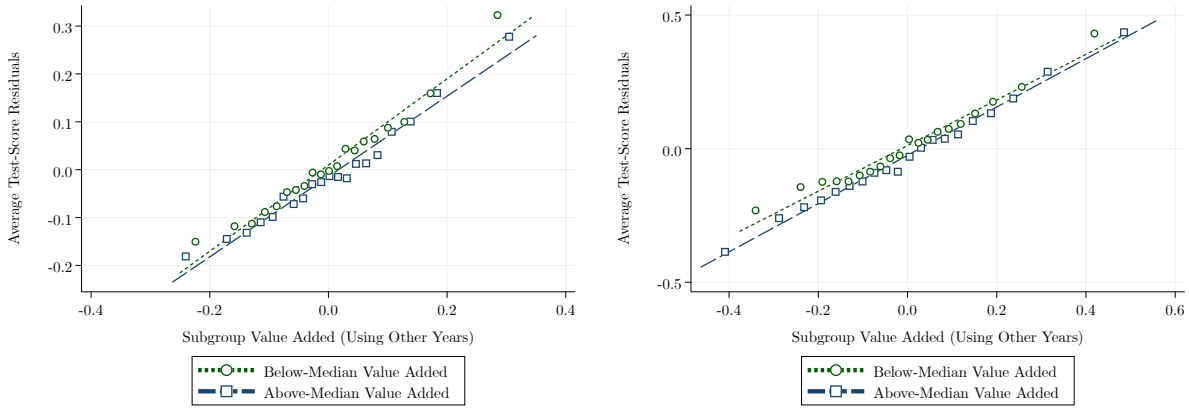
Note: This figure shows the within year-grade standard deviations in class size and composition at a district-wide level and a within-school level.

Figure A.1: Cross-Subject and Cross-Type value-added Is Much Less Correlated



Note: This figure shows our heterogeneous estimates of teacher value-added on both reading and Math test scores. Note that in this Figure Math and reading scores are plotted against each other. Each dot represents one teacher-year estimate of value-added on higher- and lower-scoring students. The correlation coefficients is for the entire population stacked by year. The dotted line shows the line of best fit with the slope reported. For reference a line with slope one is plotted in the background.

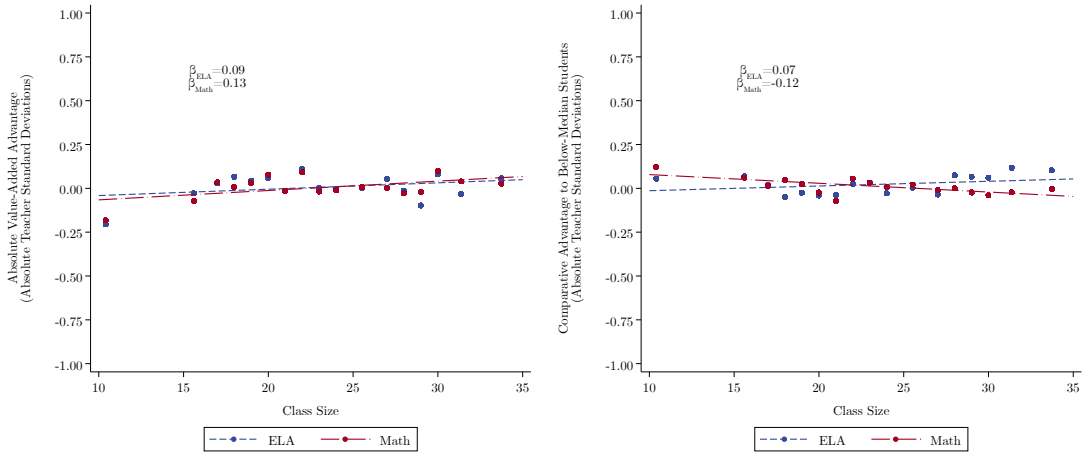
Figure A.2: Above- and Below-Median Value-added measures are Forecast Unbiased



(a) Math (b) Reading

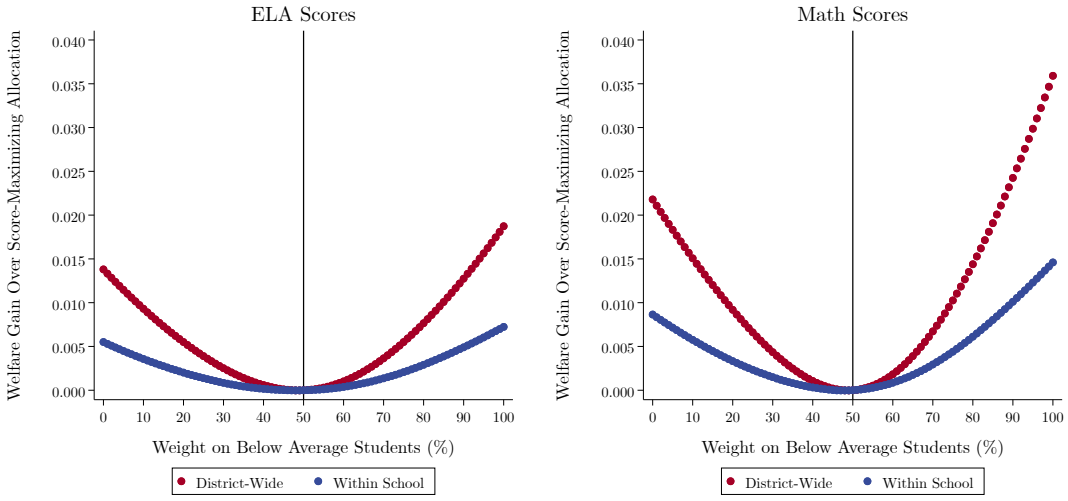
Note: These are estimates of forecast unbiasedness of both above- and below-median value added for teachers in math and reading.

Figure A.3: Value Added Only Varies Somewhat Across Class Sizes



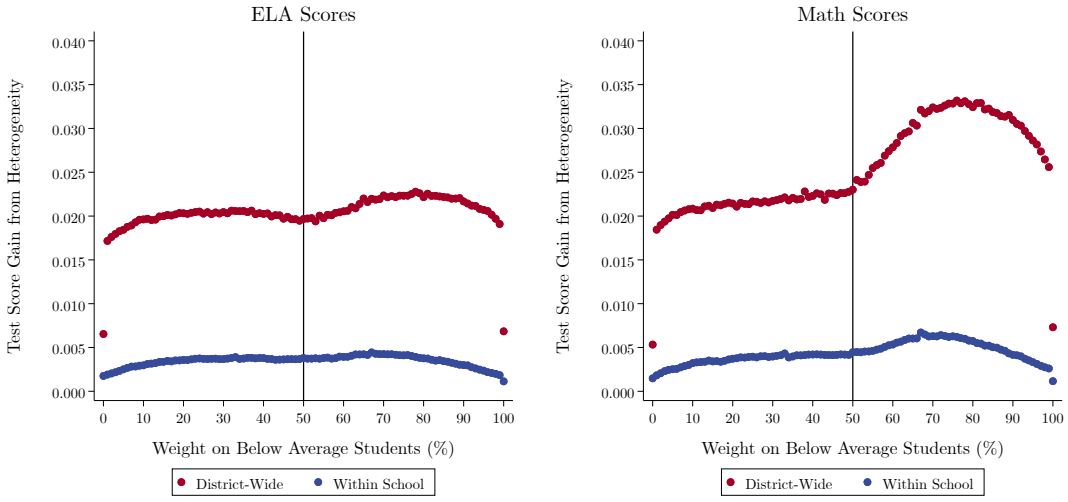
Note: This figure shows how our heterogeneous estimates of teacher value-added on both reading and Math test scores relate to class composition. The panel on the left shows teacher absolute advantage (average of value-added on higher- and lower-scoring students) and the panel on the right shows the comparative advantage (difference of value-added on below-median students minus value-added on higher-scoring students). both panels plot the ventiles of value-added (measured in teacher standard deviations in absolute advantage) over the share of number of students in each class. Both β report the change from a 25-student change in class size.

Figure A.4: Welfare Gains from Considering Distributional Objectives



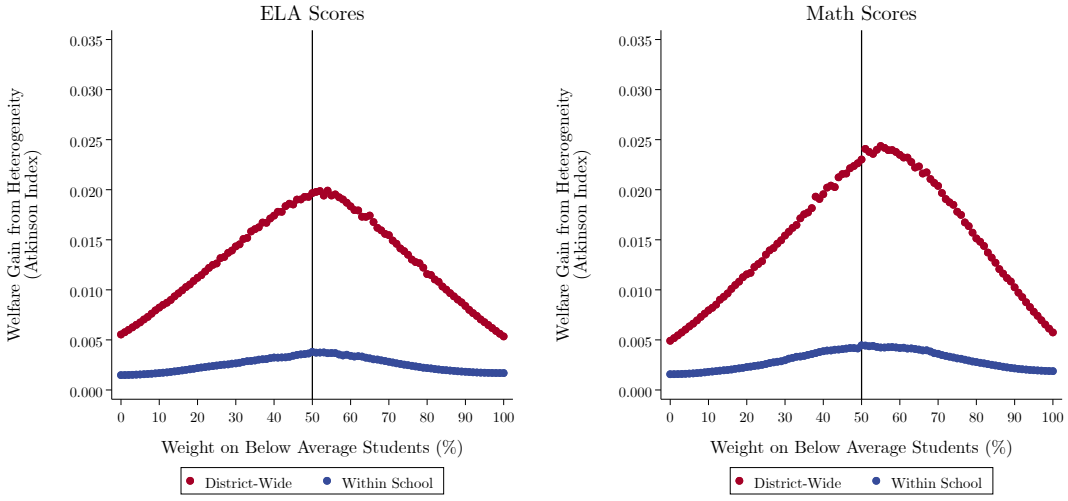
Note: This figure shows the differences in welfare attained under the score maximizing allocation and the optimal allocation using heterogeneous value-added. The unit is an Atkinson Index indifference, i.e., how much would test scores have to increase for all students to generate equivalent welfare gains. We report differences for both within-school and district-wide reallocations. At the tail ends, where the policymaker favors one group almost exclusively, the gains for the district-wide (within-school) reallocations are 85% (20%) larger in math and 50% (35%) larger in reading. Of course, the true weights for policymakers may not be near these tails, but Figure A.4 demonstrates significant potential for gains in the right setting. These potential welfare gains highlight the fact that choosing the allocation that maximizes average scores isn't necessarily a neutral choice. For example, in math it benefits higher-scoring students more.

Figure A.5: Test-Score Gains from Using Heterogeneity



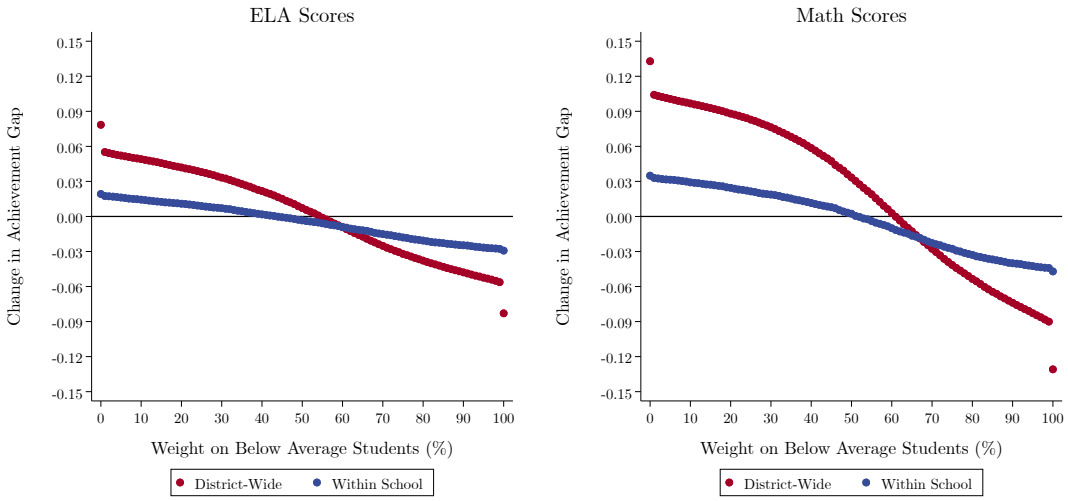
Note: This figure shows the test scores gains from using our measures of heterogeneous value-added to make allocations relative to standard measures over various social preferences.

Figure A.6: Welfare Gains from Comparative Advantage Along Distributional Objectives

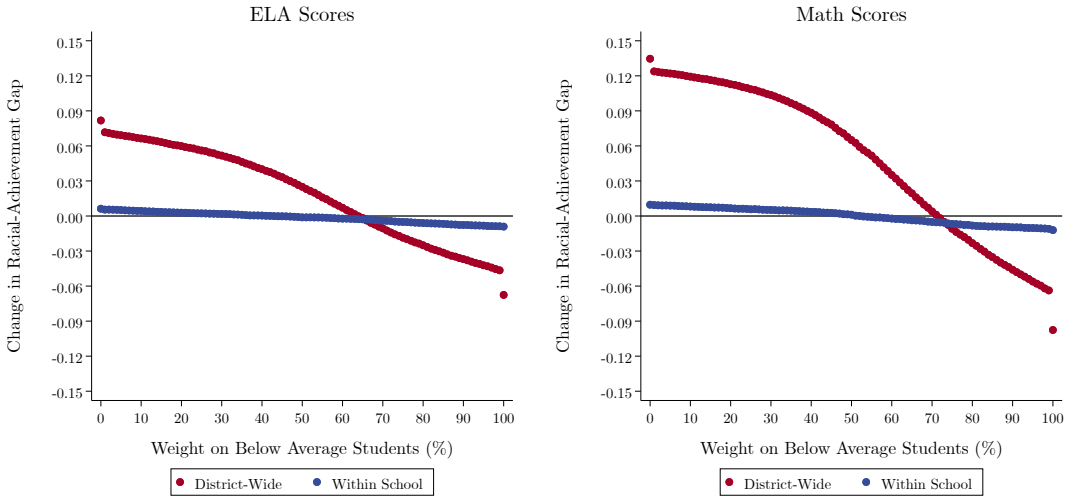


Note: This figure compares the welfare attained at the optimal allocations based on our measures of value-added with those attained at allocations based on standard value-added measures. The unit is an Atkinson Index indifference, i.e., how much would test scores have to increase for all students to generate equivalent welfare gains. We report differences for both within-school and district-wide reallocations.

Figure A.7: Reallocations Can Shrink Persistent Gaps in Student Performance



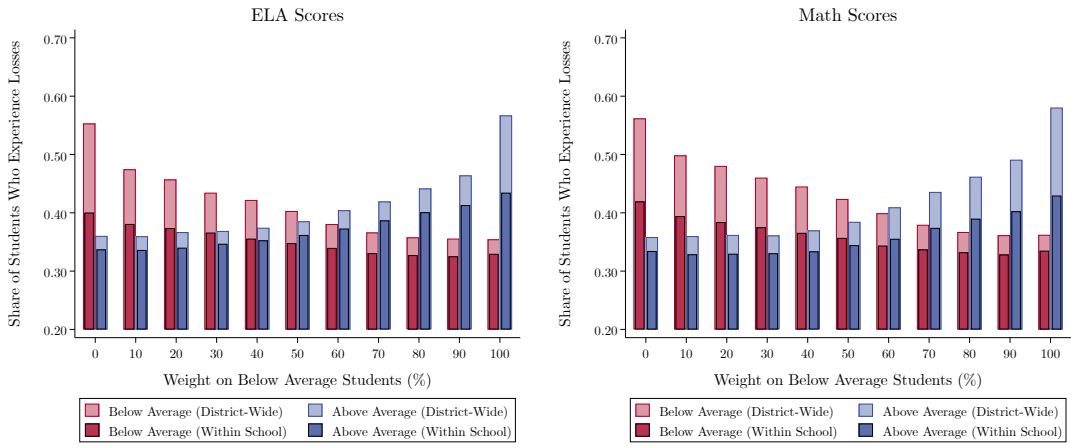
(a) Achievement Gaps



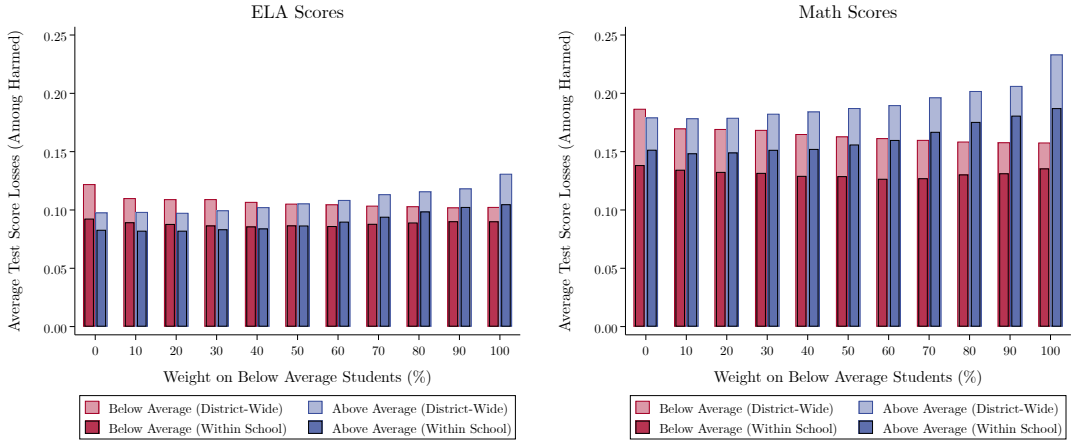
(b) Racial-Achievement Gap

Note: This figure shows how optimal reallocations would change achievement gaps between students. Each panel plots the change in the gaps of interest over the relative weights on higher- and lower-scoring students. Panel (a) displays the change in the average difference in test scores between students who scored below versus above the median in the previous year (relative to about 1.2σ), and Panel (b) displays the change in the average difference in test scores between Black and Hispanic students versus white and Asian students (relative to about 0.7σ). Both gaps are measured in student standard deviations.

Figure A.8: While Reallocations Help Many Students, They Will Harm Others



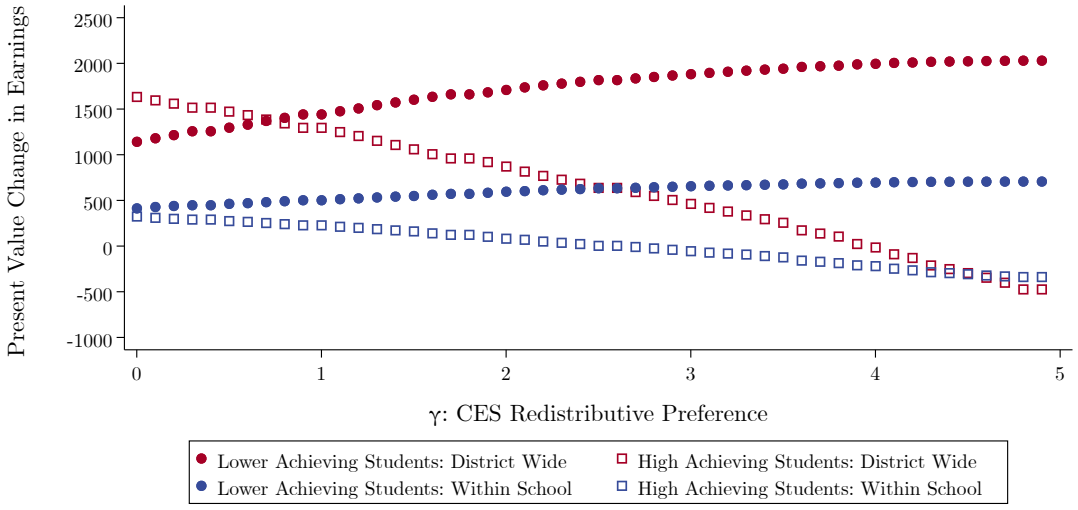
(a) Share of Students Harmed



(b) Mean Score Change among Harmed Students

Note: This figure shows information about which students are made worse off by the reallocations. Panel (a) reports the share of students whose scores would be lowered by each reallocation and Panel (b) reports the average change in scores among those harmed.

Figure A.9: Comparing to a CES Benchmark



Note: This figure shows the present-value earnings gains from optimal reallocations based off of continuous CES preferences over student types rather than discrete preferences between higher- and lower-scoring students.

B. Theory Appendix

B.1 Derivation of Equation 1

Let welfare be a weighted sum of lifetime utilities $W^{\mathcal{J}} = \sum_{i=1}^{n} \phi_i^{\mathcal{J}} U_i^{\mathcal{J}}$ where both the utilities $U_i^{\mathcal{J}}$ and weights $\phi_i^{\mathcal{J}}$ may depend on the policy $\mathcal{J}$. Assume that individual's outcomes $\boldsymbol{Y}_i$ only impacts the utility and welfare weights of that individual i. Then the expected welfare $\mathcal{W}$ under policy $\mathcal{J}$, given $\boldsymbol{Y}_i$ and $S()$ can be written

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

where the first inequality 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 $\gamma_i^{\mathcal{J}} = \frac{\mathbb{E}[\phi_i^{\mathcal{J}} U_i^{\mathcal{J}} | S_i^{\mathcal{J}}]}{S_i^{\mathcal{J}}}$.

Remark: Interpretation of γ_i. Although this definition of $\gamma_i^{\mathcal{J}}$ is fairly trivial, it has some insightful properties. For example, if we choose $S()$ as an unbiased predictor of utility, such that $\mathbb{E}[U_i^{\mathcal{J}} | S_i^{\mathcal{J}}] = S_i^{\mathcal{J}}$, then $\gamma_i^{\mathcal{J}} = \phi_i^{\mathcal{J}}$. Similarly, if welfare $\phi_i^{\mathcal{J}} U_i^{\mathcal{J}}$ is concave in $S_i^{\mathcal{J}}$, then $\gamma_i^{\mathcal{J}}$ will decrease in $S_i^{\mathcal{J}}$.

B.2 Derivation of Equation 7

Let expected welfare be $\mathcal{W}^{\mathcal{J}}$ as defined in Equation 1, $\hat{\mu}_j$ be an estimate of teacher's value added, and 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 $\widehat{\mathcal{W}}_{VA}^{\mathcal{J}} = \bar{\gamma}^{\mathcal{J}} \sum_j n_j \hat{\mu}_j$

$$\mathcal{W}^{\mathcal{J}} - \widehat{\mathcal{W}}^{\mathcal{J}}_{VA} \equiv \sum_j \sum_{i:\,\mathcal{J}(i)=j} \gamma_i^{\mathcal{J}} \mu_{i,j} - \bar{\gamma}^{\mathcal{J}} \sum_j n_j \hat{\mu}_j \tag{7}$$

$$= \sum_j n_j \left[\left(\frac{1}{n_j} \sum_{i:\,\mathcal{J}(i)=j} \gamma_i^{\mathcal{J}} \mu_{i,j} \right) - \bar{\gamma}^{\mathcal{J}} \hat{\mu}_j \right] \tag{8}$$

$$= \sum_j n_j \left[\frac{1}{n_j} \sum_{i:\,\mathcal{J}(i)=j} \gamma_i^{\mathcal{J}} \mu_i^{\mathcal{J}} - \gamma_c^{\mathcal{J}} \mu_{c,j} + \gamma_c^{\mathcal{J}} \mu_{c,j} + \bar{\gamma}^{\mathcal{J}} \bar{\mu}_j - \bar{\gamma}^{\mathcal{J}} \bar{\mu}_j - \bar{\gamma}^{\mathcal{J}} \hat{\mu}_j \right] \tag{9}$$

$$= \sum_j n_j \left[\underbrace{\gamma_c^{\mathcal{J}} \mu_{c,j} - \bar{\gamma}^{\mathcal{J}} \bar{\mu}_j}_{\text{Matching Gains}} + \underbrace{\widehat{\text{Cov}}_j(\gamma_i^{\mathcal{J}}, \mu_{i,j})}_{\text{Targeted Matches}} + \underbrace{\bar{\gamma}^{\mathcal{J}} (\bar{\mu}_j - \tilde{\mu}_j)}_{\text{External Validity}} + \underbrace{\bar{\gamma}^{\mathcal{J}} (\tilde{\mu}_j - \hat{\mu}_j)}_{\text{Estimation Error}} \right] \tag{10}$$

The first equality comes from factoring out n_j. The second from adding and subtracting $\gamma_c^{\mathcal{J}} \mu_{c,j}$ and $\bar{\gamma}^{\mathcal{J}} \bar{\mu}_j$, where $\gamma_c^{\mathcal{J}}$ and $\mu_{c,j}$ represent the average welfare weights of students in teacher $j's$ assigned class and the average match effect of teacher j on those students. The third comes from adding and subtracting $\bar{\gamma}^{\mathcal{J}} \tilde{\mu}_{c,j}$ and from the definition of covariance (we define the within class covariance as $\widehat{\text{Cov}}_j(\gamma_i^{\mathcal{J}}, \mu_{i,j}) \equiv \frac{1}{n_j} \sum_{i:\,\mathcal{J}(i)=j} \gamma_i^{\mathcal{J}} \mu_i^{\mathcal{J}} - \gamma_c^{\mathcal{J}} \mu_{c,j}$).

C. Value-Added Estimation Details

The above discussion shows the theoretical importance of measuring test score heterogeneity, but of course, measuring heterogeneity increases the variance of estimates. Weather or not it can be effectively measured to improve policy analysis is a practical empirical question. Below we cover two different methods for measuring test score heterogeneity, but first, a quick review of our benchmark traditional value-added estimation.

C.1 Estimators

C.1.1 Standard value-added

In order to reference our estimates against an up to date and rigorously tested value-added approach, we follow the baseline practices used in Chetty et al. (2014a) and implement it using the associated Stata package (Stepner, 2013). The general approach of these authors is as follows. First regress test scores $S_{i,t}$ on controls $X_{i,t}$ which gives test score residuals A_{it}. This is obtained from a regression on test scores of the form

$$S_{i,s,t} = \alpha_{j(i,s,t)} + \beta_s X_{i,t} + \epsilon_{i,s,t} \tag{11}$$

Where $X_{i,t}$ includes cubic polynomials in prior year test scores in math and reading, those polynomials interacted with student grade level, ethnicity, gender, age, lagged suspensions and absences, indicators for special education and English language learner status, cubic polynomials in class and school-grade means of prior test scores in both subjects each interacted with grade, class and school means of all the other covariates, class size and type indicators, and grade and year dummies[26]. $j(i,t)$ is the index for the teacher who has student i in her class at time t, so $\alpha_{j(i,t)}$ are year-specific teacher fixed effects.

Next, we average the residuals within each class year to get

$$\bar{A}_{jt} = \frac{1}{n} \sum_{i \in i:j(i,t)=j} A_{it} \tag{12}$$

The last step is to use the average residuals in every year but year t, denoted $\mathbf{A}_j^{-t}$, to predict $\bar{A}_{jt}$. Specifically, we choose coefficients $\psi = (\psi_i, ..., \psi_{t-1})$ to "minimize the mean squared error of the forecast test scores (Chetty et al., 2014a)"

[26]The covariates match those used in (Chetty et al., 2014a) closely. Means and standard deviations of the underlying variables appear in Appendix Table **??**.

$$\psi = \arg\min_{\psi} \sum_{j} \left(\bar{A}_{jt} - \sum_{s=1}^{t-1} \psi_s \bar{A}_{js}\right)^2 \tag{13}$$

This then gives the estimate for teacher j's value-added in year t of

$$\hat{\mu}_{jt} = \psi' \mathbf{A}_j^{-t} \tag{14}$$

C.1.2 Binned Estimator

A simple way to add heterogeneity into this model is to include an indicator for each student's type and estimate teacher affects separately for each type. This gives each teacher an estimate for each student type. We separate students into above and below median prior year test score bins. All of the above math works out essentially the same except we now have twice as many parameters to estimate. We now estimate residuals from the equation

$$S_{i,t} = \alpha_{j(i,b,t)} + \beta X_{i,t} \tag{15}$$

where $j(i,b,t)$ indicates if student i is assigned to teacher j in bin b at time t. Next we group residuals for teacher, year, bin,

$$\bar{A}_{jBt} = \frac{1}{n} \sum_{i \in i: j(i,B,t)=j} A_{it} \tag{16}$$

and we do the leave-one-out estimator with teacher bin estimates across years

$$\psi = \arg\min_{\psi} \sum_{j} \left(\bar{A}_{jBt} - \sum_{s=1}^{t-1} \psi_s \bar{A}_{jBs}\right)^2 \tag{17}$$

This then gives the estimate for teacher j's bin B value-added in year t of

$$\hat{\mu}_{jBt} = \psi' \mathbf{A}_{jB}^{-t} \tag{18}$$

We also apply statistical shrinkage, using the variance within each bin so that if the variance of one bin is higher it does not get shrunk more relative to the other bins.

C.2 Aggregating Estimates

The above method gives multiple estimates for each teacher's impact on the different types of students. For specific policy interventions, like teacher reassignment, these can be combined by summing up the conditional expected treatment with the conditional average

welfare weight such as the weights described in theorem **??**.

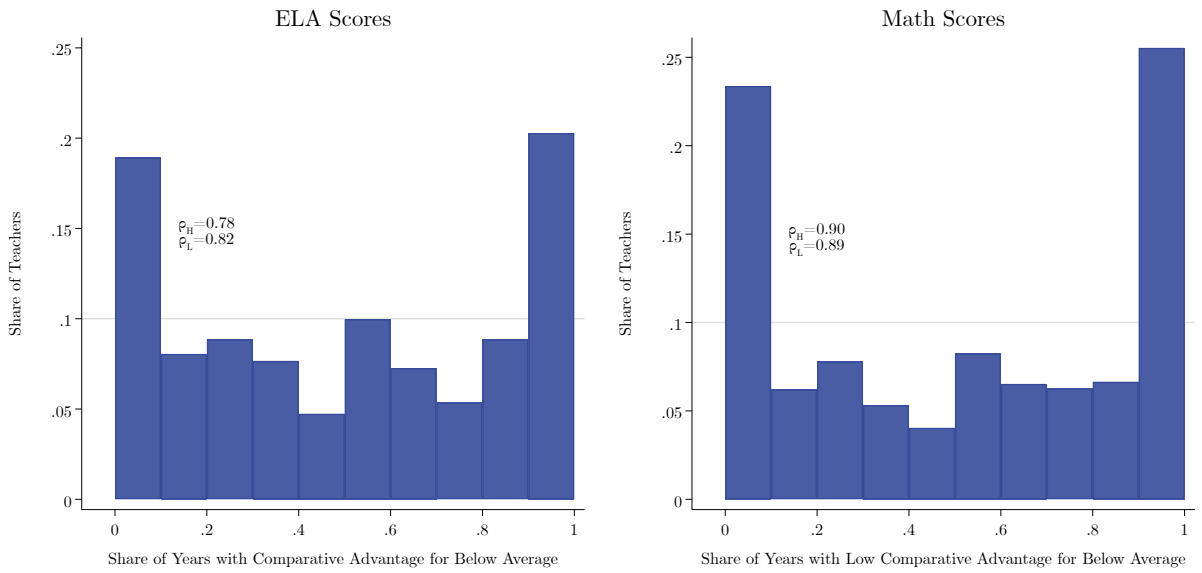
However, in some cases, value-added is also used for general teacher ranking and assessment. If teacher heterogeneity is significant, is there still a way to objectively rank teachers according to a particular set of heterogeneous welfare weights? There is not a perfect single solution since their impact depends on the class or policy environment. However, one solution that puts teachers on an even playing field is to rank teachers on the expected welfare impact they would have on an average representative class, rather than on the average impact on test scores for the class they have, which may depend on class composition, which is outside of the teacher's control and does not reflect their welfare impact.

In the discrete setting, let $\bar{\omega_k}$ and γ_k be the average proportion of students in group k and the welfare weight for group k respectively. Let $\alpha_{j,k}$ be teacher j's group specific value-added for group k. Than we can aggregate their group specific test scores as

$$VA_j = \sum_k \gamma_k \bar{\omega_k} \alpha_{j,k} \tag{19}$$

This gives the welfare benefit a teacher would have on an average class. This is the same as A_j from definition **??**. Now, choosing the average class composition for every teacher may or may not be the right normative choice. Suppose that a teacher has a big comparative advantage with high scoring students in a district with, on average, very high scoring students, but their class is primarily low scoring. What is the right way to assess their performance? They may not be bad relative to their well matched peers, which the above metric could tease out, but they may still in fact be doing a poor job helping the students they have, which the above metric ignores. This emphasizes that in a world of heterogeneity, no metric will be perfect. However, equation 19 does help to rank teachers based on what is under their control.

Figure D.10: Measures of Comparative Advantage Persistent



D. Validation and Robustness of Heterogeneous Estimates

In addition to these standard exercises we leverage the longitudinal nature of our data to show that our heterogeneous estimates capture the same correlations with long term outcomes as do standard value-added does—despite being identified off of only half of the students. In the spirit of Chetty et al. (2014b), we focus on five main outcomes: high school graduation, college enrollment in the year after twelfth grade (two-year, four-year, and any), and completion of a bachelors degree within six years of (anticipated) high school graduation. If our heterogeneous estimates corresponds to future outcomes in a similar way to standard value-added, then the predictive power has not been diminished and the estimated effects are fitting on true value-added rather than idiosyncratic noise.

To test the predictive power of value-added, we regress each outcomes teacher value-added and the controls from equation **??** in a student-subject-grade level regression. For the binned estimates, we include terms for the high- and low-bin value-added interacted with an indicator for whether the student is a high scoring:

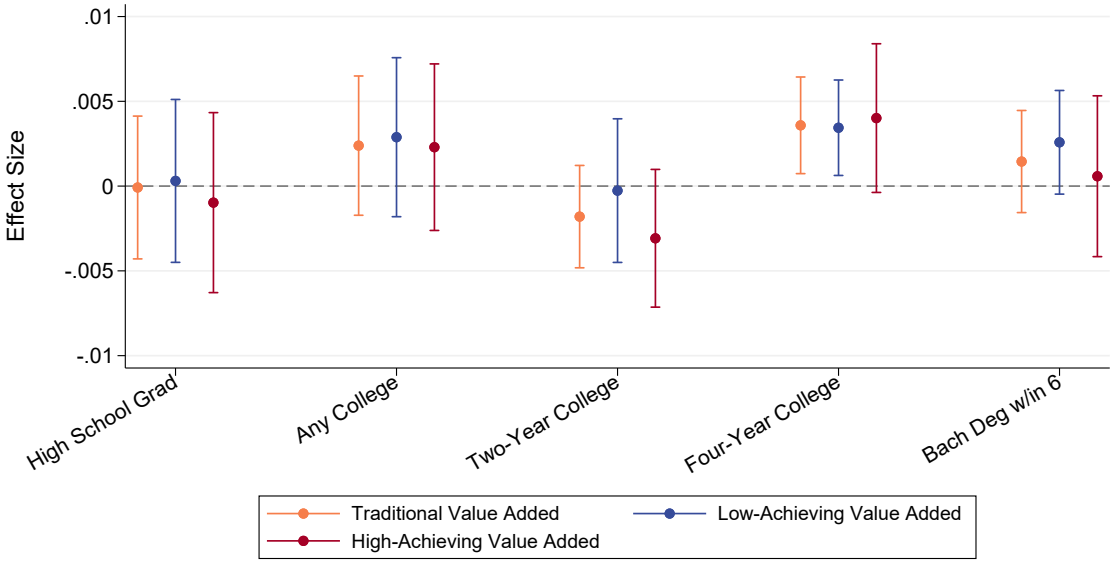
$$y_{i,j,s,t} = \tau_{VA}\hat{\gamma}^{VA}_{j,s,t}\mathbb{1}(k_i = g) + \beta_2 X_i + \nu_{i,j,s,t} \tag{20}$$
$$y_{i,j,s,t} = \sum_{g=H,L} \tau_g \hat{\gamma}^{g}_{j,s,t}\mathbb{1}(k_i = g) + \beta_3 X_i + \nu_{i,j,st}$$

This is analogous to treating the each teacher-subject-bins as a separate class where the

coefficients on value-added indicate the predictive power of high-bin value-added in each subject on high-scoring students' outcomes and low-bin value-added on low-scoring students' outcomes.

Figure D.11 reports the results from the regression in equation 20 on each outcome variable. Our results show striking similarities between traditional value-added and our estimates, despite the fact that we split our sample to estimate above- and below median effects. Surprisingly, none of the measures are predictive of high school graduation. One explanation for this might be that SDUSD has an unusually high graduation rate, averaging 90 percent for our sample, creating ceiling effects. While not statically significant, standard value-added and both of our binned estimates track closely with an increase in any college, primarily from four year college with potentially a drop in two year college, and an increase in a bachelor's degree within 6 years. We can also see that the standard errors for each student group are not actually much bigger than for the mean as a whole suggesting that the variance is loading on this achievement dimension. On a whole these effects are similar with those in Chetty et al. (2014b) and **?** for traditional value-added.

Figure D.11: Our Estimates Predict Long Term Effects as Well as Standard VA



Note: This figure compares the effect of different measures of teacher value-added on long-term outcomes. All regressions follow equation 20 and include all controls from the value-added estimation. For the outcomes, High School Grad is an indicator for whether the student graduated from high school, Two Year College is an indicator for whether the student enrolled in a two-year college within a year following high school graduation, Four-Year College is an indicator for whether the student enrolled in a four-year college within a year following high school graduation, and Any College is an indicator for either Two Year College or Four-Year College. Finally, we model an indicator for whether the student obtained a Bachelor's degree within six years of high school graduation.

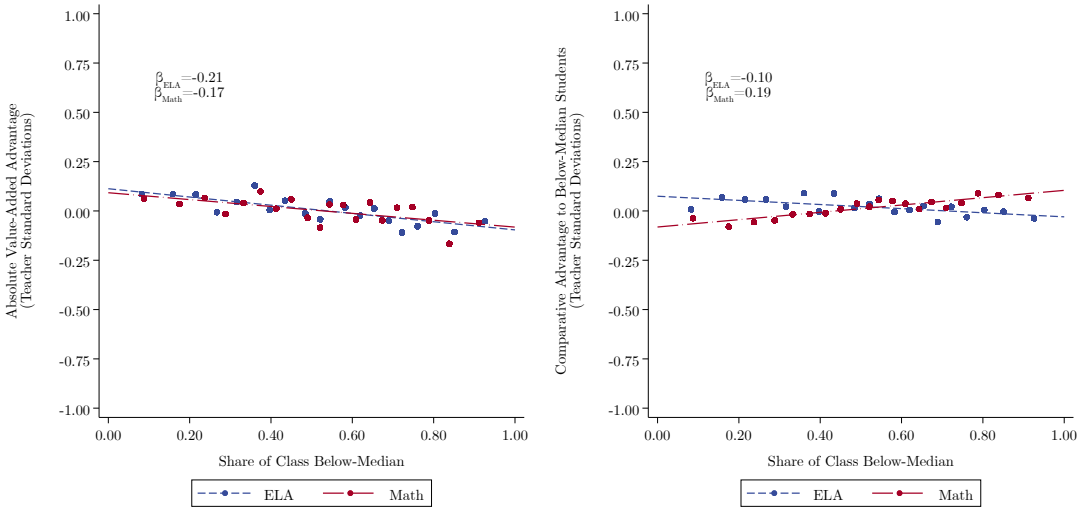
Although imprecise, these effects point to patterns in college enrollment that are inde-

pendently interesting beyond this validation exercise. For example, the effect on two-year college enrollment is higher for below-median students, which makes sense if they are more likely to be on the margin of not going to any college. On the other hand, for high-scoring students, well matched value-added may decrease the probability of two-year college enrollment and increase in the probability of four-year college enrollment. These patterns are consistent with well-matched teachers increasing the quality of post-secondary education, moving students on one margin from no college to two-year colleges and on another margin from two-year colleges to four-year colleges.

E. The Status-Quo Allocation of Teachers and Students

This section shows how teachers are allocated to classes in the status quo, considers whether this allocation is efficient or equitable, and presents descriptive evidence that there may be gains from reallocation. Figure E.12 presents a binned scatter plot of value-added for each subject over the share of lower-scoring students for that subject. Absolute advantage is reported as the average of teacher value-added on lower- and higher-scoring students, and comparative advantage is reported as the difference.

Figure E.12: Teacher value-added Only Varies Somewhat with Class Composition



Note: This figure shows how our heterogeneous estimates of teacher value-added on both reading and Math test scores relate to class composition. The panel on the left shows teacher absolute advantage (average of value-added on lower- and higher-scoring students) and the panel on the right shows the comparative advantage (difference of value-added on lower-scoring students minus value-added on higher-scoring students). both panels plot the ventiles of value-added (measured in teacher standard deviations in absolute advantage) over the share of students who are lower-scoring (i.e. have below-median lagged test scores).

These patterns suggest that classes with larger shares of lower-scoring students do not tend to have teachers with substantially different absolute or comparative advantage. First, teachers with a higher average value-added are somewhat more likely to sort into classes with higher average test scores at baseline. This suggests the current allocation is inequitable, but the effects are small. The slope only predicts that students in a class with an additional lower-scoring student in one subject will experience 0.001σ smaller gains in that subject on average. Interestingly, there is some evidence that this slightly inequitable sorting may be according to absolute advantage. Appendix Figure A.3 shows analogous results by class size revealing that better teachers teach in slightly larger classes, suggesting some allocative

efficiency from sorting better teachers in bigger classes, but again the differences are small. These two patterns are likely connected as more affluent schools tend to have higher average test scores but slightly larger classes.

Second, there is also no decisive evidence of sorting on comparative advantage. Figure E.12 also depicts the comparative advantage in teaching lower-scoring students (measured as the difference in value-added to lower- and higher-scoring students) along the class test score distribution. In math, teachers who are comparatively better at teaching lower-scoring students sort into classes with slightly larger shares of lower-scoring students, but the opposite is true in reading. Neither of these patterns is economically large. The differences by class size are similarly signed but even smaller (see Appendix Figure A.3).

The current allocation of students to classes also suggests that there will be gains from reallocations. Teachers have decidedly different effects on different students. Furthermore, teachers with large average effects are not currently being allocated to larger classes, nor are teachers being allocated according to their comparative advantage. In addition, Appendix Table A.1 reports meaningful variance in class size and class composition which both imply larger the gains from reallocation—whether within or across schools.

F. Additional Analyses about Equity

F.1 Distributional Goals

The above section shows that when the goal is to maximize average scores, using heterogeneous value-added leads to significant gains. We also know from section 4.1 that when policymakers favor one group over another, considering their distributional goals leads to significant welfare gains. Putting these together, we now address how different distributional objectives impact the gains from comparative advantage, and using heterogeneous value-added.

Using Figure 4 as a reference, we now compare the welfare from the optimal points on the inner PPF relying on mean effects and the outer PPF using heterogeneity for a given distributional goal. Reporting the difference in the Atkinson index between the optimal allocations reveals the welfare gains from using heterogeneous value-added estimates for each distributional goal. Figure A.6 reports the results. In Appendix Figure A.5, we present a simpler measure: the true (unweighted) difference in average scores for each pair of allocations.

These analyses reveal that using heterogeneous value-added matters most when the social planner has slightly egalitarian preferences. This is visible in Figure A.6 where for the district-wide reallocation the highest points on each upside-down U shape are slightly to the right of utilitarian preferences denoted with the gray line (at $\omega_L = \omega_H = 0.5$). Although the maxima, where using heterogeneous value-added is most useful, are at ω_L =0.54 for reading and 0.55 for math, the entire region between $\omega_L \in [0.30, 0.70]$ show gains equivalent to over 0.015σ of gains to all students.

The comparative advantage gains from estimating heterogeneous value-added are only large if the social planner cares about both groups. For example, if the social planner only cares about lower- or higher-scoring students ($\omega_L \in \{0.0, 1.0\}$), there are essentially no gains from comparative advantage using heterogeneous value-added. This is because lower- and higher-scoring value-added are positively correlated, so a policy that puts the highest absolute advantage teachers in the class with the most lower-scoring students will have a very similar effect on lower-scoring students to a policy that puts the teachers with the highest lower-scoring value-added in the same classes. This is visible in how close the frontiers are in Figure 4 and in the upside-down U-shape in the gains reported in Figure A.6.

The key driver of these differences are the relative shapes of the PPFs and how they affect scores. As seen in Figure 4, the best attainable allocations using standard value-added create a much flatter frontier than those using information about heterogeneity. As a result, the "price" of an additional score increase to one group is much more expensive if the social

planner relies only on information from standard value-added measures. This has direct implications for average test scores, as seen in Appendix Figure A.5. Here we depict the change in average scores generated from moving from the optimal allocation attained using standard value-added to the optimal allocation attained using our heterogeneous estimates. Rather than being U-shaped like the welfare gains, these suggest an M-shape where the score gains are biggest when on these flat regions of the interior PPF, but away from the center where average scores (and thus class sizes) are all that matter.

In summary, comparative advantage and distributional goals are both potentially important to consider, but how each effect interacts with a policymaker's welfare weights means one effect may play a much bigger role for a given policymaker. Redistribution is important when the social planner has very strong preferences for gains to one group relative to another; however, the standard measures of value-added are able to capture most of these gains because value-added heterogeneity is positively correlated within teachers. There is little scope for welfare gains from comparative advantage. Conversely, when a policymaker values gains to each group roughly equally, there is little scope for distributional gains to matter, but significant scope for welfare gains from comparative advantage. Since policy suggests some social objectives may be more nuanced, we also turn our attention to the implications of our reallocations for achievement gaps and the creation of winners and losers.

F.2 Achievement Gaps

Considering Fiugre **??**, the baseline gap between students who scored above and below the median last year is 1.27σ in reading and 1.19σ in Math. A social planner focused on raising lower-scoring students' scores without, on average, hurting higher-scoring students could shrink those gaps by 4.4 and 7.6% *every year*. The gap between Black and Hispanic students versus white and Asian students are smaller: at 0.72σ in reading and 0.63σ in Math, and these gains could be reduced by 6.5% and 9.7% per year. These changes are strikingly similar to those in Delgado (2023) where allocations are made to explicitly shrink racial gaps in math scores subject to not lowering average scores. Delgado (2023) finds a 0.068σ reduction in the racial gap with no change in average scores, but using a race blind policy our district-wide reallocations would shrink the gap by 0.064 and *raise* average test scores by 0.032σ.[27]

There are three additional points we want to highlight from this figure with implications for which gaps are effected. First, whereas both the within-school and district-wide

[27]Note that in our context larger reductions in gains are obviously possible if the social planner is willing to choose allocations that actually reduce the average scores of certain groups while staying on the frontier. While it is likely that there are interior allocations in which gaps could be further reduced, we restrict our focus to allocations that are on the frontier of gains to higher- and lower-scoring students.

reallocations could change the achievement gap, only the district-wide reallocations could meaningfully affect the racial achievement gap. This makes sense because there is more variance in racial composition across schools than within.

Second, it is interesting to note that the welfare weights that hold gaps constant vary a lot across allocations. For the within-school reallocations attaining similar gaps requires a weight on lower-scoring students between 40-43% for reading and 52-53% for Math. On the other hand, the district-wide reallocations require much larger weights on lower-scoring students. For example, it takes 55% and 61% to shrink the achievement gaps in reading and math, and even more to shrink the racial gaps: 64% and 72%. For context, this means that to control the racial-achievement gap in math, a social planner would have to forego 0.007σ in average gains.

Finally, although average-test-score maximizing reallocations ($\omega_L = \omega_H = 0.5$) within school tend to not affect either gap significantly,[28] district-wide reallocations to maximize test scores will actually expand both the achievement and racial achievement gaps. Intuitively this is because of cross-school co-variation in achievement (or race) and class size as discussed above.

F.3 Winners and Losers

As noted above, because there are so many students, no reallocation—even one creating large average gains—is a Pareto gain in the sense that it helps, or leaves unaffected, all students. Despite the net gains from matching teachers to their comparative advantages and putting stronger teachers in larger classes, reallocations will assign some students to less effective teachers or to teachers who are a worse match for them (despite the teacher being a better match for their class).

Before communicating these results, we want to highlight the fact that *any* allocation of teachers to students will assign some students better teachers than others. In that sense the "harms" presented here should be benchmarked by the fact that in the status quo roughly one third of students are assigned to a teacher with below-median value each year (among teachers teaching the relevant grade in the student's school), and for these students, the average "loss" (relative to the expectation) is about 0.10 student standard deviations in their scores on tests of each subject.

With that context in mind, Appendix Figure A.8 shows that just as some students experience lower test score growth because of the year-to-year allocations of teachers in the status quo, some also receive lower value-added teachers in our reallocations. For example, the optimal within-school reallocations assign between 35-38% of students to lower value-added

[28] In fact, if anything they would slightly shrink the achievement gap.

teachers, with 39-47% for the district-wide reallocations. Unsurprisingly, more egalitarian allocations reduce the achievement gains of higher-scoring students relative to the status quo whereas more elitist allocations reduce the gains to lower-scoring students. Appendix Figure A.8 also reports the average achievement loss among students who are harmed. In the optimal district-wide (within-school) allocations, students who receive lower value-added teachers than they would in the status quo experience 0.104-0.120σ (0.085-0.099σ) smaller reading testing gains on average and 0.173-0.204σ (0.140-0.165σ) smaller math gains on average, per year. While these figures sound large in terms of educational interventions, it's important to remember that they are relatively similar to the "losses" that are occurring in the status quo. Our reallocations change which students receive teachers with lower absolute advantage or poorly matched comparative advantage, but on average these changes are more than offset by even larger average gains to other observably similar students.

One implication of this depiction of winners and losers is that our reallocative policies have a strong redistributive component. For a social planner who only cares about higher- versus lower-scoring students this consideration is irrelevant, but in practice districts may want to preserve some horizontal equity.[29] For example, because our reallocations tend to put teachers with higher absolute advantage in larger classes and because larger classes tend to be in schools with more higher-scoring students, our optimal reallocations will tend to benefit lower-scoring students in these schools slightly more than lower-scoring students in schools with lower average achievement. As discussed in Section 2, this may be troubling if the policymaker has preferences over multiple dimensions of student characteristics. For example, this could be problematic if the policymaker is most concerned about lower-scoring students in schools with lower achievement.

The fact that there are indeed winners and losers among students, in addition to the observation that teachers, administrators, and teachers' unions—by revealed preference—weakly prefer the status quo to any reallocation raises the question of welfare implications from these reallocation policies. Can schools reallocate teachers in ways that matter for welfare? How could they make such reallocations incentive compatible for families and teachers? What would be the cost of smoothing such incentive compatibility constraints? And would the reallocation still be worth doing? These are questions we consider in the following section.

[29] At least relative to the status quo. In an obvious sense, the opportunity cost of the current allocation is that it harming (or at least not benefiting) many students that a different allocation could be making better off.