Effect Heterogeneity and Optimal Policy: Getting Welfare Added from Teacher Value Added

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

September 8, 2023

Abstract

Though ubiquitous in research and practice, mean-based "value-added" measures may not fully inform policy or welfare considerations when policies have heterogeneous effects, impact multiple outcomes, or seek to advance distributional objectives. In this paper we formalize the importance of heterogeneity for calculating social welfare and quantify it in an enormous public service provision problem: the allocation of teachers to elementary school classes. Using data from the San Diego Unified School District we estimate heterogeneity in teacher value-added over the lagged student test score distribution. Because a majority of teachers have significant comparative advantage across student types, allocations that use a heterogeneous estimate of value-added can raise scores by 34-97% relative to those using only standard value-added estimates. These gains are even larger if the social planner has heterogeneous preferences over groups. Because reallocations benefit students on average at the expense of teachers' revealed preferences, we also consider a simple teacher compensation policy, finding that the marginal value of public funds would be infinite for bonuses of up to 14% of baseline pay. These results, while specific to the teacher assignment problem, suggest more broadly that using information about effect heterogeneity might improve a broad range of public programs—both on grounds of average impacts and distributional goals.

[*] Department of Economics, University of California, San Diego: teastmond@ucsd.edu

[†] Department of Economics University of Michigan: njmather@umich.edu

[‡] National Bureau of Economic Research: ricksmi@umich.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, 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

When evaluating policies, programs, and institutions researchers often rely on mean impacts. While means are powerful summary measures, they can also mask economically important information. This paper seeks to understand how measuring heterogeneity can more fully inform welfare measures and better optimize policy choices. We ask two main questions. (1) Theoretically, when does heterogeneity (in effects, outcomes, and social preferences) matter for maximizing a social objective? (2) Empirically, how large are the welfare gains from using heterogeneous rather than average estimates of impacts to evaluate and refine public policy?

Although these questions have many applications, we explore them in the context of value-added scores for elementary school teachers. Many have used value-added scores (regression adjusted means) to measure the effects of teachers and schools (see reviews in Angrist et al., 2022; Bacher-Hicks and Koedel, 2022); 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). We choose the elementary school setting because of mounting empirical evidence that value-added scores are both *multidimensional* and *heterogeneous* in the education context. For example, teachers affect student outcomes in multiple dimensions such as math and reading scores (Condie et al., 2014), attendance and suspensions (Jackson, 2018), and work ethic and learning skills (Petek and Pope, forthcoming). Furthermore, teachers also have heterogeneous effects on different types of students defined by factors such as race and gender (e.g., Dee, 2005; Delhommer, 2019; Delgado, 2022) and socioeconomic status (Bates et al., 2022). Similar patterns have been found in health-related value-added (e.g. Hull, 2020).

This paper applies and extends insights from theoretical welfare economics to overcome the limitations that arise from multidimensionality and heterogeneity, allowing us to empirically evaluate the optimal allocation of teachers to classes based on this information. The critical issue from a social welfare perspective is that in the presence of multidimensionality and heterogeneity, value-added measures only partially order the welfare of an allocation of teachers to students. Intuitively, this is because of ambiguity about whether the definition of a "better" teacher should prioritize gains in math versus reading scores or gains for high-achieving versus low-achieving students (See the impossibility-like results in Condie et al., 2014). Fortunately, whereas research in value-added has identified these problems, research in public finance has a long history of using welfare functions to aggregate over the heterogeneous effects of policies. We extend such insights from welfare economics for two purposes. First, we characterize the shortcomings of relying on mean-oriented measures of policy effects such as standard value-added to make welfare considerations in general. Then the bulk of the paper evaluates the optimal allocation of teachers to classes using measures of heterogeneous value-added that produce scalar, welfare-relevant statistics.

Our theoretical results show two ways that ignoring effect heterogeneity can lead to inaccurate inference about both policy counterfactuals and how policy can be improved. First, bias arises when mean effects are not externally valid to match effects from the policy. For example, imagine a medical treatment that did not have serious side effects in the population in general. If we are considering a policy that would target this treatment to new high-risk patients, it is not clear whether the impact will be the same. Second, bias also arises from the covariance across the target population of the heterogeneous effects of a policy and an individual's welfare weights. For example, consider a tax reform that raises post-tax incomes by \$3000 to the richest 50% of households but reduces incomes by \$1000 for the poorest 50% of households. Policymakers may consider this reform undesirable for equity reasons even though it increases average incomes. These biases can both be reduced or eliminated by estimating conditional average treatment effects along appropriate observable dimensions and allowing for heterogeneous welfare weights. When optimizing policy, correcting this bias can lead to significant gains through comparative advantage and allow policymakers to direct interventions towards people with the highest marginal welfare benefit.

These theoretical results highlight an interesting contribution of our paper. As empirical policy evaluations become increasingly common, our theoretical results characterize the trade-offs implicit in relying on mean impacts. For example, using mean effects to predict the welfare of an allocation is biased in general because welfare depends not just on program impacts and welfare weights but the covariance of the two. Interestingly, this insight is reminiscent of similar results in optimal corrective taxation of heterogeneous consumption externalities (like alcohol). Griffith et al. (2019) show that the optimal corrective tax is the average consumption externality *plus* the covariance between individual contributions to the externality (the effect) and demand elasticities (the weight). Furthermore, in the externality context, conditioning (in this case tax differentiation by product) also reduces the bias, as it can in our setting.[1] The importance of heterogeneity and conditioning in these theoretical settings raises questions about whether using average "sufficient statistics" is appropriate when heterogeneous estimates could inform differentiated policies like corrective taxation of heterogeneous *production* externalities (Hollingsworth and Rudik, 2019; Fell et al., 2021;

[1] The second insight is technically a generalization of the first, which was originally suggested in Diamond (1973).

Sexton et al., 2021). Crucially, we speak to these trade-offs by showing how both biases can be reduced by estimating conditional average treatment effects along observable dimensions to allow for heterogeneity in impacts.

Motivated by the importance of heterogeneity in general, we estimate heterogeneity in teacher value-added along the achievement distribution in the San Diego Unified School District, the second largest district in California. We find large gains from using heterogeneity to more optimally allocate teachers to students. In particular, we use the methods pioneered by Delgado (2022) to estimate the value-added of all third- through fifth-grade teachers on student math and English language arts (ELA) scores allowing for heterogeneous effects on students who had above- and below-median scores the previous year. Although these measures of value-added are correlated with standard (i.e. homogeneous value-added) measures, we find substantial heterogeneity. For example, the average within-teacher difference in value-added across groups (i.e. comparative advantage) is as large as 53% (48%) of a standard deviation in mean value-added for ELA (math). We use these estimates to consider welfare gains from two sets of possible policies: reallocating teachers to classes without changing school assignment or allowing for school reassignment.[2] There are enormous gains from reallocation. Over the course of third to fifth grade, using heterogeneous measures of value-added to improve district-wide teacher assignments could raise student math scores by 0.17 student standard deviations on average and ELA scores by 0.12. For context, both changes are roughly equivalent to an intervention improving all teachers' value-added by 30% of the (teacher) standard deviation in the relevant subject.

In this process, our paper makes three innovative contributions to the literatures on value-added and teacher value-added. First, we demonstrate how important achievement is as a dimension of effect heterogeneity in our education context. Whereas many papers have found evidence of "match effects" between students and teachers sharing observable characteristics like gender or race (Dee, 2005; Delhommer, 2019), other results reveal that these match effects only explain part of the heterogeneity in teacher effects on the same dimensions (Delgado, 2022). Our results suggest that focusing on demographic match is incomplete because it overlooks how instructional differentiation along the achievement distribution (well documented in the education literature) interacts with these characteristics. This insight reflects other evidence from health economics that in general lagged outcomes are one of the most important dimensions for match effect heterogeneity (as in Dahlstrand, 2022).

Second, our results highlight how combining information from multiple outcomes sub-

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

stantially improves the welfare gains from reallocations. 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. To this end, we aggregate teacher effects using estimates of the differential impact of elementary school gains in math and ELA on lifetime earnings from Chetty et al. (2014b). Back of the envelope calculations suggest that over three years the allocation of teachers that maximizes present-valued lifetime earnings would generate over $4000 in present valued earnings per student or over $83.7 million in total.[3] Whereas interventions in the education literature have often focused on math scores for a variety of reasons (Chetty et al., 2014a; Delgado, 2022; Bates et al., 2022; Ricks, 2022), our contribution is accounting for the separate marginal effects of math and reading outcomes, which generates 34% larger wage impacts (value-added of $21 million) relative to focusing only on math.

Third, these results have implications for the discussion of using value-added in teacher (and doctor and hospital) compensation and extend our understanding of the welfare implications of such policies. Motivated by the large 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 SDUSD teacher with 10 years of experience). For 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. These ideas combine insights from two literatures on teacher labor markets: one focusing on dismissal (Hanushek et al., 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., 2022, 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.

Taken together, our results highlight the first-order importance of considering heterogeneity in empirical welfare analysis. In our theory we show how the gains possible from allocations based on heterogeneous effects may be 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 only using the standard value-

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

added measure. While the critical role of comparative advantage has been acknowledged for centuries, our contribution to welfare theory is in connecting treatment effect heterogeneity, comparative advantage, and social preferences. These connections capture and formalize the growing understanding that heterogeneity is a key consideration for allocating scarce resources according to a social objective by means of targeting. This has been explored theoretically (Kitagawa and Tetenov, 2018; Athey and Wager, 2021) and is reflected in a recent explosion of empirical inquiry about 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 suggest that in these settings and others ignoring heterogeneity may have serious welfare ramifications and that considering heterogeneity in effects and social preferences presents a clear path forward for future welfare 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 contains our estimation procedure and a description of value-added in the San Diego Unified School District. Section 4 leverages our welfare theory to explore the reallocation of teachers to classes and measures the welfare gains from using information about heterogeneity. Finally, Section 5 draws the pieces together to explore the implications for welfare and Section 6 concludes.

2. A Welfare Theory of value-added

This section formalizes the implications of estimating mean-oriented statistics for use in welfare analyses and the benefits of estimating heterogeneous impacts. We begin by showing how a welfare-theoretical framework can allow a social planner to aggregate over multidimensional policy impacts on a heterogeneous population. Second, we show how relying on average effects and average welfare weights can lead to biased welfare estimates. This bias has two sources: average treatment effects have imperfect external validity in different allocations (for example assigning teachers to classes with different compositions), and average welfare weights ignore heterogeneous gains to groups with different welfare weights (for example, differential valuation of an identical test-score increase for struggling versus advanced students). Third, we show how measuring heterogeneity along key dimensions can minimize the bias. Finally, we show graphically how correcting this bias leads to better policy optimization through comparative advantage and targeting interventions towards the recipients with the highest marginal benefit.

2.1 Welfare with Heterogeneity and Multidimensionality

Consider a social planner selecting a policy $p \in \mathcal{P}$. This policy could be assigning teachers to classes (our application), defining an eligibility threshold for a means-tested program like health insurance, or choosing between various public works projects. The welfare under policy p is a function of the lifetime utilities U_i^p and welfare weights ϕ_i^p of each person i under each policy p. With a population of size n welfare is

$$\mathcal{W}^p = \sum_{i=1}^{n} \phi_i^p U_i^p$$

If the policy p has heterogeneous effects on utility for different people, using welfare weights ϕ_i^p is a long-standing method to allow the social planner to aggregate over individuals and recover a scalar measure of welfare.

In practice neither policymakers nor economists observe lifetime utility directly. Instead, they usually rely on observable outcomes Y like earnings, health outcomes, or test scores as proxies. We let the social planner evaluate policies using a "score function" $S_i^p = s(\boldsymbol{Y}_i^p, \boldsymbol{X}_i)$ which produces an individual-level score for the policy based on observable outcomes and characteristics. Note that while this score could represent any social objective, identifying the expected lifetime utility or earnings would be particularly useful in many cases (see the related work on surrogate indices by Athey et al., 2019). Just as the welfare weights allow the social planner to aggregate over the heterogeneous effects of the policy, the score function allows the social planner to aggregate over the multidimensional effects of the policy.

Under this setup, a policymaker can evaluate each policy p based on observable outcomes. Assuming an individuals' outcomes $\boldsymbol{Y}_i^p$ only impact their own utility and weights, the expected change in welfare from the status quo ($p = 0$) to policy p is

$$\Delta \tilde{\mathcal{W}}^p \equiv \sum_{i=1}^{n} \gamma_i(S_i^p, S_i^0) \Delta S_i^p \tag{1}$$

where $\gamma_i(S_i^p, S_i^0)$ is a new welfare weight and ΔS_i^p is the effect of policy p on individual i's score. The weight γ_i^p reflects the average welfare gain from marginal score changes over $[S_i^0, S_i^p]$, incorporating the change in expected utility and the relevant welfare weights, ϕ_i^p. A detailed explanation of this derivation can be found in Appendix B.1.

Unfortunately, estimating this welfare metric has a major complication: The effects of the policy ΔS_i^p and the proper weights γ_i^p are both individual specific. The impact of the policy on the score, ΔS_i^p, and the impact of the score on lifetime utility, γ_i^p, may both vary from student to student. Even though these individual-level measures provide a more accurate theoretical framework, using individual welfare weights and individual outcomes to assess policy is typically not feasible. Because of this limitation, policies are often evaluated with aggregate measures. We now characterize the bias that this aggregation produces and how estimating heterogeneous effects can reduce that bias.

2.2 Bias from Ignoring Match Effects or Individual Welfare Weights

Empirical analyses often simplify the weights and treatment effects to means in order to measure welfare. This approach multiplies an estimate of the average treatment effect of a policy $\widehat{ATE}^p$ with the average welfare weight for the impacted population (see intuition in Hendren and Sprung-Keyser, 2020). Assuming the average welfare weight is known $\mathbb{E}[\gamma^p] = \frac{1}{n}\sum_{i=1}^{n}\gamma_i(S_i^p, S_i^0)$, this approach allows for two sources of bias.[4] First, because the true ATE^p is rarely known (and never known *ex ante*), other estimates such as rules-of-thumb and estimates from different times or populations are used. For example, in the value-added setting a teacher's average impact on a different class in the past is often used to infer their impact on another class in the future, introducing bias. Second, as shown in Appendix B.2, the welfare weights that convert a true ATE^p into welfare are a function of the joint distribution of the individual-level treatment effects and individual welfare weights. By instead using the simple population mean $\mathbb{E}[\gamma^p]$, more bias is introduced. In general, these simplifications lead to a biased measure of welfare:

Theorem 1. If welfare is estimated using the product of an average outcome from a different population $\widehat{ATE}$ and an average welfare weight $\mathbb{E}[\gamma^p]$, then the estimate will contain the following bias relative to the more general benchmark in Equation 1:

$$\begin{aligned}\textbf{Average Bias}_{ATE} &= \frac{\Delta\tilde{\mathcal{W}}^p}{n} - \mathbb{E}[\gamma^p]\widehat{ATE} \\ &= \mathbb{E}[\gamma^p]\left(\mathbb{E}[\Delta S^p] - \widehat{ATE}\right) + \mathrm{Cov}(\gamma^p, \Delta S^p)\end{aligned}$$

Proof in Appendix B.3

With the equation for the bias in hand, we see that these common simplifications lead to two sources of bias. First, one source of bias comes from the difference in the expected change in our outcome of interest, and the $\widehat{ATE}$ estimate used. While these statistics could differ for any reason relating to the external or internal validity of our estimate, our paper is most interested in a specific concern with external validity: Whether averages of heterogeneous

[4] In practice the average welfare weight needs to be estimated as well, which could introduce a third source of bias, so we assume that policymakers have prior knowledge about the average welfare weight.

effects apply in different populations. For example, if teachers have heterogeneous impacts on students, then estimating the average treatment effect on their current class will not give an unbiased estimate of their average impact on a class of very different students. If, for example, we change the class composition to better match the teacher's comparative advantage, their average impact will increase. A more formal explanation of this impact can be seen in Appendix B.4.

Second, using the population average welfare weight ignores any covariance between welfare weights and treatment. While not the case in general, there are some situations where the covariance would be zero. For example, when the effects of a policy are uniform (or random) there can be no covariance. Perhaps more relevant to policy the covariance will also be zero when there is no variation in welfare weights among the impacted population. This may approximately hold, for example, for targeted programs like SNAP, Medicaid, and TANF. The covariance is likely to matter in many other settings. For example, in our setting teacher reassignment has the potential to disproportionately help low-performing students. If low-performing students have higher welfare weights, the covariance term in the bias would be positive and means would understate the value of the reallocation.

2.3 The Case for Estimating Heterogeneity

Measuring heterogeneous impacts along key dimensions can lower the bias outlined above. By choosing features that explain the most variation in welfare weights and policy impacts, we may be able to lower the bias significantly. In practice, this method requires estimates of the conditional average treatment effect and welfare weights by subgroup ($\widehat{CATE}(x)$ and $E[\gamma^p|x]$) rather than using average treatment effects and weights. Incorporating this, the bias can be characterized in the following way:

Theorem 2. If mean welfare is estimated using the weighted mean of a conditional average treatment effect $\widehat{CATE}(x)$ and a conditional average welfare weight $E[\gamma^p|x]$ weighted by the fraction of the population with characteristic x, P_x, the mean welfare estimate will contain the following bias:

$$\begin{aligned}\textbf{Average Bias}_{CATE} &= \frac{\Delta\tilde{\mathcal{W}}^p}{n} - \sum_X P_x E[\gamma^p|x]\widehat{CATE}(x) \\ &= \sum_x P_x \left(\text{Cov}(\gamma^p, \Delta S^p|x) + E[\gamma^p|x]\left(\mathbb{E}[\Delta S^p|x] - \widehat{CATE}(x)\right)\right)\end{aligned}$$

If the features in x are chosen carefully, both portions of the bias can be lowered while still being identifiable. To be more precise, we will again consider the two bias terms separately and compare them to the unconditional counterpart in Theorem 1.

First, consider the covariance terms. The covariance term in Theorem 1 has been replaced by the weighted sum of conditional covariance terms. Using the law of total covariance, we can see that this portion of the bias will be smaller after conditioning, when

$$\left|\sum_X P_x \mathrm{Cov}(\gamma^p, \Delta S^p|x)\right| < \left|\sum_X P_x \mathrm{Cov}(\gamma^p, \Delta S^p|x) + \mathrm{Cov}(\mathbb{E}[\gamma^p|x], \mathbb{E}[\Delta S^p|x])\right| = \left|\mathrm{Cov}(\gamma^p, \Delta S^p)\right| \tag{2}$$

This means that when the average within group covariance between γ^p and ΔS^p is smaller than the total covariance, the bias will be reduced. The middle term breaks up the total covariance into two parts. The first term is the within group covariance, and the second is the covariance of the group means. To better connect these terms to applications, it is helpful to think through cases. First, if both of these terms are the same sign, the condition will be met. Consider a case where we condition on pre-test scores, like our paper, but race also impacts γ and is not conditioned on. If the gains from a teacher allocation are positively (or negatively) correlated with both the welfare weights on both pre-test scores and race, the condition is met. Now suppose they are opposite signs. That is, the gains are positively associated with test score and negatively associated with the welfare weights on race or visa-versa. In this case, the inequality may or may not be satisfied. It will still be satisfied when

$$2 * \left|\sum_X P_x \mathrm{Cov}(\gamma^p, \Delta S^p|x)\right| < \left|\mathrm{Cov}(\mathbb{E}[\gamma^p|x], \mathbb{E}[\Delta S^p|x])\right| \tag{3}$$

Put simply, this holds when the within group covariance is small relative to the group mean covariance. In keeping with our example, the within group covariance would be small if the unconditioned feature, race, either does not impact γ^p very much after conditioning on pretest scores, has little association with ΔS^p after conditioning on pretest scores, or their relationship happens to be randomly distributed after conditioning on pre-test scores. The group mean covariance will be large if the conditioned factor, pre-test-scores, plays a large role in the relationship between γ^p and ΔS^p. For example, suppose pre-test groups with large welfare weights also see large test score gains because teachers are sorted according to their comparative advantage along the pre-test dimension.

Now to consider the second term. As before, this could come from any external or internal validity issue with $\widehat{CATE}(x)$, but we focus on the bias from population changes interacted with heterogeneous treatment effects. If a teacher has different impacts on different types of students, for example, and the class composition changes, their average impact will change. By conditioning on the observable, x, we can adjust for compositional and treatment effect differences over X. The new estimator takes a teacher's average impact on group x and weights that impact by the composition of their new class. The remaining bias, then, would need to come from differences in treatment effects along other dimensions and variation in composition within a group x across classes. Pulling out the terms, this will be smaller when the following holds.

$$\sum_x P_x E[\gamma^p|x]\left(\mathbb{E}[\Delta S^p|x] - \widehat{CATE(x)}\right) < \mathbb{E}[\gamma^p]\left(\mathbb{E}[\Delta S^p] - \widehat{ATE}\right) \tag{4}$$

A more formal treatment can be seen in Appendix B.5.

Putting these ideas together, there are two special cases that are helpful to think through. first, the case where welfare weights really only depend on x. For example, if x is pretest scores and the policymakers want to treat every student with the same pre-test score equally. In this case, the first term goes to zero since there is no covariance within test score groups. There could still, however, be differences in treatment effects and class composition within a test score group x. For example, if teachers have differential impact by race (Delgado, 2022). This would lead to a non-zero value for the second term. If there is no heterogeneity within x, either because the treatment effects are the same or the class compositions are the same within x, the second term would also be zero and we would have a completely unbiased estimator. These special cases help to highlight how the first term is driven by the policymaker's re-distributive preferences while the second is driven by the heterogeneous treatment effects and compositional differences between sup-populations.

Given these differences, it is worth noting that there is no reason one could not condition the welfare weights and the estimates on different subsets of $\boldsymbol{X}$. for example, $E[\gamma^p|x_1]$ $\widehat{CATE(X_2)}$. It might be the case that a variable is not meaningful in the welfare weight, but is a factor in estimating an accurate treatment effect. While this could be done, we focus on the case where the same variable, pre-test scores, is being considered for both.

2.4 Graphical Intuition of the Welfare-Relevant Components

Having illustrated how to reduce bias for welfare estimates of a given policy intervention, this section considers the welfare gains from decreased bias when comparing different policies. We present a simple example with two groups to show how heterogeneous estimates allow welfare improvements relative to evaluations based on means. For simplicity of exposition, we assume that all effect heterogeneity and heterogeneity in social preference relates to these two groups. This highlights three channels for gains from reallocations—some of which are only possible by estimating heterogeneity.

We illustrate these three channels for improving welfare in Figure 1. The two axes of Figure 1 depict the average change in the score function for two groups. In our example it would depict the average change in math scores for lower- and higher-scoring students. Connecting these two axes are two production possibility frontiers (PPFs—depicted as curves). Allocations between the origin and "PPF: ATE" are possible by using information about mean effects that capture absolute advantage—such as a teacher's average test-score value-added on students.[5] In our setting this would mean assigning teachers with higher overall value-added to larger classes, and teachers with lower value-added to smaller classes. Allocations within the "PPF: $CATE$" are possible by using information about heterogeneous effects that capture both absolute and comparative advantage. In our setting this would mean also assigning teachers to classes with larger shares of the group they have a comparative advantage in teaching. This PPF is at least weakly dominant because it allows for additional gains from matching teachers to classes in ways that leverage their heterogeneous value-added across student groups.

Now consider a policymaker with indifference curves corresponding to the dotted lines. The slope of these indifference curves indicates the relative preferences given to one group versus the other. In this example, the slope is higher than -1, indicating that the policymaker places greater weight on group 1. Figure 1 presents the status quo and three possible reallocations (a white box and colored circles) and their corresponding welfare (indicated with dashed indifference curves).

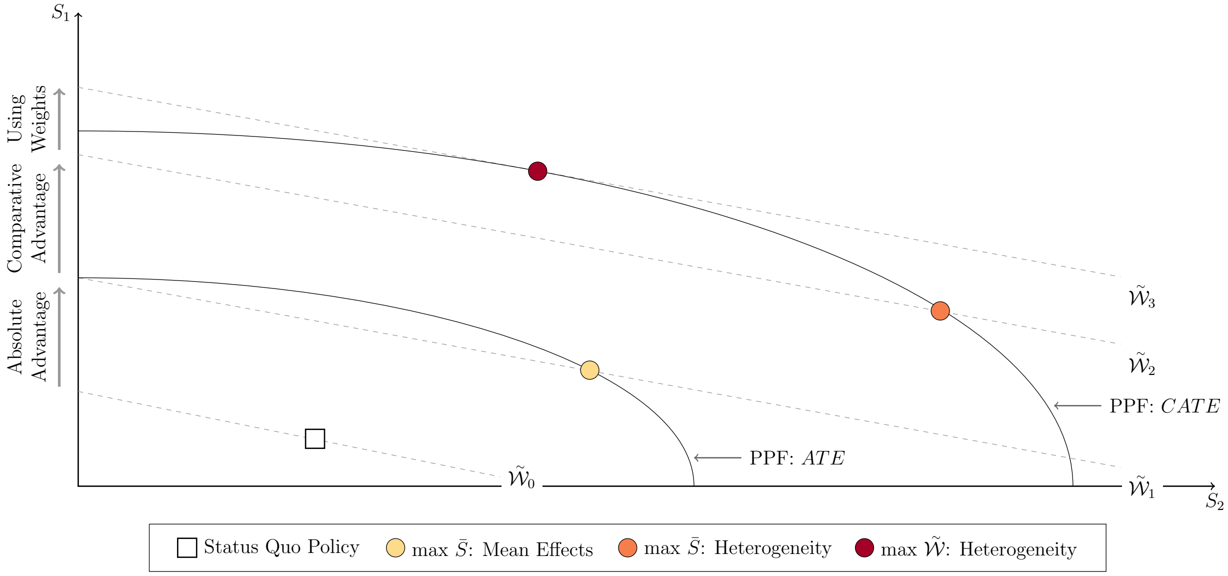
First, a policymaker trying to maximize test scores (despite having re-distributive goals) using standard value-added measures can experience welfare gains from the absolute advantage of teachers. Figure 1 represents this reallocation as a movement from the white box to the yellow circle on PPF: ATE with welfare gains corresponding to a move from $\tilde{\mathcal{W}}_0$ to $\tilde{\mathcal{W}}_1$[6]. This movement reflects the gains from making allocations based on absolute advantage.

Second, a policymaker maximizing test scores with heterogeneous estimates of teacher value-added (but still ignoring their re-distributive preferences) can experience further gains from the comparative advantage of teachers. With heterogeneous estimates, the policy makers can assess how a teacher would impact students in each group in addition to students

[5] Technically, a valid value-added estimator is only a consistent estimate for this parameter as the set of students a teacher teaches approaches a representative sample.

[6] Note that, in our case, for these gains to be non-zero, two things must be true: it must be the case that (1) some classes have different sizes, and that (2) some teachers have different value-added scores. If these conditions are met a policymaker would expect to increase the scores for students in both groups by assigning higher-value-added teachers to the larger classes. Such reallocations can lead to meaningful impacts in the real world setting we use, where class size averages about 27 with a standard deviation of about 6.

Figure 1: Absolute Advantage, Comparative Advantage, and Social Preferences Contribute to Welfare



Note: This figure illustrates the welfare gains allocations using heterogeneous effects and welfare weights. The two axes present the outcome score of interest, S, for individuals of two types. The graph contains two production possibility frontiers and some indifference curves. The interior production possibility frontier is attained by allocations made with the constant-effects model, like traditional value-added measures. These mean estimates could enable welfare gains from allocations based on the absolute advantage (possibly weighted by social preferences). The second, dominant frontier is attained by allocations using information about effect heterogeneity and, thus, comparative advantage. The indifference curves show the welfare value of four allocations: (1) the status quo, (2) the average-score maximizing allocation using mean effects, (3) the average-score maximizing allocation using heterogeneous effects, and (4) the welfare maximizing allocation using heterogeneous effects.

on average. This knowledge would allow them to reallocate teachers based on absolute and comparative advantage, indicated as a movement from the white box to the orange circle on PPF: CATE with welfare gains corresponding to a move from $\tilde{\mathcal{W}}_0$ to $\tilde{\mathcal{W}}_2$.[7] Compared to the allocation on PPF: ATE, the gains from $\tilde{\mathcal{W}}_1$ to $\tilde{\mathcal{W}}_2$ reflect the additional gains from making allocations based on comparative advantage.

Finally, a policymaker can produce further welfare gains by directly considering their distributional goals. In our example, the policymaker wants to focus on lower-scoring students for educational remediation (although a focus on higher-scoring students, perhaps for prestige, is also possible). If this is the case, both score-maximizing allocations are suboptimal. This loss is visualized in Figure 1 where the indifference curves at $\tilde{\mathcal{W}}_1$ and $\tilde{\mathcal{W}}_2$ are not tangent to either PPF. As such, the policymaker can increase welfare by trading off the possible test-score gains for one group against gains to the other groups. The optimal consideration moves them to the red point, with the largest welfare of $\tilde{\mathcal{W}}_3$.

Although each of these pieces could generate large welfare gains in theory, whether there are meaningful gains from estimating heterogeneity in practice remains an empirical question. For example, if teacher effects are homogeneous or highly correlated there would be no gains from making allocations based on comparative advantage. Furthermore, even if there are differences or distributional objectives, if the status-quo allocation already takes them into account, there would be no gains from reallocations since the welfare gains have already been captured. The remaining sections of the paper measure the amount of heterogeneity in teacher impacts and describe the welfare effects of possible reallocations.

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

Having established how measuring effect heterogeneity could be useful for informing welfare and policy, this section sets the groundwork for determining to what extent heterogeneity in teacher value-added matters in practice for the allocations of teachers to classes in elementary school. 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 how it is at play in the status quo allocation of teachers to classes.

[7] Note that, in our case, for these gains to be larger than the gains from absolute advantage, two more things must be true: it must be the case that (1) some classes have different compositions of student types, and (2) that some teachers have different value-added on each type of student. If these conditions are met a policymaker would expect to further increase the scores for students in both groups by assigning better matched teachers to classes.

3.1 Background and Administrative Data

To consider socially optimal allocations of teachers to classes, 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 1,816 teachers who are the main instructors in third, fourth, or fifth grade classes in the 2002-03 through 2012-13 school years.[8] We link all teachers to their students each year and we restrict our attention to students with test scores in both English Language Arts (ELA) and math for two consecutive years. This leaves us with 196,452 student-year observations in 10,447 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 statistics and information about the current allocation of teachers to classes in Section 3.4.

3.2 Estimation Overview

We use the data from San Diego Unified to evaluate the importance of estimating heterogeneity in optimally assigning teachers to classes. While there are many dimensions over which we could estimate heterogeneous effects, we focus on lagged student scores. Specifically, we estimate the value-added of each teacher on the Math and ELA scores of students with below-median scores (lower-scoring students) and students with above-median scores (higher-scoring students). Our theory suggests that to be welfare improving the dimension we choose should capture a lot of the variance in impacts and be relevant to the social planner. We estimate heterogeneity along the achievement distribution because it meets these criteria.

First, measuring heterogeneity in teachers' effects on lower- and higher-scoring students captures the most salient dimension of instructional heterogeneity. This intuition is not just based on anecdotes; indeed, the large education literature about instructional differentiation suggests that teaching lower- and higher-scoring students requires very distinct skills. See for instance the large literature on differentiated instruction (see Betts, 2011; Duflo et al., 2011; Tomlinson, 2017, for review and examples). Furthermore, while many papers have found evidence of "match effects" between students and teachers sharing observable characteristics like gender or race (Dee, 2005; Delhommer, 2019), results from Delgado (2022) shows that these match effects only explain part of the heterogeneity in teacher effects on students of different genders and races. This suggests that focusing on demographic match may be overlooking something key. We suggest that the most relevant dimension is related to differentiation along the test-score distribution.

[8] We limit to these years because the state-mandated tests were stable and comparable over these years.

Second, policymakers often expressly identify achievement as a dimension over which they have heterogeneous valuations of gains. For example, quintessential US policies like the federal No Child Left Behind Act of 2001 directly focused on accountability for and proficiency among lower-scoring students. The stated goal was to focus on raising the lower bound of student test scores, calling for corrective action based on whether the lowest performing groups met state standards.[9] At the same time, many national, state, and local policies promote gains to lower-scoring students while expressing nondiscriminatory, identical preferences for students of different genders, races, and socioeconomic statuses conditional on their achievement.

3.2.1 Standard value-added

For our traditional value-added estimates we follow the approach in Chetty et al. (2014a) and implement it with associated Stata package (Stepner, 2013). The details are presented in Appendix C, but the general approach has three steps. First, we estimate the effects of student i's characteristics in year t, $X_{i,t}$, on test scores in subject s, $S_{i,s,t}$, in a regression of the form:

$$S_{i,s,t} = \beta_s X_{i,t} + u_{i,s,t}$$

Second, we obtain the average of the residuals implied by β_s by class and year:

$$\bar{A}_s^{j,t} = \frac{1}{n_{j,t}} \sum_{i:\mathcal{J}(i,t)=j} \left[S_{i,s,t} - \hat{\beta}_s X_{i,t} \right]$$

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

$$\hat{\tau}_s^{j,t} = \hat{\boldsymbol{\psi}}_s \bar{\boldsymbol{A}}_s^{j,-t} \tag{5}$$

The main assumption necessary to interpret these estimates as causal effects is that class-level shocks and idiosyncratic student-level variation are conditionally independent and a stationary process (given the controls, $X_{i,t}$). It must also be the case that the variance in teacher value-added is stationary (as outlined in Chetty et al., 2014a, —again formal details are in Appendix C).

To the end of establishing this conditional independence, we follow the controls of Chetty et al. (2014a), documented to have unbiased estimates of teacher effects. In our setting $X_{i,t}$

[9] The fact that these policy objectives often find broad cross-partisan support could lead one to conclude that all policymakers have somewhat egalitarian preferences and that disagreements are not questions of direction but only magnitude.

includes cubic polynomials in prior year test scores in math and ELA, those polynomials interacted with student grade level, as well as controls for ethnicity, gender, age, the lagged percentage of days absent, indicators for past special education and English language learner status, cubic polynomials in class and school-grade means of prior test scores in both subjects (also interacted with student grade level), class and school means of all the other covariates, class size, and grade and year indicators.[10]

3.2.2 Heterogeneous value-added

For our estimates of heterogeneous value-added, we follow the approach pioneered in Delgado (2022) and applied in Bates et al. (2022), implemented with extensions we made to the Stepner (2013) Stata package. The details are also presented in Appendix C, but the general approach also has three steps. The first step is identical, with the addition of indicators for group g to $X_{i,t}$ We then obtain the average of the residuals implied by β_s by class, type, and year:

$$\bar{A}^{j,t}_{g,s} = \frac{1}{n_{j,t,g}} \sum_{i:\mathcal{J}(i,t)=j,g_i=g} \left[S_{i,s,t} - \hat{\beta}_s X_{i,t}\right]$$

Finally, we estimate leave-year-out (jackknife) measures of teacher impact by predicting $\bar{A}^{j,t}$ with the residuals in all other years using the observed auto-covariance.

$$\hat{\tau}^{j,t}_{g,s} = \hat{\boldsymbol{\psi}_{g,s}} \bar{\boldsymbol{A}}^{j,-t}_{s} \tag{6}$$

Here the main assumption necessary to interpret these estimates as causal effects is that, class-*type*-level and student-level variation are conditionally independent and stationary processes (as derrived in Delgado, 2022, —again formal details are in Appendix C). Note that we differ from Delgado (2022) in one way: We impose a zero-covariance assumption about the idiosyncratic teacher value-added components across groups, similar to the assumptions implicit in the measurement of value-added across subjects in both Chetty et al. (2014a) and Delgado (2022) for internal consistency.

3.3 Heterogeneity Highlights the Importance of Comparative Advantage

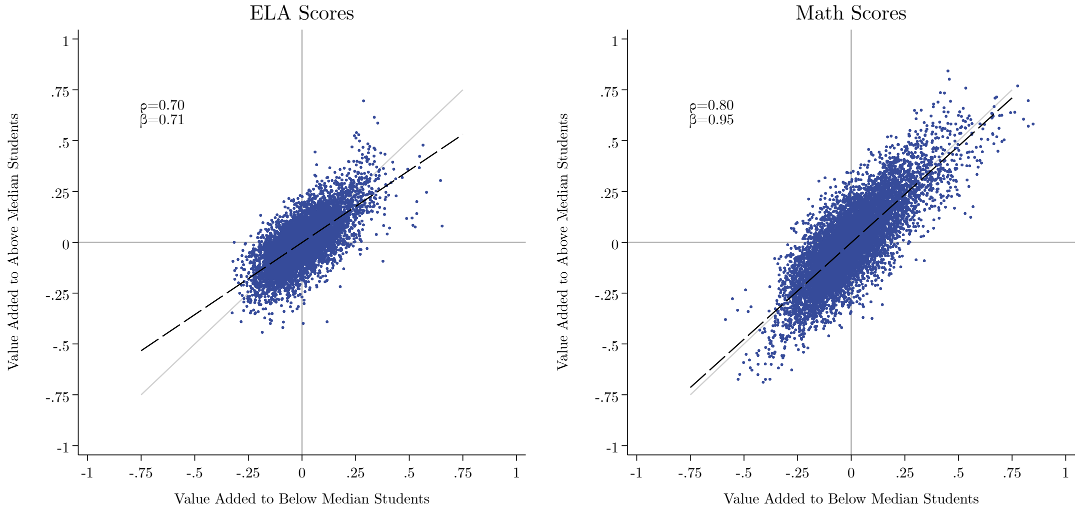
We use these techniques to estimate the heterogeneous effects of 1,816 teachers on 109,125 lower-and higher- scoring students from 127 elementary schools in SDUSD. These teachers

[10] The only notable difference from the controls in Chetty et al. (2014a) is their inclusion of information about free and reduced price lunch, which we omit in our research because of restrictions that SDUSD imposes on researchers' use of this information due to their perception of federal regulations on use of student level subsidy information.

taught grades 3-5 in the 2002-03 to the 2012-13 school years. In this section, the mean value-added is normed to zero for each group, reflecting both the economic intuition that for the average student the "outside option" for the teacher she or he has is the average teacher and the econometric identification argument in Chetty et al. (2014a) implicit in our identifying assumptions.

We depict the main value-added results in Figure 2. This Figure reports two scatter plots—one for ELA and one for math—where each point represents one teacher. The teachers 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 English Language Arts (ELA) and Math test scores. Each dot represents one teacher-year estimate of value-added on high- and low-scoring students. The correlation coefficients is for the entire population stacked by year. The dashed line shows the line of best fit with the slope reported. For reference a line with slope one is plotted in the background.

Visual inspection of Figure 2 illustrates the 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 by comparing teachers along the gray 45-degree line. Teachers above and to the right generate larger testing gains compared to teachers below and to the left. Comparative advantage can also be seen visually. Teachers with dots above the gray 45-degree line have a comparative advantage in teaching higher-scoring students, and teachers with dots below that line have a comparative advantage in teaching lower-scoring students. The size of the average comparative advantage is large: 53% the size of the cross-teacher standard deviation in standard teacher value-added for ELA and 48% for math.

The differences within and between teachers are what will generate gains for the reallocation exercises. We estimate that teacher value-added to higher- and lower-scoring students is correlated at 0.7 for ELA and 0.8 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., 2022) but larger than those by race (0.7 for math and 0.4 for ELA in Delgado, 2022). Furthermore, our theoretical framework suggests there is value in combining information from multiple outcomes. In that light, it is also worth noting that the cross-subject correlations are lower. For example, Figure A.1 shows that the cross-subject, cross-group correlations are both around 0.6, suggesting even larger gains from cross-subject comparative advantage.

It is also interesting to note that Figure 2 reveals that value-added to math is much more dispersed than value-added to ELA. This is consistent with evidence from similar value-added papers (e.g., Chetty et al., 2014a). Our results further show that teachers' value-added is more highly correlated across achievement groups for Math than for ELA. This is also consistent with absolute advantage being more important and variable with Math teaching than with ELA teaching.

3.3.1 Validation and Robustness

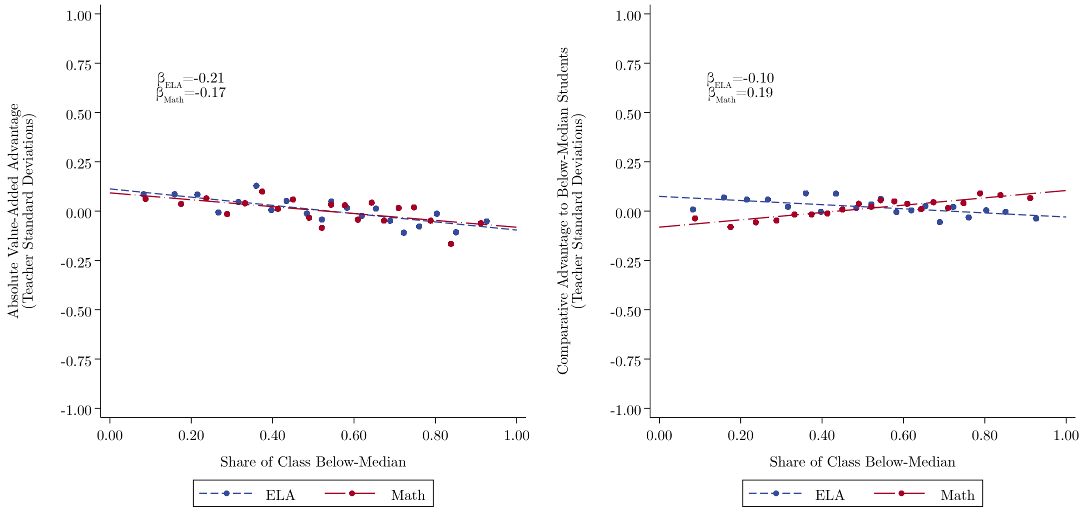
Although these results suggest striking patterns of comparative advantage, our reallocation exercises and welfare estimates would be meaningless if these estimates reflected idiosyncratic noise rather than persistent heterogeneity within and across teachers. Although the use of shrinkage assuages these concerns, we also perform three additional exercises demonstrating the stability and credibility of our heterogeneous estimates. Each result reinforces our confidence that the value-added scores are fitting systematic patterns in causal differences and not just idiosyncratic noise.

First, Appendix Figure D.6 reports 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 is between 0.78-0.90 for all estimates. Additionally, Appendix Figure D.7 leverages the longitudinal nature of our data to show that heterogeneous value-added estimates carry the same information about long term outcomes as traditional value-added estimates (Chetty et al., 2014b). These results show striking similarities between the effects of our estimates and traditional value-added. Furthermore, estimates for each student group are no less precise suggesting that the variance is loading on the dimension of heterogeneity we specified.

3.4 The Status-Quo Allocation of Teachers and Students

This section shows how teachers are allocated to classes in the status quo, whether this allocation is efficient or equitable, and presents descriptive evidence that there may be gains from reallocation. Figure 3 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 3: Teacher value-added Only Varies Somewhat with Class Composition



Note: This figure shows how our heterogeneous estimates of teacher value-added on both English Language Arts (ELA) 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. Overall 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.2 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 larger classes tend to be in more affluent schools with higher average test scores.

There is also no clear evidence of sorting on comparative advantage. Figure 3 also depicts 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 are sorting into classes with slightly larger shares of lower-scoring students, but the opposite is true in ELA. Neither of these patterns is economically large. The differences by class size are similarly signed but even smaller (see Appendix Figure A.2). The combination of heterogeneity in teacher effects and the absence of significant sorting in the status quo suggest large gains from reallocation.

The current allocation of students to classes also suggests that there will be gains from reallocations. Variance in class size and class composition will both increase the gains from reallocation. Appendix Table A.1 reports the standard deviations of class size and the share of higher-scoring students in math and ELA at a district-wide level and within schools (controlling for variation by grade and year), revealing ample variation even within school. This suggests that although reallocating teachers across schools necessarily allows for bigger test-score gains, much of the potential gains may be achievable by reallocating teachers within their current school and grade.

4. Efficiently Allocating Teachers to Classes

Although our general theoretical framework could be applied in many settings, with estimates of the heterogeneous teacher effects we now use our theory to consider the public service provision problem of allocating teachers to classes. This section defines the allocation problem, presents the gains possible under the optimal allocations, and compares the gains obtained from using our estimates relative to using standard value-added measures.

We parameterize the social objective $\tilde{\mathcal{W}}$ using higher- and lower- scoring students to compare different allocations and find the relevant optima. Let $\mathcal{J} : (i,t) \rightarrow j$ 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):

$$\max_{\mathcal{J}\in\mathscr{J}} \tilde{\mathcal{W}}(\mathcal{J};\omega) = \max_{\mathcal{J}\in\mathscr{J}} \frac{1}{N_{i,t}} \sum_{(i,t)} \omega_L \, L_{i,t} \, \hat{\tau}_L^{\mathcal{J}(i,t)} + (1-\omega_L)\,(1-L_{i,t})\,\hat{\tau}_H^{\mathcal{J}(i,t)} \tag{7}$$

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. The set $\mathscr{J}$ is the social planner's choice set made up 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—something we will 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, 2022). 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., 2022), 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.

We solve this allocation problem for two sets of possible reallocations: within-school and district-wide. For both, we restrict $\mathscr{J}$ so that every year the students in each class and the grade assignments of each teacher do not change. We leave class composition fixed so that changes in within-class peer effects do not contaminate the outcomes in predicted counterfactual allocations. For the within-school reallocation we further require that teachers do not change schools. Whereas this within-school problem can be solved easily by iterating over school-grade(-year) cells, the district-wide reallocation problem has over 3×10^{1830} allocations to search over. 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. The social planner problem in equation 7 can be re-characterized as a mixed-integer linear programming problem and solved using the COIN-OR Branch and Cut solver implemented by the Python package Pulp (see, for example, DeNegre and Ralphs, 2009).

4.1 Allocations Incorporating Heterogeneous Impacts Increase Test Scores

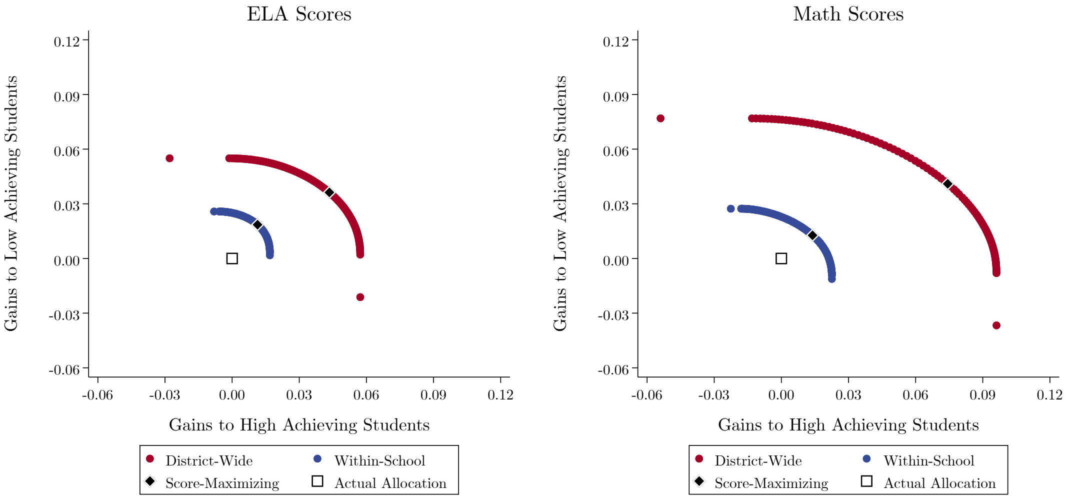
We create a production-possibility frontier (PPF) for the gains to each group from the within-school and district-wide reallocations. To do this, we solve the optimization problem in Equation 7 for 101 different values of the social weights ω_L ranging from 0.0 to 1.0. We then recover the average value-added received by lower- and higher-scoring students and calculate the gain beyond the status quo. By comparing the optimal gains attained under different weights, this analysis characterizes how reallocation gains to lower-scoring students trade off with those to higher-scoring students, creating the PPFs.

We depict these production-possibility frontiers in Figure 4. We plot the PPF for change in ELA 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 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 4 reveals three striking patterns. First, there are large gains possible from both reallocations. For example, in the district-wide reallocation a social planner seeking to raise average scores (i.e., a utilitarian planner with $\omega_L = \omega_H = 0.5$) could increase both lower- and higher-scoring students' scores by 0.04 student standard deviations. Gains from math are even larger: 0.04 for lower-scoring students and 0.07 for higher. Similarly, the simpler within-school reallocation could raise ELA and Math scores for both groups by more than 0.01 standard deviations. Recalling that these represent one-year gains, a policy that optimally allocated teachers could increase average math scores by 0.12σ in ELA and 0.17σ in math.[11] These are large gains—almost identical 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 et al., 2009; Hanushek, 2011; Chetty et al., 2014b).

[11] Where the annual means and standard deviations scores are normalized by those in the entire state of California.

Figure 4: Optimal Allocations Can Create Large Gains to High- and Low-scoring Students

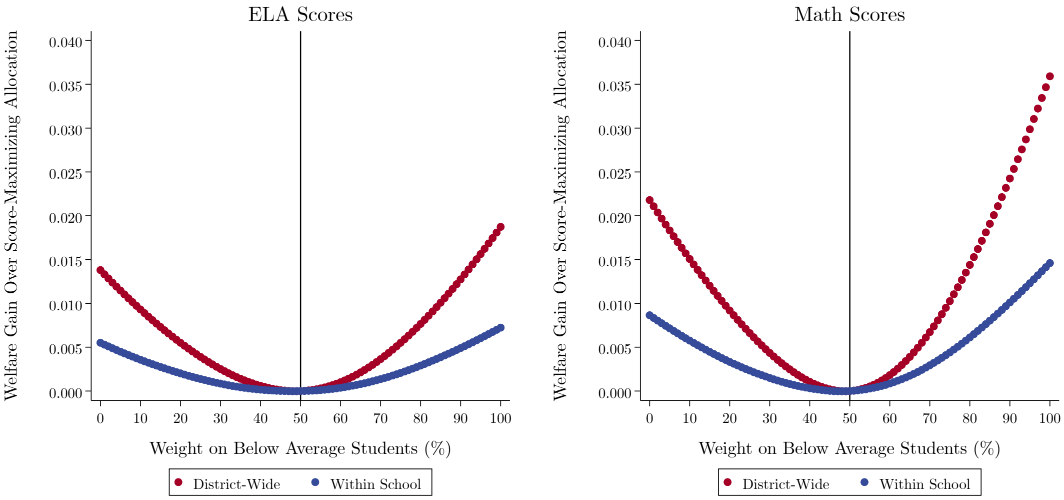


Note: This figure shows the test score gains from optimal allocations relative to the status quo. Two production possibility frontiers are presented, one for reallocating teachers within school-grade cells and one reallocating teachers across schools (still within grade). Each PPF is constructed by finding the optimal allocation given relative weights on lower- and higher-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.

The second pattern visible in Figure 4 is that the curvature of the PPFs demonstrates the value in explicitly considering the distributional goals of a policymaker. 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.

We compare the total welfare achieved under an optimal allocation for a given set of welfare weights (the optimal point on a PPF in Figure 4 for a given indifference curve) to the test-score maximizing allocation (the black diamond mark on the relevant PPF). To normalize these welfare gains, we construct an "Atkinson index" type measure such that the social planner would be indifferent between the optimal allocation and an allocation where every student experienced a given test score gain. Figure 5 shows the difference in this Atkinson index for each allocation on the comparative advantage frontier compared to the test-score maximizing allocation. As expected, the gains are small for similar weights and grow as the social planner favors one group more or less. 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 ELA. Of course, the true weights for policymakers may not be near these tails, but Figure 5 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 5: 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.

Estimating these gains highlights three interesting implications for our understanding of teacher allocations. First, the gains to math scores are larger than the gains to ELA scores. This is because the variance in teacher value-added on math is larger as shown in Figure 2 and in prior work (e.g., Chetty et al., 2014a). This suggests that for one-subject reallocations like Bates et al. (2022), it is indeed better to focus on math in order to raise average scores. Second, the allocations that optimize math scores and ELA scores are distinct. This is because the teachers that are the best at teaching each group of students math are not always the best at teaching those students in ELA. As such, the gains highlighted in papers that do reallocations using one subject at a time like Delgado (2022) and Bates et al. (2022) only give a lower bound to the gains from using information on both outcomes simultaneously. This will motivate our analyses in Section 5 where we aggregate gains over multidimensional outcomes. Finally, note that the largest possible gains to each group are different. 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 compared to lower-scoring students. This class-size dimension becomes particularly important when comparing these allocations to those made using only information about absolute advantage from traditional value-added estimates.

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. Furthermore, teachers have strong preferences over locations (Boyd et al., 2005a) and schools (Bates et al., 2022) that could impede some allocations from being incentive compatible. Finally, 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).

4.2 What Value Does Estimating Heterogeneity Add?

The previous subsection quantified large gains from teacher reallocations, but how much of these gains would be possible without knowing the heterogeneous effects? If all of these gains simply come from moving better teachers to larger classes, there is no need to estimate heterogeneous effects. To evaluate the importance of estimating heterogeneity, we compare the best allocations using heterogeneous estimates with those possible using only standard estimates of value-added. This allows us to decompose the welfare gains from the best allocations into the absolute advantage, comparative advantage, and redistribution components.

To find the optimal allocations with the standard value-added we use the same set of social objective functions and same solution concept, but we replace the estimates of each teacher's value-added on both higher- and lower-scoring students with the standard estimates:

$$\max_{\mathcal{J}\in\mathscr{J}} \widetilde{\mathcal{W}}_{VA}(\mathcal{J};\omega) = \max_{\mathcal{J}\in\mathscr{J}} \frac{1}{N_{i,t}} \sum_{(i,t)} \omega_L \, L_{i,t} \, \hat{\tau}_{VA}^{\mathcal{J}(i,t)} + (1-\omega_L)\,(1-L_{i,t})\,\hat{\tau}_{VA}^{\mathcal{J}(i,t)} \tag{8}$$

where $\hat{\tau}_{VA}^{j}$ is the standard value estimate described in section 3.2.1 and where we again solve the problem for 101 different values of the social weights ω_L ranging from 0.0 to 1.0. Intuitively, the gains from using absolute advantage as captured in the standard measures come from putting the higher value-added teachers in larger classes to maximize average scores—or using ω_L-weighted class size when the social planner has heterogeneous preferences over groups' gains. The gains attained and reported at each point are calculated using our heterogeneous estimates to avoid compromising the external validity of our score predictions that would occur if using standard estimates to predict the effect of sending teachers to very different classes.

4.2.1 Estimating Heterogeneity Increases Average Test Scores

As illustrated in Figure 1, using heterogeneous value-added could increase average scores beyond what is possible using standard value-added via comparative advantage. This subsection explores the extent to which information about comparative advantages can raise average scores in practice. We document large gains beyond what can be accomplished using the information about absolute advantage that standard value-added measures provide.

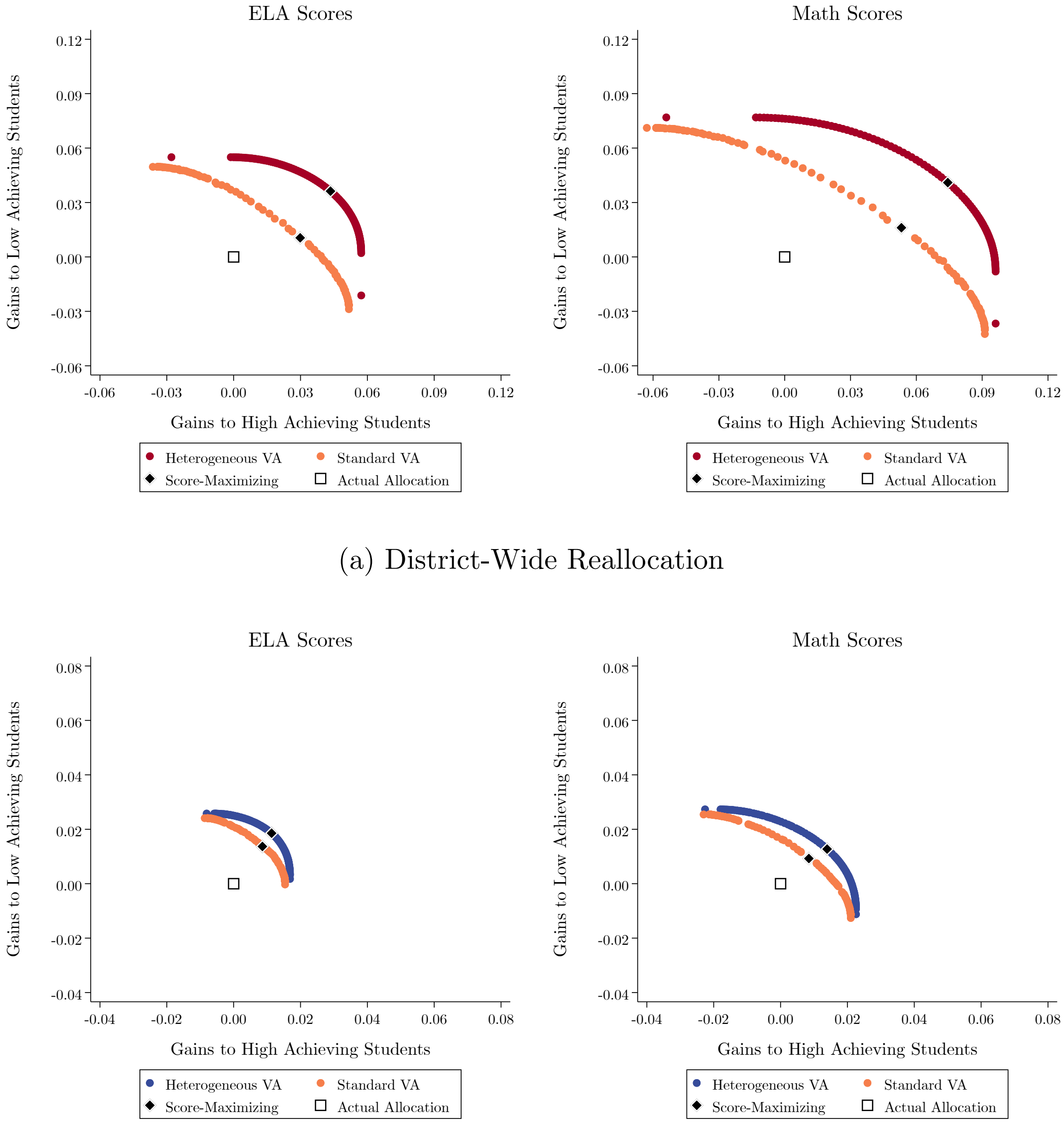
To approach this question, we depict and compare the production-possibility frontiers for average achievement gains to each group using heterogeneous and standard value-added in Figure 6. Here again each point presents the average change in lower-scoring students' test scores in the optimal allocations (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—which is obtained by placing the highest value-added teachers in the largest classes.

Note that the empirical results in Figure 6 are analogous to the theoretical depiction in Figure 1. For each panel the outer PPF presents the changes in test scores possible by using information about both absolute and comparative advantage based on the heterogeneous teacher effects whereas the interior PPF presents the changes in test scores possible by using only the information about absolute advantage contained in standard value-added estimates. Again, the current allocation is denoted with a square.

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 generate 97.3% higher ELA scores and 66.4% higher Math scores than allocations using only absolute advantage. These are large gains: an average gain of 0.020σ in ELA or 0.023σ in Math for students in the district would be an impressive policy victory, especially considering this policy could be implemented year-over-year for compounding gains. Gains to the within-school reallocations are smaller in absolute terms, but comparative advantage is still critical. Using heterogeneous effects boosts average ELA scores by 34.1% and math scores by 50.3% (both about 0.0045σ).

Interestingly, even for a social planner trying to maximize average scores the choice between standard and heterogeneous value-added measures has striking distributional implications in the district-wide allocations. On one hand, the average-score gains from reallocations using only information about absolute advantage (from standard value-added) are concentrated among higher-scoring students. For example, the higher-scoring students' gains of 0.03σ in ELA and 0.05σ in Math are almost exactly three times larger than the cor-

Figure 6: 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.

responding gains to lower-scoring students. On the other hand, the large gains from using comparative advantage in the district-wide reallocations accrue disproportionately to lower-scoring students. For example, the 0.02σ ELA gain is split almost 0.03σ to lower-scoring students and just over 0.01σ to higher-scoring students. Figure 6 depicts these observations visibly: Whereas the expansion path from the status quo through the two PPFs is almost linear for the within-school reallocations in Panel (b), it is extremely non-linear for the district-wide reallocations Panel (a). These asymmetries motivate a direct focus on the equity implications of using heterogeneity.

4.2.2 The Interaction of Distributional Goals and Comparative Advantage

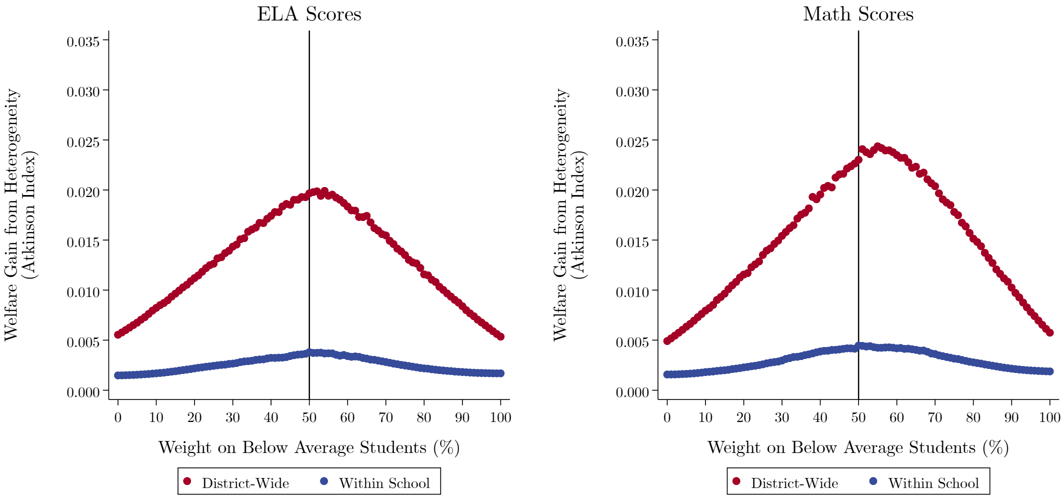
The above section shows that when the goal is to maximize average scores, using heterogeneous value-added leads to significant gains. We also know from section 4.1 that when policymakers favor one group over another, considering their distributional goals leads to significant welfare gains. Putting these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

Using Figure 6 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 7 reports the results. In Appendix Figure A.3, 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 7 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 ELA 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

Figure 7: 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.

lower-scoring value-added in the same classes. This is visible in how close the frontiers are in Figure 6 and in the upside-down U-shape in the gains reported in Figure 7.

The key driver of these differences are the relative shapes of the PPFs and how they affect scores. As seen in Figure 6, 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.3. 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.

4.3 Other Equity Implications from Reallocations

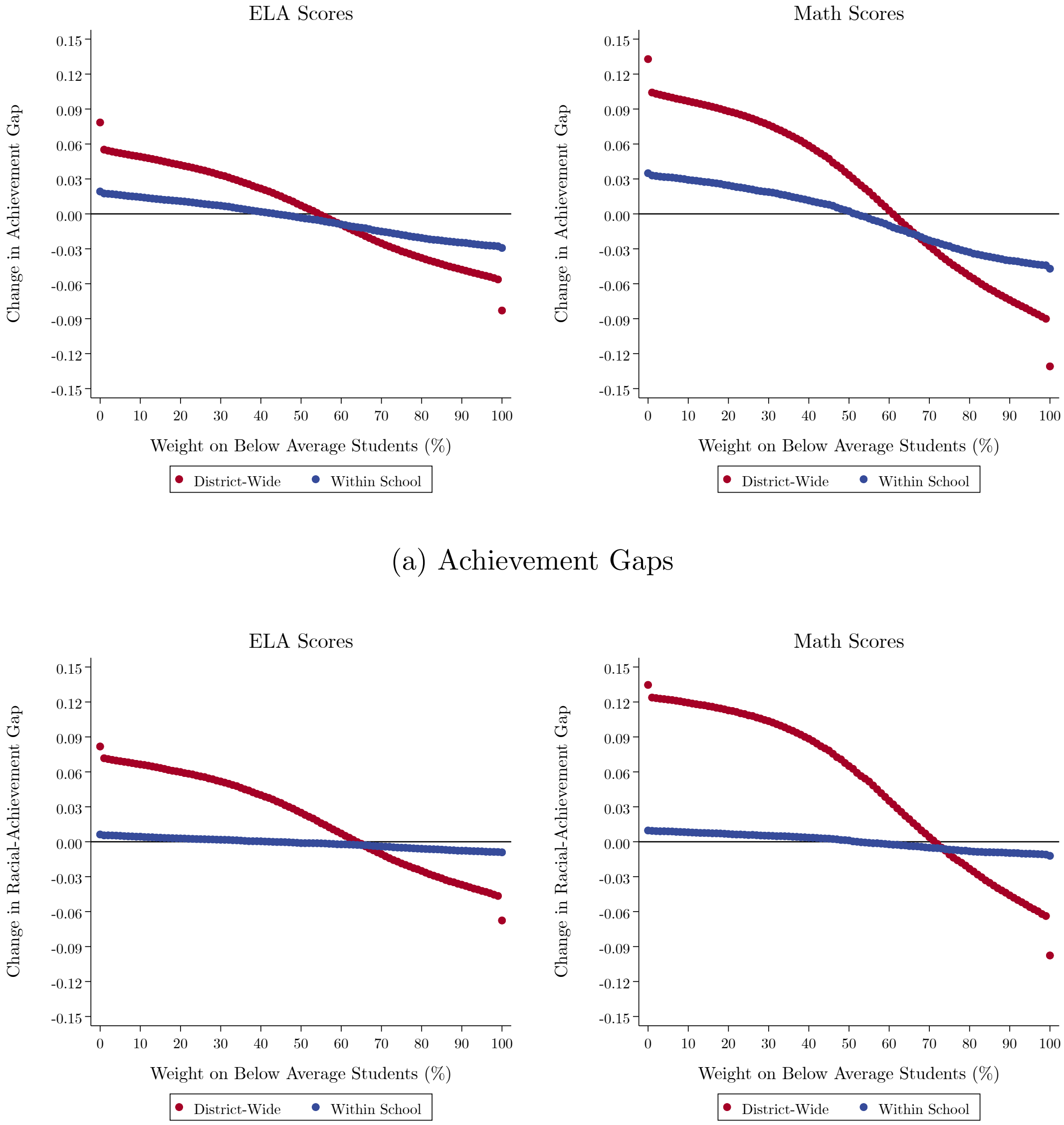
Having described the optimal reallocations and decomposed the welfare gains from them, our final task is to explore other equity implications that the proposed reallocations would have. Specifically, we study how our reallocations affect overall achievement gaps and racial achievement gaps, and we describe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

4.3.1 Shrinking Achievement Gaps

Many education policies—including those that motivated our welfare theory—propose interventions that will lower the achievement gaps between lower- and higher-scoring students. To consider this we plot out the change in two policy-relevant achievement gaps in Figure 8. First, in Panel (a) we show how the optimal within-school and district-wide reallocations for each ω_L would change the achievement gap between students who performed above and below median in the previous year. We also report similar changes in the racial achievement gap in Panel (b). We define this gap as the difference in average scores between Black and Hispanic students versus White and Asian students. Interestingly, we show that our completely race-blind policies can reduce average racial test score gaps just as much as the race focused reallocations in Delgado (2022).

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. For example, the baseline gap between students who scored above and below the median last year is 1.27σ in ELA 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 ELA 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 (2022) where allocations are made to explicitly shrink racial gaps in math scores subject to not lowering average scores. Delgado (2022) finds a

Figure 8: 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.

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.[12]

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 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 ELA 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 ELA 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 utilitarian, test-score maximizing reallocations ($\omega_L = \omega_H = 0.5$) within school tend to not affect either gap significantly,[13] 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.

4.3.2 Reallocation 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

[12] 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.

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

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.4 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 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.4 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 ELA 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.[14] 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—

[14] 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.

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.

5. From value-added to Welfare Added

We have provided a welfare theory, estimated the relevant parameters, and demonstrated the test score gains from reallocations along a single subject. Our empirical findings so far can be interpreted as statements about a popular outcome of interest, test scores. With some assumptions, however, our findings on test score gains can be interpreted as an unbiased, or less biased than the mean, welfare estimate using our welfare theory.

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, forthcoming; 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. This assumption is analogous to the "no spillovers" condition assumed in Section 2. 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, we need to consider the bias terms from Theorem 2. First, consider the covariance term. It is important to remember that this term is dependent on the policymaker's welfare weights. As mentioned above, the covariance terms would be zero if our policymaker truly cared about only average lower- and higher-scoring students. If this is not the case, for a completely unbiased estimate, we need the conditional covariance of the true welfare weights (that consider all factors important to the policymaker) and student gains to be uncorrelated. We know that different allocations impact racial test score gaps and that gains from some reallocations accrue to lower-scoring students primarily in higher-scoring schools. While the estimates may not be unbiased in this case, satisfying Equation 2 would still ensure they are better than simple means. Conditioning on additional factors like race and school average scores could further assuage these concerns, but for tractability, we stick to conditioning on test scores.

Next, we consider the estimation bias between our estimated conditional average treatment effect and the truth. While we know teacher impacts differ along different dimensions (Delgado, 2022), we believe conditioning on test scores captures much of the variation without over-fitting. While race also plays a role, finding common support for all teachers can be practically challenging. Gender may play a role in teacher impacts as well; however, gender composition does not change significantly between most classes, limiting the bias introduced by teacher heterogeneity.

There are still two significant shortcomings that we address in the following section. First, these teachers teach both ELA and Math, and so an optimal reallocation policy would consider the impact on both 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) of how value-added affects lifetime earnings.

The second shortcoming to address is the impact of reallocations on teachers. We need to consider the welfare component attributable to teachers' disutility from the reallocations. We treat teacher's preferences as an incentive compatibility constraint and assume they will need to be compensated 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 at least some may prefer their new assignments,[15] 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., 2022), but highly structured wages in teacher labor markets often make this difficult in practice. We will focus on the marginal value of public funds (MVPF, Hendren and Sprung-Keyser, 2020) for a hypothetical universal bonus program.

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.

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

5.1 Students: Earnings Implications of Reallocations

We begin with the welfare implications for students under the assumptions outlined above. These results are most closely tied to our previous analyses focused on student gains. 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 ELA on student earnings from Chetty et al. (2014a). They estimate that a one standard deviation increase in ELA scores in elementary school generates an additional \$1,524 in earnings in early adulthood and that the corresponding gains in Math are \$650.

Because of the fundamental trade-off between the facts that our reallocations generate larger gains in math, but gains to ELA 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, ELA and Math.[16]

$$\widetilde{\mathcal{W}}(\mathcal{J};\omega) = \frac{1}{N_{i,t}} \sum_{(i,t)} \sum_{s} \omega_s \left[L_{i,s,t}\, \hat{\tau}_{L,s}^{\mathcal{J}(i,t)} + (1 - L_{i,s,t})\, \hat{\tau}_{H,s}^{\mathcal{J}(i,t)} \right] \tag{9}$$

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 9 presents the trade-offs between gains to average Math and average ELA 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 ELA value-added are somewhat correlated, this allocation would also raise ELA scores by 0.019σ (0.005σ within schools). The focus on math scores only, however, forgoes large ELA gains. This could be particularly problematic as ELA 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

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

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 ELA and math scores equal to the relative price. We illustrate this graphically in Panel (a) of Figure 9 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 ELA gains, but do not focus exclusively on ELA value-added because the value of marginal gains to ELA 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 (2022) and Bates et al. (2022) which both focus only on math. Surprisingly, our results overturn this intuition. We will discuss the details of how we obtain these numbers below, but we find that 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 ELA 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

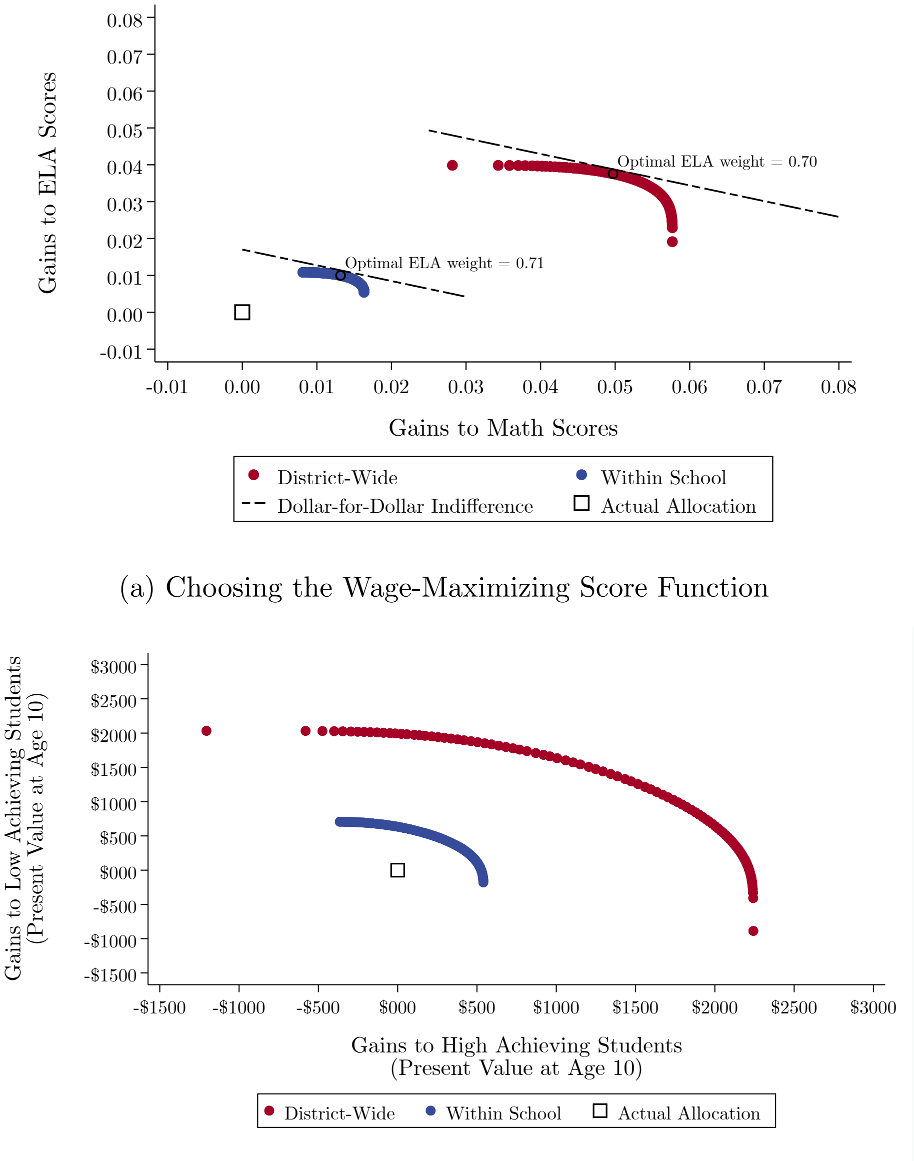
With information about the income-optimal score function in hand, we return to the question of optimal policy with heterogeneous social preferences. Combining all of the pieces we define a new social welfare function to optimize

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

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.[17]

[17] Note that because not all students are low scoring in Math and ELA the achievement weight ω_L may not apply uniformly to each student. In practice this means that there are four implicit weights generated by

Figure 9: 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 ELA scores to estimate the frontier of possible earnings gains. Panel (a) displays the PPF of math versus ELA 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 ELA + 0.25 Math (present valued at age 10).

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 9 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 students of each type each year (on average), this implies the value of the reallocation across all students is \$27.9 million. 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.

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.5, 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 dollar 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 dollar gains attainable, a discussion of how to do so is of first-order importance. The following subsection provides that discussion.

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.

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

$$MVPF^j(b) = \frac{\sum_i (1-t)\Delta S_i^p)}{N_j b - t\Delta S_)^p} \tag{10}$$

where $(1-t)\Delta S_i^p$ 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 $t\Delta S_i^p$ 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).[18]

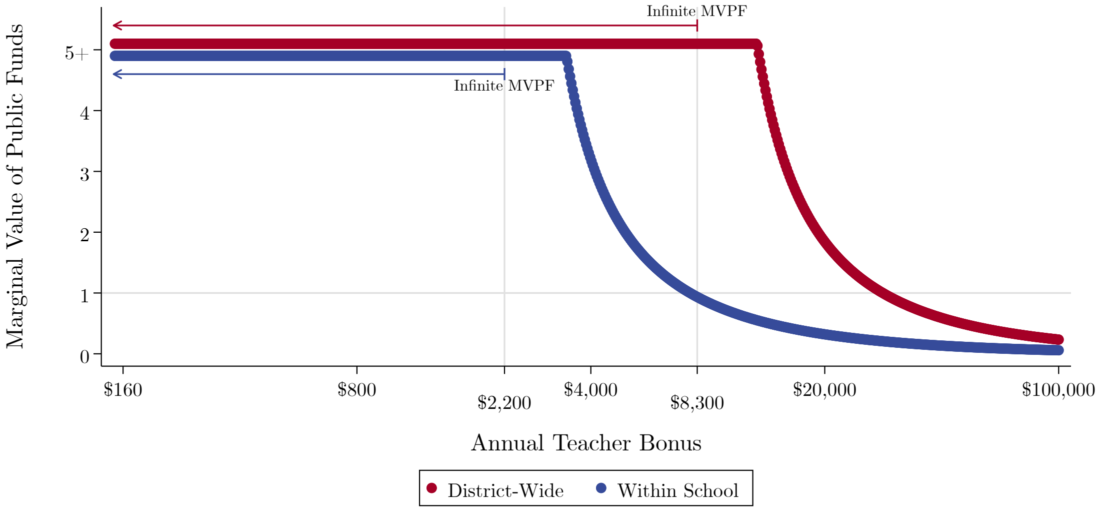
We combine our estimates of present-value monetary gains with data from the Opportunity Atlas (Chetty et al., 2018) to calculate these MVPF empirically. For the changes in earnings, we focus on the utilitarian, earnings-maximizing, within-school and district-wide reallocations as described in the previous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

[18] 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.

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 10.

We present the results in Figure 10. Figure 10 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.

Figure 10: 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.

The main takeaway from Figure 10 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 second number is particularly striking because despite being noninvasive the within-school reallocation is still generating substantial gains.

A second important insight from Figure 10 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., 2022)—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.

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. Conclusion and Implications for Policy

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 et al., 2009; Hanushek, 2011; Chetty et al., 2014b; Delgado, 2022). 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, A., P. A. Pathak, J. Schellenberg, and C. R. Walters (2020): "Do parents value school effectiveness?" American Economic Review, 110, 1502–39.

Abrams, D. S. and A. H. Yoon (2007): "The luck of the draw: Using random case assignment to investigate attorney ability," University of Chicago Law Review, 74, 1145.

Agrawal, D., W. Hoyt, and T. Ly (2023): "A New Approach to Evaluating the Welfare Effects of Decentralized Policies," Working paper.

Alatas, V., R. Purnamasari, M. Wai-Poi, A. Banerjee, B. A. Olken, and R. Hanna (2016): "Self-targeting: Evidence from a field experiment in Indonesia," Journal of Political Economy, 124, 371–427.

Angrist, J., P. Hull, and C. R. Walters (2022): "Methods for Measuring School Effectiveness," .

Athey, S., R. Chetty, G. W. Imbens, and H. Kang (2019): "The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely," Tech. rep., National Bureau of Economic Research.

Athey, S. and S. Wager (2021): "Policy learning with observational data," Econometrica, 89, 133–161.

Bacher-Hicks, A. and C. Koedel (2022): "Estimation and interpretation of teacher value added in research applications," Working paper.

Bates, M. D., M. Dinerstein, A. C. Johnston, and I. Sorkin (2022): "Teacher Labor Market Equilibrium and Student Achievement," Tech. rep., National Bureau of Economic Research.

Betts, J. R. (2011): "The economics of tracking in education," in Handbook of the Economics of Education, Elsevier, vol. 3, 341–381.

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

Bhatt, M. P., S. B. Heller, M. Kapustin, M. Bertrand, and C. Blattman (2023): "Predicting and Preventing Gun Violence: An Experimental Evaluation of READI Chicago," Tech. rep., 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, 24, 113–132.

Chan, D. C., M. Gentzkow, and C. Yu (2022): "Selection with variation in diagnostic skill: Evidence from radiologists," The Quarterly Journal of Economics, 137, 729–783.

Chandra, A., A. Finkelstein, A. Sacarny, and C. Syverson (2016): "Health care exceptionalism? Performance and allocation in the US health care sector," American Economic Review, 106, 2110–2144.

Chetty, R., J. N. Friedman, N. Hendren, M. R. Jones, and S. R. Porter (2018): "The opportunity atlas," Opportunity Insights.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a): "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates," American Economic Review, 104, 2593–2632.

——— (2014b): "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood," American Economic Review, 104, 2633–79.

Condie, S., L. Lefgren, and D. Sims (2014): "Teacher heterogeneity, value-added and education policy," Economics of Education Review, 40, 76–92.

Dahlstrand, A. (2022): "Defying Distance? The Provision of Services in the Digital Age," Tech. rep.

Dee, T. S. (2005): "A teacher like me: Does race, ethnicity, or gender matter?" American Economic Review, 95, 158–165.

Delgado, W. (2022): "Heterogeneous Teacher Effects, Comparative Advantage, and Match Quality: Evidence from Chicago Public Schools," .

Delhommer, S. (2019): "High School Role Models and Minority College Achievement," Tech. rep.

DeNegre, S. T. and T. K. Ralphs (2009): "A branch-and-cut algorithm for integer bilevel linear programs," in Operations research and cyber-infrastructure, Springer, 65–78.

Diamond, P. A. (1973): "Consumption externalities and imperfect corrective pricing," The Bell Journal of Economics and Management Science, 526–538.

Doyle, J., J. Graves, and J. Gruber (2019): "Evaluating measures of hospital quality: Evidence from ambulance referral patterns," Review of Economics and Statistics, 101, 841–852.

Duflo, E., P. Dupas, and M. Kremer (2011): "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya," American economic review, 101, 1739–1774.

Einav, L., A. Finkelstein, and N. Mahoney (2022): "Producing Health: Measuring Value Added of Nursing Homes," Tech. rep., National Bureau of Economic Research.

Fell, H., D. T. Kaffine, and K. Novan (2021): "Emissions, transmission, and the environmental value of renewable energy," American Economic Journal: Economic Policy, 13, 241–72.

Finkelstein, A. and M. J. Notowidigdo (2019): "Take-up and targeting: Experimental evidence from SNAP," The Quarterly Journal of Economics, 134, 1505–1556.

Glazerman, S., A. Protik, B.-r. Teh, J. Bruch, and J. Max (2013): "Transfer Incentives for High-Performing Teachers: Final Results from a Multi-site Randomized Experiment," Tech. rep., U.S. Department of Education.

Griffith, R., M. O'Connell, and K. Smith (2019): "Tax design in the alcohol market," Journal of Public Economics, 172, 20–35.

Hanushek, E. A. (2011): "The economic value of higher teacher quality," Economics of Education review, 30, 466–479.

Hanushek, E. A. et al. (2009): "Teacher deselection," Creating a new teaching profession, 168, 172–173.

Harrington, E. and H. Shaffer (2023): "Estimating prosecutor effects on incarceration and reoffense," Tech. rep., Working Paper.

Hendren, N. and B. Sprung-Keyser (2020): "A Unified Welfare Analysis of Government Policies," 135, 1209–1318.

Hollingsworth, A. and I. Rudik (2019): "External impacts of local energy policy: The case of renewable portfolio standards," Journal of the Association of Environmental and Resource Economists, 6, 187–213.

Hull, P. (2020): "Estimating hospital quality with quasi-experimental data," Tech. rep., Working Paper.

Hussam, R., N. Rigol, and B. N. Roth (2022): "Targeting high ability entrepreneurs using community information: Mechanism design in the field," American Economic Review, 112, 861–98.

Ida, T., T. Ishihara, K. Ito, D. Kido, T. Kitagawa, S. Sakaguchi, and S. Sasaki (2022): "Choosing Who Chooses: Selection-Driven Targeting in Energy Rebate Programs," Tech. rep., National Bureau of Economic Research.

Imberman, S. A. and M. 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, 91, 104–121.

Ito, K., T. Ida, and M. Tanaka (2021): "Selection on Welfare Gains: Experimental Evidence from Electricity Plan Choice," Tech. rep., National Bureau of Economic Research.

Jackson, C. K. (2018): "What do test scores miss? The importance of teacher effects on non–test score outcomes," Journal of Political Economy, 126, 2072–2107.

Jacob, B. A. and L. Lefgren (2007): "What do parents value in education? An empirical investigation of parents' revealed preferences for teachers," The Quarterly Journal of Economics, 122, 1603–1637.

Johnson, A. C. (2021): "Preferences, Selection, and the Structure of Teacher Pay," Working paper.

Kitagawa, T. and A. Tetenov (2018): "Who should be treated? empirical welfare maximization methods for treatment choice," Econometrica, 86, 591–616.

Krueger, A. B. (1999): "Experimental estimates of education production functions," The quarterly journal of economics, 114, 497–532.

Norris, S. (2019): "Examiner inconsistency: Evidence from refugee appeals," University of Chicago, Becker Friedman Institute for Economics Working Paper.

Petek, N. and N. G. Pope (forthcoming): "The Multidimensional Impact of Teachers on Students," Journal of Political Economy.

Ricks, M. D. (2022): "Strategic Selection Around Kindergarten Recommendations," Tech. rep.

Rothstein, J. (2010): "Teacher quality in educational production: Tracking, decay, and student achievement," The Quarterly Journal of Economics, 125, 175–214.

——— (2015): "Teacher quality policy when supply matters," American Economic Review, 105, 100–130.

Sexton, S., A. J. Kirkpatrick, R. I. Harris, and N. Z. Muller (2021): "Heterogeneous solar capacity benefits, appropriability, and the costs of suboptimal siting," Journal of the Association of Environmental and Resource Economists, 8, 1209–1244.

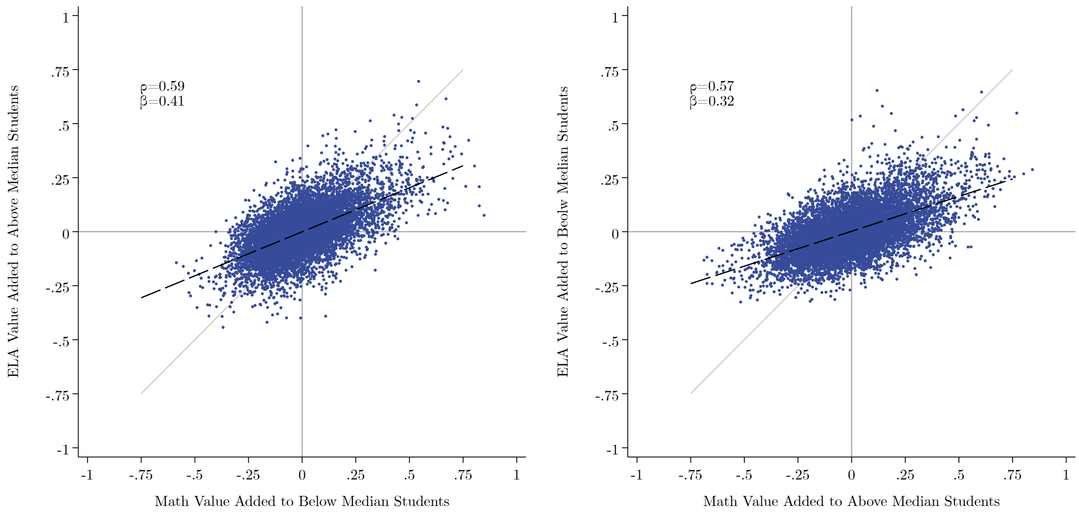
Staiger, D. O. and J. E. Rockoff (2010): "Searching for effective teachers with imperfect information," Journal of Economic perspectives, 24, 97–118.

Stepner, M. (2013): "VAM: Stata module to compute teacher value-added measures," Statistical Software Components, Boston College Department of Economics.

Tomlinson, C. A. (2017): How to differentiate instruction in academically diverse classrooms, ASCD, third ed.

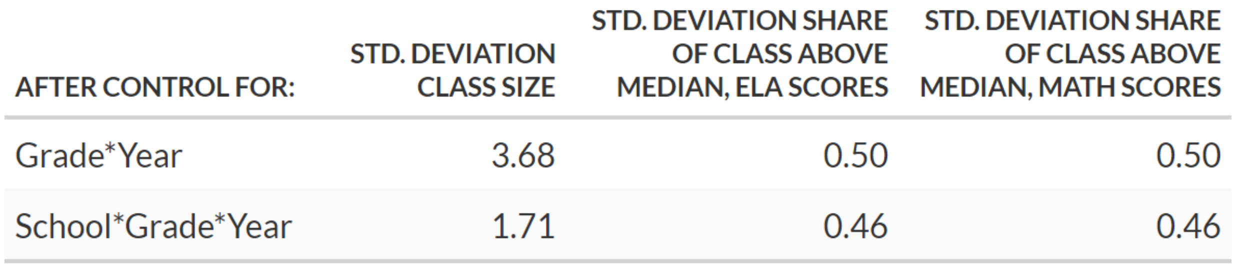
A. Additional Tables and Figures

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 English Language Arts (ELA) and Math test scores. Note that in this Figure Math and ELA 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.

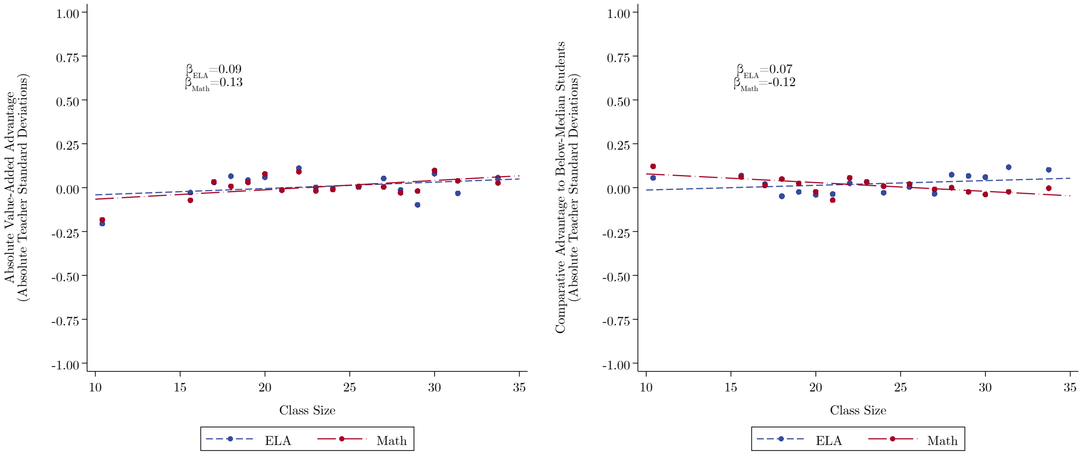
Table A.1: The Standard Deviation of Class Size and the Share of Students in the Class Who Are High-Scoring in ELA and Math



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

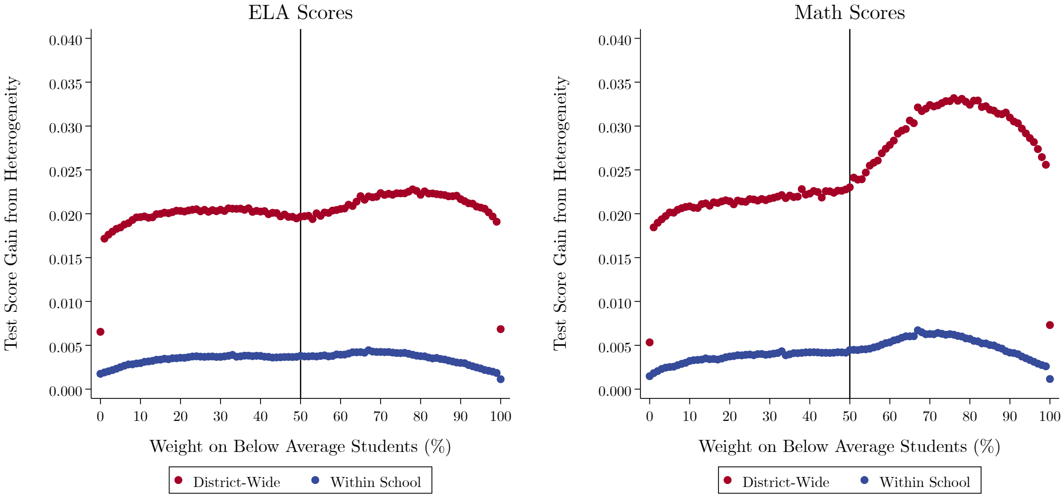
Note: This figure shows the within year-grade standard deviations in class size and composition at a district-wide level and a within-school level.

Figure A.2: value-added Only Varies Somewhat Across Class Sizes



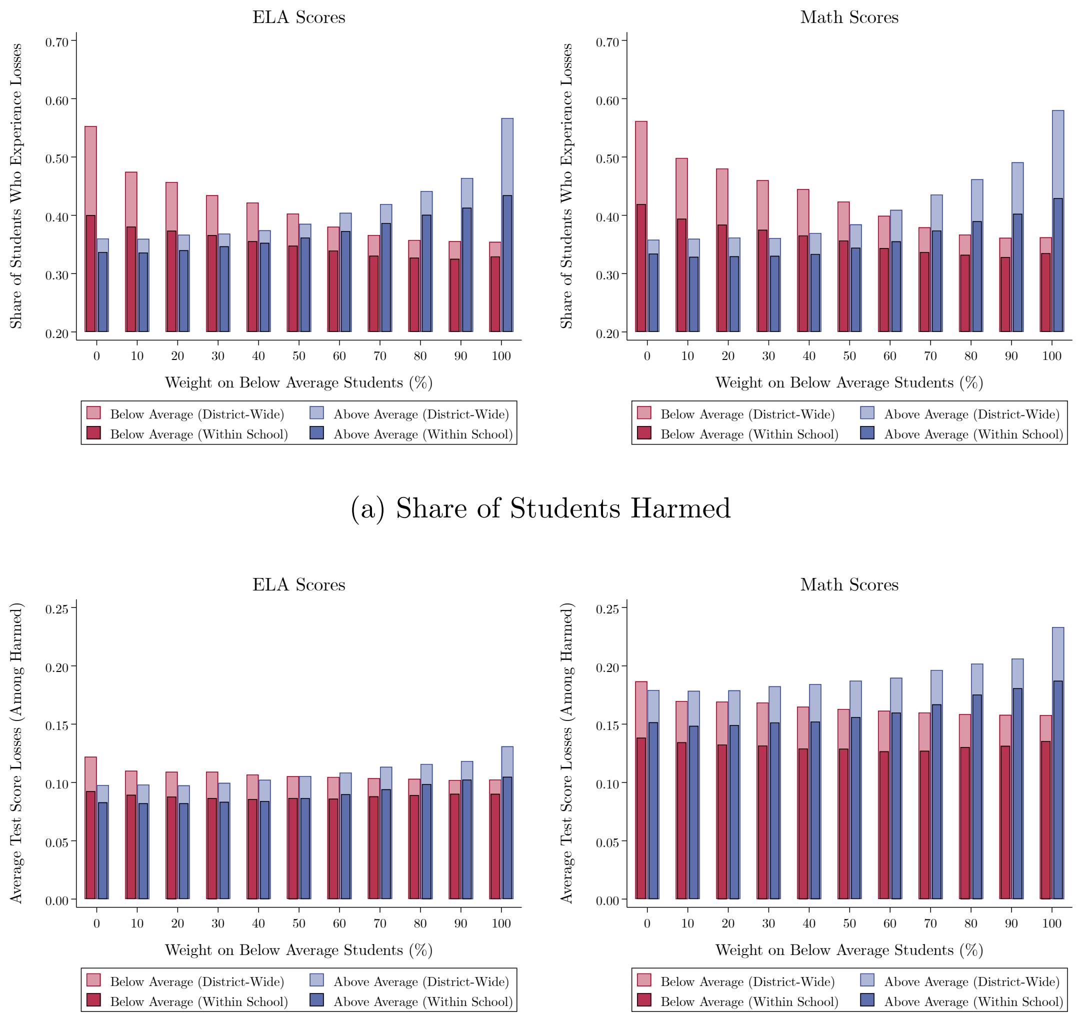
Note: This figure shows how our heterogeneous estimates of teacher value-added on both English Language Arts (ELA) and Math test scores relate to class composition. The panel on the left shows teacher absolute advantage (average of value-added on higher- and lower-scoring students) and the panel on the 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.3: 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.4: While Reallocations Help Many Students, They Will Harm Others

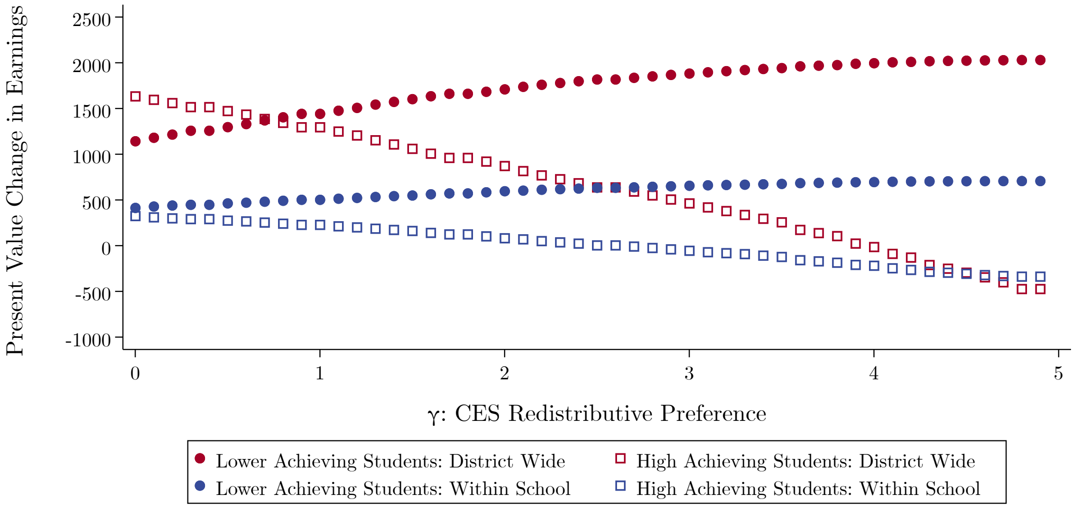


(a) Share of Students Harmed

(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.5: 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 From Test Scores to Welfare Details

Below is a more detailed version of definition 1

Proof. If a change in an individual's outcomes $\boldsymbol{Y}_i$ only impacts the utility and welfare weights of that individual i, then for a given score function S, the expected change in welfare $\Delta\tilde{\mathcal{W}}^j$ from the status quo policy $(j = 0)$ to policy j is

$$\begin{aligned}
\Delta\tilde{\mathcal{W}}^j &\equiv \mathbb{E}[\mathcal{W}^j|\boldsymbol{S}^j] - \mathbb{E}[\mathcal{W}^0|\boldsymbol{S}^0] \\
&= \sum_{i=1}^{n} \mathbb{E}[\psi_i^j U_i^j|S_i^j] - \mathbb{E}[\psi_i^0 U_i^0|S_i^0] \\
&= \sum_{i=1}^{n} \frac{\mathbb{E}[\psi_i^j U_i^j|S_i^j] - \mathbb{E}[\psi_i^0 U_i^0|S_i^0]}{\Delta S_i^p} \Delta S_i^p \\
&\equiv \sum_{i=1}^{n} \gamma_i(S_i^j, S_i^0)\Delta S_i^p
\end{aligned}$$

The last line is simply redefining the first term as a test score welfare weight $\gamma_i(S_i^j, S_i^0)$. $\boldsymbol{S}^j$ is the vector of test scores for every student under policy j. This means the expectations on the first line are conditional on the entire vector of test scores. This means the relationship between test scores and utility is fully flexible, and each student's utility can be uniquely impacted by a given test score change. Note that γ_i is an average over test score points for a given student, not an average across students. To understand this term, it is helpful to think through a simple example. Suppose $\mathbb{E}[\psi_i^j U_i^j | S_i^j] = S_{it}$ for all students. That is, expected welfare is linear in test scores. In this case, $\gamma_i(S_i^j, S_i^0) = 1$ because all students gain 1 util per score over the entire range of scores, and test scores are equivalent to welfare. Although welfare weights are often based off of earnings or earnings ability, the implication of definition 1 is that we can theoretically apply weights to a short term outcomes like test scores, rather than utility, and still have an unbiased estimate of welfare. Of course, in practice, getting individual weights is likely impossible. The later theory sections address the best way to overcome this problem with conditional aggregation, but definition 1 provides a ground truth reference that incorporates a large amount of of potential heterogeneity, individual differences.

B.2 Welfare Weighting the ATE

Using a similar approach to Hendren and Sprung-Keyser (2020), the following equation shows how it is possible to estimate welfare from an average treatment effect if the proper weight is applied

$$\Delta \mathcal{W}^j \tag{11}$$

$$= \int_0^1 \gamma_i(S_i^j, S_i^0) \Delta S_i^p \mathrm{d}i \tag{12}$$

$$= \frac{\int_0^1 \gamma_i(S_i^j, S_i^0) \Delta S_i^p \mathrm{d}i}{\int_0^1 \Delta S_i^p \mathrm{d}i} \int_0^1 \Delta S_i^p \mathrm{d}i \tag{13}$$

$$= \tilde{\gamma}^j ATE^j \tag{14}$$

The trouble is that the first term, $\tilde{\gamma}^j$ depends, not just on the test score welfare weights γ_i, but also on the joint distribution of those weights with the changes in test scores for policy j. It is a complex object that involves a deep understanding of the distribution of heterogeneous impacts resulting from policy j. If a policymaker already has this deep knowledge, it is not clear how much giving them the average treatment effect will help.

B.3 Theorem 1 proof

Proof.

$$\begin{aligned}\textbf{Average Bias}_{ATE} &= \frac{\Delta\tilde{\mathcal{W}}^j}{n} - \mathbb{E}[\gamma^p]\widehat{ATE} \\ &= \frac{1}{n}\sum_{i=1}^{n}\gamma_i(S_i^j, S_i^0)\Delta S_i^p - \mathbb{E}[\gamma^p]\widehat{ATE} \\ &= \mathbb{E}[\gamma^p\Delta S^p] - E[\gamma^p]\widehat{ATE} \\ &= \mathbb{E}[\gamma^p]\mathbb{E}[\Delta S^p] + \text{Cov}(\gamma^p, \Delta S^p) - E[\gamma^p]\widehat{ATE} \\ &= \text{Cov}(\gamma^p, \Delta S^p) + \mathbb{E}[\gamma^p]\left(\mathbb{E}[\Delta S^p] - \widehat{ATE}\right)\end{aligned}$$

The first line is how we are defining bias. It is the benchmark with individual heterogeneity minus our common estimator of the mean welfare weight and the average treatment effect. The second line comes from definition 1. The third line comes from recognizing that the first term in line two is the population average, or expectation, of $\gamma^p\Delta S^p$. The fourth line uses the general definition of covariance, that is $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[y]$. The last line just rearanges the terms.

B.4 Averate Treatment Effect Bias Explained

The specific source of average treatment effect bias we are consider can be a concern for any policy j that involves assigning specific sub-treatments d (teachers) to subsets of the population of size K_d^j (classes). First note that the average treatment effect is the following weighted average of sub-treatment effects ATE_d^j

$$ATE^j = \frac{1}{n}\sum_d K_d^j ATE_d^j$$

The bias comes in from incorrect estimates of the average sub-treatment effect (teacher impact) ATE_d^j characterized by the following

$$ATE_d^j - \widehat{ATE}_d^j = \frac{1}{K_d^j}\sum_{i=1}^{K_d^j}\Delta S_i^d - \frac{1}{K_d^0}\sum_{l=1}^{K_d^0}\Delta S_l^d$$

Here we can see the bias comes from different individual impacts between the existing class and the class in the policy counterfactual. It is helpful to think through the two cases where this difference goes to zero. First, if there is no treatment effect heterogeneity. For example, a teacher impacts all students equally on average and so $\Delta S_i^d = \Delta S_l^d \quad \forall \quad i, l$. Second, even if there is treatment effect heterogeneity, if the classes have similar characteristics the means may still be the same. For example, a teacher may be very bad at teaching English language learners (ELA). However, if both classes have the same fraction of ELA students, the teacher's mean impact will be the same.

B.5 Conditional Average Treatment Effect Bias Explained

The bias in the second term will be lower after conditioning when

$$\mathbb{E}[\Delta S^p] - \widehat{ATE} > \sum_x P_x \left(\mathbb{E}[\Delta S^p | x] - C\widehat{ATE(X)} \right) \tag{15}$$

As in the previous section, we can zero in on a specific teacher or sub-treatment and see that, for a given teacher, conditioning reduces bias when

$$ATE_d^j - \widehat{ATE}_d^j \tag{16}$$

$$= \frac{1}{K_d^j} \sum_{i=1}^{K_d^j} \Delta S_i^d - \frac{1}{K_d^0} \sum_{l=1}^{K_d^0} \Delta S_l^d \tag{17}$$

$$> \sum_X P_{dx}^j \left(\frac{1}{K_{dx}^j} \sum_{i=1}^{K_{dx}^j} \Delta S_i^d - \frac{1}{K_{dx}^0} \sum_{l=1}^{K_{dx}^0} \Delta S_l^d \right) \tag{18}$$

$$\sum_X P_{dx}^j \left(\widehat{ATE}_{dx}^j - \widehat{ATE}_{dx}^0 \right) \tag{19}$$

The left side is the difference in mean treatment effects between the baseline class and the counterfactual class, as described above. The right hand side is the difference in the mean treatment effects for a given x, weighted by the portion of students in the counterfactual class in group x. Bias in this case comes from differences within a group x between the baseline and counterfactual treatment effects. There is no longer any bias from differences in the fraction of students with characteristics x. If a teacher is worse at teaching struggling students, for example, and their new class has many more struggling students, the left hand side will overestimate their impact on the new class. The right hand side will only be biased if there is variation within performance groups in both the teachers impact and the student compositions. For example, teachers may have different impacts on students based on race, even within a pretest group, and racial composition could differ across class (Delgado, 2022).

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

$$S_{i,s,t} = \alpha_{j(i,s,t)} + \beta_s X_{i,t} + \epsilon_{i,s,t} \tag{20}$$

Where $X_{i,t}$ includes cubic polynomials in prior year test scores in math and ELA, 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[19]. $j(i,t)$ 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

$$\bar{A}_{jt} = \frac{1}{n} \sum_{i \in i : j(i,t) = j} A_{it} \tag{21}$$

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)"

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

$$\psi = \arg\min_{\psi} \sum_{j} \left(\bar{A}_{jt} - \sum_{s=1}^{t-1} \psi_s \bar{A}_{js}\right)^2 \tag{22}$$

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

$$\hat{\mu}_{jt} = \psi' \mathbf{A}_j^{-t} \tag{23}$$

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

$$S_{i,t} = \alpha_{j(i,b,t)} + \beta X_{i,t} \tag{24}$$

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,

$$\bar{A}_{jBt} = \frac{1}{n} \sum_{i \in i : j(i,B,t) = j} A_{it} \tag{25}$$

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

$$\psi = \arg\min_{\psi} \sum_{j} \left(\bar{A}_{jBt} - \sum_{s=1}^{t-1} \psi_s \bar{A}_{jBs}\right)^2 \tag{26}$$

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

$$\hat{\mu}_{jBt} = \psi' \mathbf{A}_{jB}^{-t} \tag{27}$$

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 2.

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

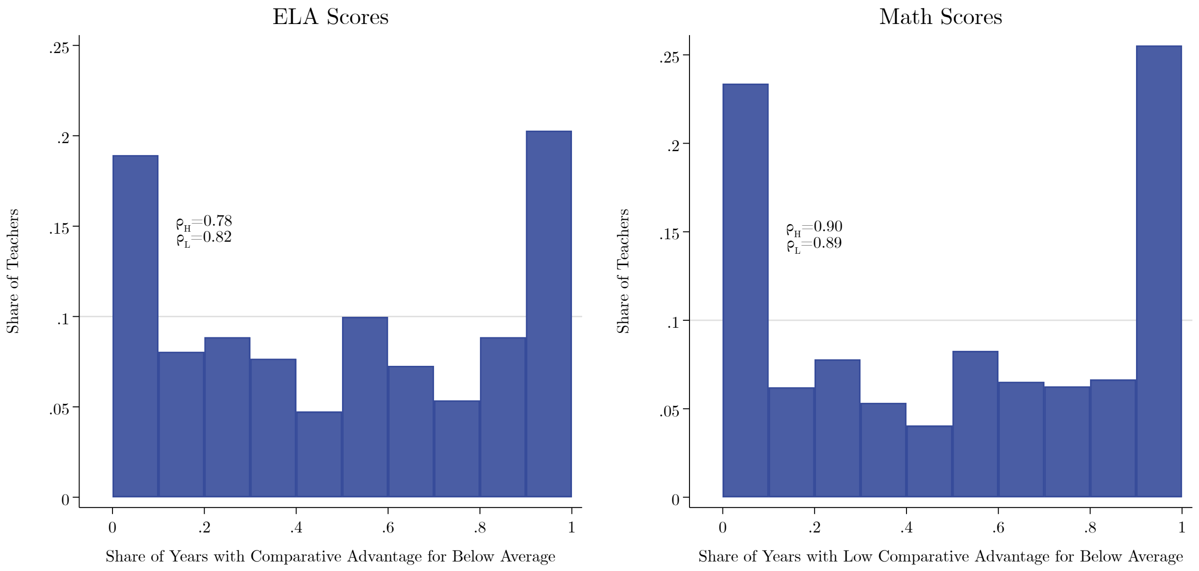
$$VA_j = \sum_k \gamma_k \bar{\omega_k} \alpha_{j,k} \tag{28}$$

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 28 does help to rank teachers based on what is under their control.

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

Figure D.6: Measures of Comparative Advantage Persistent



are fitting on true value-added rather than idiosyncratic noise.

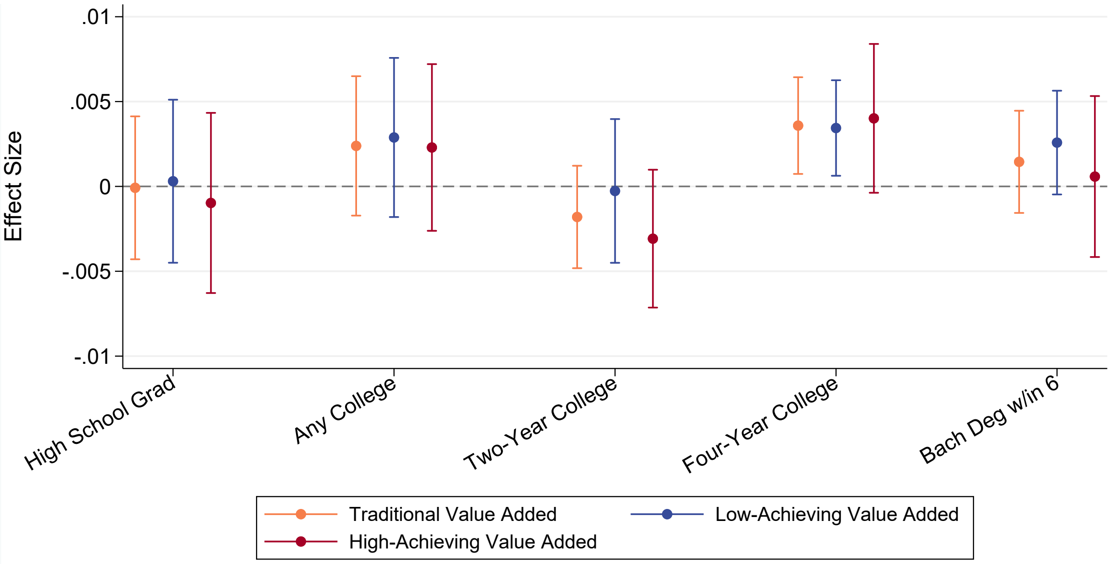
To test the predictive power of value-added, we regress each outcomes teacher value-added and the controls from equation **??** in a student-subject-grade level regression. For the binned estimates, we include terms for the high- and low-bin value-added interacted with an indicator for whether the student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

$$\begin{aligned} y_{i,j,s,t} &= \tau_{VA}\hat{\gamma}^{VA}_{j,s,t}\mathbb{1}(k_i = g) + \beta_2 X_i + \nu_{i,j,s,t} \qquad (29) \\ y_{i,j,s,t} &= \sum_{g=H,L} \tau_g \hat{\gamma}^{g}_{j,s,t}\mathbb{1}(k_i = g) + \beta_3 X_i + \nu_{i,j,st} \end{aligned}$$

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.7 reports the results from the regression in equation 29 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.7: 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 29 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 independently 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.